Reflections over teaching a class on data analysis with Coursera (written in both English and Russian)

Boris Mirkin
Professor, Higher School of Economics Moscow RF
Professor Emeritus, Birkbeck University of London UK

Contents Page

1. My class setting

I am a Russian scientist specializing in methods for cluster analysis of data presented in various formats.

Back in sixties, I graduated from both graduate and PhD programs in Computer Science from a provincial university in Russia.

Then I shifted to data analysis, which at that time was considered part of statistics. Two things contributed to my work: (i) a very fragmentary knowledge of international developments in the field, (ii) a computer science perspective on data, rather than the classical probabilistic perspective. Therefore I and my peers in Russia developed a somewhat different approaches to many a classical problem. I think that these developments are useful in practical data analysis.

This includes my general views on the goals and structure of data analysis as a teaching subject. My teaching style is highly affected by the fact that for many years I was teaching this material to students in Birkbeck University of London master's program in Computer Sciences. Many of them are quite versatile in coding though having a rather modest knowledge of mathematics concepts, therefore needing a less challenging, than usual, mathematics notation and explanation. Both my views and my teaching experience are reflected in my textbook "Core concepts in data analysis" published by Springer 2011.

I do think that my views, however weird they may seem to an educated specialist, do not contradict the common principles but supplement them in a beneficial for effective data analysis manner. Therefore, I was pleased when my university, NRU HSE Moscow, included my course among the first batch of Russian-proposed classes on Coursera. Having in mind my previous teaching experiences in London and Moscow, I developed a set of eight week-long lectures oriented at a student who is not quite good in mathematics. A difficult task of preparation of such slides that could serve as both an authoritative concept source and hands-on instruction for computations has been accomplished by me in March-April 2014. The teaching assistant, an HSE PhD program student, Ekaterina Chernyak, my collaborator on several other projects, helped me in this. Then the two of us addressed an absolutely unexpected problem. Because of large numbers of students expected, hundreds if not thousands of them, an unusual requirement emerged: to prepare such a system of knowledge control that could work automatically with no participation of either of us. We filled in the eight weeks with in-built quizzes, computation assignments and subjects for essays to be peer reviewed. We decided to give each student a randomly selected sample from the same data set for our computation assignments so that we could check only how well the student's answers match the answers given by our own computation. Unfortunately, Coursera site cannot support transfers of data sets to students. Therefore, we had to use an externally made tool for this. Luckily, Nikolay Vyahhi, an experienced Coursera user, had developed a platform, STEPIC, initially oriented for teaching bioinformatics classes, which could be used for this task; Nikolay proposed us his help which we accepted with gratitude, so the issue was solved.

Much helpful was the Springer publisher who made available pdf files of the relevant chapters from my textbook (2011) to subscribers free of charge, as well as by making concessions over the price for those who wanted to purchase a hard print or e-version of the book. Similarly, MathWorks company provided subscribers with tutorials for Matlab as well as a requested by me version of Matlab free of charge. To be fair, I should point out that I did not try to approach the Springer or MathWorks with my requests; just the opposite - their representatives, A. Birukau and C. Chitale, respectively, approached me asking how they could help. Of course, much help was received from the Department of HSE responsible for logistics and the entire Coursera project, Ms. Eugenia Kulik and Mr. Alex Mazurov, especially with the view that we were just one of a dozen or more classes to be delivered.

To my surprise, the class, scheduled at April-June 2014, received more than 57000 subscribers, much, much more than I had ever expected.Two greatest sources were the United States (28%) and India (16%). Other countries, like China, Canada, Russia contributed about 4% each. The professional level was rather high too; for example, 9% had a PhD degree. Most were full-time employed professionals (60%); full-time students made 23%. Gender distribution was somewhat biased though: only 23% were women (on average in Coursera, 40%). Well, about 22 thousand of 57 did not bother to visit the course site at all. Of those who did visit, about 22 thousand dropped off immediately. I guess my advert was somewhat misleading regarding the level of maths required (see later on this).

Then the numbers of students having watched the videos steadily declined from about 15 thousand in the first week to slightly less than 4000 in the eighth week. The number of students submitting answers to quizzes and assignments was even smaller, falling by the end of week 3 to just under a thousand. Yet students actively corresponded with

each other over various forums discussing unclear or erroneous statements in the course, and, especially, in homework tasks, as well as expressing opinions over various class-related aspects. The number of students who have passed all, including the final test, is about 400 of which two thirds with distinction. Of course, this is a small fraction of the subscriber numbers. Yet it is a huge personal achievement: perhaps this number is greater than the total number of conventional students who completed my class for the past decade.

Unfortunately, while actually running the class, I did not see student forums at all. According to the Coursera protocol, the actual running is supervised by the TA, HSE recent graduate, Ekaterina Chernyak, who approached me rather rarely, usually when several students claimed that it was something wrong with this or that formulation. Now I have looked through most forums and I feel overwhelmed with the scope of issues raised in the discussion. Further on I am going to review the subjects discussed in their relation to the class.

2. A general view of the discussions on student forums

A set of critical comments went from students who joined the class mistakenly and did not expect that much computation work; some stated this explicitly, some implicitly. For example, some referred to my poor English that made the quizzes incomprehensible (although this claim was refuted by others). Some well formulated but rather unusual tasks, such as evaluation of the precision of the regression equation, also were initially attributed to issues of the Russian-English translation. Other drop-outs cited using a "commercial" package MatLab instead of anything publicly available, although the package was available to the subscribers free of charge and, moreover, we specifically advertised that marks did not depend on the programming tool utilized and we did not care of that.

Much critics concentrated on blunders and errors that I admitted on slides, especially in the first lectures. Some typos have been found in the Springer published text (2011). I accept all of them with gratitude. I do apologize for them. I have corrected corresponding places so that these errors should not emerge anymore.

Technical issues, especially in the beginning, also led to questions and discussions. These went from questions of how to get MatLab or the textbook and how much that would cost - although the free availability of both is very clearly outlined in the menu - to comments on issues of formatting the answers to test questions and homework to get them through at submitting.

The issue of peer grading has led to a more or less animated discussion. Coursera admits three types of tests to be given to students while lecturing: quizzes, programming assignments, and peer assessments. To make our class independent of computational platforms and programming languages, we decided that all the assignments should request computations and check the results rather than the ways of obtaining them. This has been well received by the students. To address a cumbersome task of inclusion of peer assessments, we decided to limit peer assessments to writing essays describing main concepts in the class only. The peer reviewing, that is, marking each essay by a designated group of students, is a distinctive feature of Coursera teaching process. To make it as independent of subjective mood changes of the peers, as possible, we decided to instruct peer markers so that they should base their mark on just only checking for the presence in the essay of one or two features mentioned in the definition of the concept under consideration. Also, we deliberately lowered the contribution of peer grades to 10% of the total mark. Appendix 1 is a sample of unedited expressions of opinion on the issue of peer assessment by students. The reader can see how the initial worst expectations are steadily changing for more favorable views towards recognizing the peer assessment as a useful teaching device.

Several threads on forums praised various fragments of the class as well as the class as a whole. The depth and magnitude of such comments make me think, that overall I have managed to get my message through. A representative sample of such messages is in Appendix 3. Only those with an informative part have been put there. A few students, of higher standing I guess, explicitly expressed their wishes to participate in the next runs of the course as TAs - this is noted and much appreciated.

3. Helpful discussions and lessons learnt

To shorten my exposition, I am going to list those of my assumptions that have been so greatly challenged by the students' comments, that I am going to take care of the assumptions when and if the class is run for the second time.

A. The role of typos and blunders

Having written and published a relatively large collection of research papers and monographs, I hold a rather fatalistic view that typos and blunders are an inevitable evil which cannot be avoided however many times one checks their own text. Moreover, when teaching a conventional class, an occasional blunder may be useful as a device to make the audience more attentive to contents of the learning material and, thus, more active. This happens when one or two bright students notice a blunder and voice their concerns, so that they and other students come to participate in the process of correcting the blunder by the instructor. Yet on a distance-based course, after seeing a few blunders, a student who has encountered an unclear claim or task may prefer to think that the issue is not because of her/his misunderstanding but just one more blunder by the instructor. Therefore, blunders are a real evil at distant learning to hinder the learning process - they must be avoided at any cost.

I do hope that by now, after taking into account the student comments, my slides are blunder-free (well, almost!) and what remains to be done is to revise my presentation accordingly.

B. Examples and instruction in tests

Many complaints were related to the lack of flexibility in formats of answers to testing tasks accepted by Coursera, and further on, by STEPIC tool, which made students to waste their efforts and energy on technicalities rather than on substance. Some felt that the time limit of 5 minutes for downloading the data sent them for computational tests was too restrictive. Many complained that they could not understand the questions in quizzes, I think because of lack of understanding of the class material, although many of them suspected that the meaning was lost in translation from Russian to English. My view is that, after more than a decade of teaching in the University of London, my language is not that bad; moreover, some students in Birkbeck's introductory classes did tell me that they preferred me teaching (over some mother-tongue English speaking counterparts) because I speak loudly and slowly. It should be probably mentioned that the Coursera web site did not advertise my international standing among subscribers.

I think this all can be addressed by using more instruction and exemplary answers alongside the tests.

C. Mathematical precision and positioning the class

My image of "an ideal student" for my class is highly affected by my teaching experience in the Department of Computer Science, Birkbeck University of London. The Birkbeck College is a brand-name for a part-time university in the UK. Correspondingly, my customary students were programmers who sought obtaining a Master in CS diploma. These students were not terribly good at mathematics; occasionally I encountered a BSc in Computer Science from a UK University who had never heard of the concept of derivative, the base of calculus. Yet all or almost all of them were good in computing. Therefore I had to accommodate the class contents in such a way that such students did not much suffer of their lack of mathematical background. Accordingly, under the implicit assumption that no serious data analysis can be conducted without computing, I advertised my class as not requiring a deep mathematical background, yet I failed to mention that a heavy computational load was involved. Therefore a few individuals lacking both mathematical and computational backgrounds were lured into subscribing to the class, but of course they recognized very soon that they were not able to follow the mathematics and bear the computational load and dropped out. To address this issue, I should be more specific in my description of course contents and explicitly tell that a student is expected to be able to take and make a number of computing assignments.

Yet there have been criticisms, from the opposite end, that the course lacks a mathematical precision.

Indeed, consider a seasoned student who has experience in learning a class or two in mathematical statistics. Such a class proceeds as follows: first, a mathematical structure is defined, then its properties are mathematically stated if not proved, then this structure applies to a number of tasks in statistical estimating or hypotheses testing. My class does not follow the suit: it is of data analysis concepts and their use in advancing into real-world data summarization

or correlation issues. Mathematics here is but a tool. For example, when explaining the concept of linear regression of y over x, mathematical statistics considers a bivariate distribution on the plane of (y,x) pairs, defines the concept of conditional expectation of y over x, proves that this minimizes the quadratic error and derives a corresponding decomposition of the unconditional variance of y, after which states that the conditional expectation is a linear function of x when the distribution is Gaussian. A parameter of this distribution, the correlation coefficient, has something to do with both the slope of the linear function and the squared error. Then come applications of these to testing statistical hypotheses of the model such as the hypothesis that the slope/correlation coefficient is zero. This all is quite elegant, nice mathematics, all things are proven, no questions emerge; one can move to studying another chapter. In the context of data analysis, I am trying to show what can be derived from data without the use of the model of bivariate distribution. I arrive at the concept of data-based correlation coefficient (which would be just a sample-based estimate of the theoretical correlation coefficient in the mathematics-statistics treatment) without ever using the Gaussian assumption of the distribution. Then I look at the values of correlation coefficient at some data examples and I see that the coefficient can be zero, hiding the real correlation just because the sample is not homogeneous, or as large as possible just because of outliers, even if it is zero on the base sample. These well justify a famous saying that there are three levels of lies: just a lie, a damn lie and statistics. These issues are glossed over in mathematical statistics classes - they teach models, but not in my class - I teach concepts.

The following discussion presented here with three unedited messages, continued in Appendix 2, although differently worded, in fact concerns the difference between my class and mathematical statistics classes.

Hans Tunggajaya·

I find the course lacking of mathematical precision. I'm not talking a really rigorous mathematics, but the level of Ng's Machine Learning course, which I find quite balanced between intuition and rigor.

3votes received.

Anonymous·

Introductory fact # 1: I have been doing this course really well, so I am not frustrated, angry or something like that.

Introductory fact # 2: I have attended a number of machine learning and statistics classes both online and at my universities, so I am quite familiar with the subject.

Bottom line: This class is by far the messiest one I have ever had about this subject. I really do not understand why the Professor is avoiding to use a bit more formal math in the slides and why he does not spend a bit more time explaining some formulas. For example, my view: All variables written in the slides should be visited and explained, at least mentioned. Otherwise, their existence is just pointless (take this sentence with a grain of salt). He apparently thinks that this informal way, without using a lot of math, would help the crowd, but this deliberate exaggerating at avoiding math is making the whole thing just messier, in deed. If someone want to use data analysis, and I believe that the guys who are attending this class attempt to do, it is not unreasonable to expect from that person that he or she should be prepared to do some more serious math. Quite opposite of his view, I think - people would even appreciate a bit more formal approach to clarify the foggy concepts, as being presented here. This is a highly quantitative subject, and it should be presented as one. If the time for the class is limited, as it is here, it is better to have, say, 5 concepts quite well explained and in detail, than 10 concepts just barely glanced over within the same amount of time. Furthermore, even a bit more serious issue: the fact that the Professor often uses his own notations and his unconventional views for some methods in an introductory course, like this one, is not helping at all. The introductory course should use common views and widely accepted concepts to put people on the stable ground, to give them some knowledge that could be easily transferred to other courses and the literature in the field; whereas some more advanced courses could afford to use some non-standard, hipster like if you want, interpretations. I am deeply grateful for this class and for the effort of the Professor and his assistants, but I cannot be honest and say that I am enjoying this class, I am taking it just to obtain a certificate, which I will mention in my CV. P.S. Peer assessments are just a joke, really. Writing a sentence or two is not helping at anything. Essays, like the first one, are very helpful, but the two sentence responses are simply not. My plain, honest opinion.

4votes received.

Wilko Dijkhuis·

Wow, I'm starting to like this course very much.

However, as a polytechnic teacher of elementary statistics, I have one big reservation. The description of this course suggests that it is accessible to statistical novices. Some might hope that this is a gentle (i.e. non-mathematical) introduction. To those who are lured into this course by this suggestion, I – a professional teacher of gentle introductions to statistics –have one urgent advice: abandon all hope.

This is a master class well suited for those that already are familiar with the subject but want to have a second – data analysis centered - look at the foundations of their field.

Yes this course is not mathematical. But here that means two things:

1) the praxis of actually doing data-analysis is leading. The end is not to give interesting new mathematical proofs. The end is to use existing mathematics to do interesting data-analysis.

2) The underlying mathematics is loosely indicated, but not rigorously developed (a horror to all real mathematicians; in that sense the approach is anti-mathematical instead of non-mathematical)

PS For those that want a gentle intro I can warmly recommend: http://exexstats.tumblr.com/post/58597571228/tips-for-www-learning-statistics-intro-to

14votes received.

Appendix 2 contains a copy of the message by the Anonymous above and a follow-up discussion refuting the major points raised in this message and explaining how the class contents should be perceived.

My conclusion: I should better accentuate the main tasks of the core data analysis as well as the goals of my class. As stated by Wilko Dijkhuis above, "The goal is not to give interesting mathematical proofs but rather to use mathematics to do interesting data analysis." In this, I should clearly state that a mathematical background is needed as a prerequisite. Perhaps some knowledge in computers should be required too.

D. Unusual perspectives

As data analysis models involve feature values themselves rather than their probabilistic distributions - the main device in the classical statistics, they can bring unusual insights into the nature of the concepts. I tend to highlight these when they are rather elementary such as the meaning of the celebrated Pearson chi-squared contingency coefficient as an association index, not just a criterion for independence testing as it is considered usually. However, less straightforward implications remain unnoticed. One such item has been mentioned on the forum by some concerned students.

This is of multivariate analysis of nominal features (the subject of an in-lecture quiz). I thought that I clearly indicated a path for using the quantitative analysis methods explained in lectures by enveloping nominal categories as 1/0 binary variables. This however was entirely missed by most students, perhaps because this advice went over a specific data table, not as a general statement. When they looked at the internet, they found no advice either. This makes me think of adding a section explaining how to apply methods under study to nominal and mixed nominal-quantitative data.

Of course, the task of explanation and justification of other unusual statements remains one of my first priorities.

This is especially important because many students, when encountering some new or unclear concept, use Google or another search engine to see how this concept is defined and considered in other sites. Luckily, other sites recently began considering deep relations between Principal component analysis and SVD decomposition, the cornerstone of my treatment of the subject. Yet some other useful concepts such as Quetelet index for capturing association between categories are not widely circulated, which may frustrate some listeners too. In this regard I can only wish that the students would be more attentive to the materials in my lectures. Say, I provided examples of wrong values of the correlation coefficient because of no homogeneity or outliers and referred to these cases as examples when

statistics is the biggest lie. Yet a student explaining this paradox to another student preferred to refer to the authority of other websites rather than to these examples.

E. The humor

Many students pointed out that listening to my lectures was sometimes entertaining because of accidental humoristic overtones.

Dipanjan Sarkar· 2 months ago

I agree, the lectures are short but quite informative and the professor doesn't make it seem too boring by adding a bit of humor here and there. Nice topics to be covered too, looking forward to the remaining weeks of the course.

8votes received.

Dan Northrup. 7 July (an extract from a personal message)

Dear Dr. Mirkin,

I really enjoyed your class and the many jokes you made through the lecture. Please continue to offer such courses and know that they are very well appreciated by your students.

This reaction appeared somewhat unexpected to me because I tried to be more or less formal in my Coursera lectures. However, having looked through the videos I found funny moments indeed, as, say, when I give examples of the power law and mention Google and Amazon sites as the most popular and my own site as the least popular. Another example is my usage of the catch-phrase "Clash of civilizations" when referring to a relatively minor, from the life-and-death point of view, difference in perspectives on data analysis maintained by data miners and statisticians. Probably indeed an acute feeling of differences in scales leads me sometimes to expressing this or that as a kind of grotesque.

HSE surveyed students before and after the class (K.A.Kuzminykh). Main results are presented in the Supplement in the format of two collections, one of comments of what they did not like at the class, and the other of comments of what they did like in the class. There are also two reference graphs summarizing and visualizing the opinions, as derived by E. Chernyak and myself out of these collections by using a method that we had proposed earlier for the analysis of text collections. These confirm and detail the opinions expressed in these reports.

4. Conclusion

Overall, apart from the tiredness of the hard labor I was engaged at when preparing my slides and rethinking their contents over and over again, I feel a deep satisfaction of the course and I am going to further improve it in the directions pointed out in the previous section. Specifically, I should take maximum care of:

- Avoiding typos and blunders at any cost;
- Providing more pointed examples, templates and instruction at knowledge control;
- Properly positioning the class
- Giving more explanations to differentiate between mathematical and non-mathematical parts of the course;
- Better highlighting the core elements of data analysis and the goals of my class;
- Giving a wider range of clearly outlined unusual implications of my approaches to the data analysis;

- Continuing the humoristic component in the lectures.

I think I should explore the option of making some parts a bit lengthier, as was clearly suggested by some students on forums. Probably I was too much concerned with the request that a part should not be longer than 10 minutes.

Besides, I am thinking of developing a software package to cover the methods under study in such a way that not only results of applying them to data can be seen but intermediate results found at individual steps of the algorithms as well.

Of course a better feedback system while running the class would be of help in solving immediate issues. Probably, a forum directly oriented at me with my replies to, say, a dozen messages a week, on a specified week-day could be of value as well. I am not quite optimistic on this since students' comments indicate that our course is friendlier than many others in this regard. Probably I should spend more time in thinking how could we integrate in practice efforts of individuals who wish to take part in the TA duties.

This is especially important because the Teaching Assistant on this course is much busy during all the 9-10 week period of actually running the class. We should restructure the TA's responsibilities in such a way that they can be performed by different people independently. This would allow us to use help of the international volunteers.

Appendix 1. Discussion of peer assessments on a forum: is it like going to a dog show where none of the judges has ever seen a dog before?

Anonymous· 2 months ago

Those peer assessment are joke. I just give up now !! I regard this course waste of time ....too many failures here!

-14votes received.·

Anonymous· 2 months ago

I think the knowledge here is very useful. I have issues with peer assignments in all of the classes that implement it but that does not stop me from learning what I can from this very smart professor who has probably forgotten more about this then I will ever learn.

I just do like I did in school - hop on the B-train so I can concentrate on learning and not jumping though unnecessary hoops for a grade. Learn the subject, prove that I know it and move on with life. My theory of life.

12votes received.

Francesco Trentini· 2 months ago

You can ignore peer assessment if you want. It's up to you. And if you care about certificates, you can still get one "with distinction" because the peer homeworks account for just 10% of the final grade according to the policy.

6votes received.

Anonymous· a month ago

I don't like classes with peer assessment either, but think the Prof has much of value to say. I hope he does a second iteration and drops the PA.

3votes received.

Thomas Vimont· a month ago

Can someone explain what is wrong with peer assessment ? It is the first course I take where I use peer assment. I like it because it forces me to think about some not so hard questions and make me explain in a clear way my thoughts. Usually I think I know something and I have difficulties when I actually try to explain it. For this week, I do not have a clear idea of the answer so I have to look for some insights on the internet and learn new stuff.

To me it is useful but everyone does not have the same background and objective so maybe it is not for you.

3votes received.·

Anonymous· a month ago

Thomas, peer assessment is a lot like going to a dog show where none of the judges has ever seen a dog before. (I am indebted to someone whose name I forget for this explanation.)

Some peer assessors take themselves too seriously and try very hard to mark down other students' work. I almost failed one course because of a couple of stupid peer graders. It is very de-motivating when a large part of your grade is determined by self-important idiots. (Of course, it is also very de-motivating when the computer does not recognize your correct answers.)

I include myself among those who do not really know what a dog is. So why should I be appointed a judge?

I think multiple-choice quizzes are the safest way to avoid unfair grading.

P.S. I don't like the down-voting feature on Coursera fora. People should be allowed to express their opinions without getting down-voted by fellow students. Down-voting makes the fora nastier and ruder than they need be.

2votes received.

Thomas Vimont· a month ago

Thank you for the explanation. I can understand the frustration but as said above, peer assessment is only a small part of the mark in this course.

I have the other bias when I mark : if I feel that one has tried to produce an answer and that it is understandable, I give a good mark even if I gave the same mark to someone who made a much better answer. If enough people mark, it should converge to the true mark ;-)

Anonymous· a month ago

Love the dog show analogy. That is EXACTLY how I feel about peer assessment. (No offense to some here who probably really are experts), This course I'm not taking for the certificate (time issues) so not a big deal. I find the lectures interesting, and I bought his book to get deeper into it. Buying the book also gets him a little money which is my way of thanking him for sharing his expertise for free.

Holger Wenzel· a month ago

Sometimes peer grading can be unjust or you feel that some of your peers gave you an unjust grading.

This happens pretty often when a course is taken by advanced students who write their assignments in such a way that their less advanced fellow students don't understand them an give them a bad grade. Then the discussions can turn pretty nasty.

Just sticking to the question asked in the assignment and explaining your answer is such a way that it is intelligible to students who know nothing about the subject than what was taught in the course so far normally gets you around this problem.

In my view the peer grading can teach you a very important point: Think about your audience and write in a way that it is easy to understand for exactly this audience.

I like the way this course handles the peer assignments. They are present, but are weighted pretty low. So it is not a catastrophe if one of the peers gives you a bad grade, just because he is in a bad mood.

6votes received.

Marc Chaplin 9 July 2014

Thank you to Prof Mirkin and Ekaterina Chernyak.

The course offered a very interesting slant on some great topics, especially Quetelet and the entire PCA section. I really felt that the peer review exercises also contributed a lot towards the learning process - which is not the case with many other online courses - and hope others experienced likewise.

It was also very refreshing to see the high level of involvement of staff on the forums - great customer service!

Appendix 2. Discussion of teaching specifics of the class by experienced attendees

Anonymous· a month ago

Introductory fact # 1: I have been doing this course really well, so I am not frustrated, angry or something like that.

Introductory fact # 2: I have attended a number of machine learning and statistics classes both online and at my universities, so I am quite familiar with the subject.

Bottom line: This class is by far the messiest one I have ever had about this subject. I really do not understand why the Professor is avoiding to use a bit more formal math in the slides and why he does not spend a bit more time explaining some formulas. For example, my view: All variables written in the slides should be visited and explained, at least mentioned. Otherwise, their existence is just pointless (take this sentence with a grain of salt). He apparently thinks that this informal way, without using a lot of math, would help the crowd, but this deliberate exaggerating at avoiding math is making the whole thing just messier, in deed. If someone want to use data analysis, and I believe that the guys who are attending this class attempt to do, it is not unreasonable to expect from that person that he or she should be prepared to do some more serious math. Quite opposite of his view, I think - people would even appreciate a bit more formal approach to clarify the foggy concepts, as being presented here. This is a highly quantitative subject, and it should be presented as one. If the time for the class is limited, as it is here, it is better to have, say, 5 concepts quite well explained and in detail, than 10 concepts just barely glanced over within the same amount of time. Furthermore, even a bit more serious issue: the fact that the Professor often uses his own notations and his unconventional views for some methods in an introductory course, like this one, is not helping at all. The introductory course should use common views and widely accepted concepts to put people on the stable ground, to give them some knowledge that could be easily transferred to other courses and the literature in the field; whereas some more advanced courses could afford to use some non-standard, hipster like if you want, interpretations. I am deeply grateful for this class and for the effort of the Professor and his assistants, but I cannot be honest and say that I am enjoying this class, I am taking it just to obtain a certificate, which I will mention in my CV. P.S. Peer assessments are just a joke, really. Writing a sentence or two is not helping at anything. Essays, like the first one, are very helpful, but the two sentence responses are simply not. My plain, honest opinion.

4votes received.

Anonymous· a month ago

I actually appreciate the fact that he's giving us his unconventional view. Otherwise, why bother to have him teach it - just some guy from the local community college can cover the conventional basics. What for me is amazing about these MOOCs is that you can take a class from true experts and get a look into their heads so to speak. Sometimes a bit more math would probably help, but as a first iteration, he's probably trying to find a good balance.

Anyway, I'm the anonymous above who doesn't see value in peer revues, and haven't down voted you or anything because I like honest feedback and opinions. I'm not working toward a certificate due to time limitations though I've earned five from Coursera, HarvardX and CalTechX.

2votes received.

Anonymous· a month ago

I would say that the interesting fact about this course is precisely that he is challenging all that we know from other courses!

If you have been around data analysis and people doing machine learning you may have noticed a certain degree of bigotry... Sometimes this expresses itself in another wild flight of mathematical gibberish which nobody seems to care except to get a paper into a conference: many times I find that I would rather somebody told me what they were trying to accomplish than just present me with the empteenth matrix equation of some quirky loss function.

What professor Mirkin is here telling us is no rocket science either conceptually or mathematically and most of us seem to have the knowledge to supply the missing formulation. But I like his insights and accumulated wisdom (yeah, he is way data-wiser than I am, at least!).

I think you should take the insights and gloss over the messier aspects. Every course has them, if you look back on them a little bit! ;)

5votes received.

Francesco Trentini· a month ago

Agreed... I'm following the course precisely because it is a well informed unconventional presentation. If I wanted conventional I would have gone elsewhere.

To those requiring more "rigor" take time to grab a copy/e-copy of the textbook, it has much more detail and a pleasant format and presentation. Lectures have an obvious time tradeoff to obey.

3votes received.

Anonymous· a month ago

Indeed! What called my attention to professor Mirkin's work was his 1995 book which is still one of the most inspiring references on clustering. This is the earliest place where I've found "co-clustering" or "bi-clustering"... Does anybody know of an earlier one?

As an aside, already, I have found a couple of suggestions for papers just by listening to his lectures. A waste of time? I should say not!

5votes received.

Sushil Bhattacharjee· 13 days ago

Introductory fact: the two introductory facts stated by the Anonymous author who started this sub-thread also apply to me:)

It's not true to say that everyone in this course wanted to have a mathematical presentation of the subject. I quite liked the way Prof. Mirkin presented the topics in this course. I think he tried to show us how to "eyeball" the data; how to draw qualitative conclusions from all the quantitative analysis. In addition to the lectures he has made available the relevant chapters of his book. People looking for a rigorous mathematical treatment of the subject could always refer to the text-book.

Anonymous· a month ago

The course is NOT waste of time. It 's experience and knowledge. I agree bout peer assesment. But the course is very good. Thanks a lot to prof. Mirkin! I know what I say . I also work in this area plus I am a Fullbright professor.

4votes received.

Appendix 3. Messages of praise and gratitude[1]

Anonymous· 2 months ago

Pity I chose the wrong set of courses to get started with my data science learning... discovered this course too late (just today).

I simply loved the week1 videos, and Prof. Mirkin's approach in general.

Any idea about the next offering? If it's going to be within next couple of months or so, I'll wait; otherwise I would probably get started with this course right away...

Thanks!

Arvind Tiwary· a month ago

Superb coverage of the distribution of the mean of a sample in chapter 2 of the book. Great to see that even if the underlying population is not normal and has power law or log distribution the sample mean will converge to Gaussian. The example of mixed distributions was also great.

This alone makes this course worthwhile for me. I have got much more then I was hoping for.

THANKS Professor

Anonymous· a month ago

Professor - what a great insight in iK-Means! Simple and effective. Thank you for that. Brilliant.

Craig Milhiser· 11 June 2014

Professor and Ekaterina,

Thank for you a great course. I enjoyed learning new ways to approach some problems that I thought I understood and new skills. I liked that you did not treat all topics like a survey course but instead you went into some depth with the topics. I also liked that the videos show the professor.

I know it is a great effort to prepare the videos, assignments, participate in the discussion boards and adjust what you can during the course. I appreciate your effort and I learned a lot from you.

philip law· 20 June 2014

I greatly appreciate Professor Boris Mirkin's course. He is very insightful and the choice of topics and the presentations are superb. His presentation is often from an unique perspective different from ordinary textbook presentation of the same topics.

Neill White· 20 June 2014

I concur with the kudos. I think the instructor did what he intended: expose and elucidate the core concepts of data analysis. Most of the core concepts were explained at such a level that we could write the functions ourselves (SVD being the lone exception). The lectures were very well thought out and even a little entertaining. Thank you to Prof Mirkin and TAs.

[1] Only those related to my work are here; those praising the work and feedback by E. Chernyak are not listed.

Francisco J. Valverde-Albacete 21 June 2014

I entered the course to become better acquainted with Prof. Mirkin's views on data analysis and I must say that it has fulfilled my expectations. Information and knowledge you can find in books, but insights and wisdom come mostly from the scientist-practitioner.

Thanks very much for the course! (And your books too!)

Mustakeem Khan 21 June 2014

Great course. It was a challenge. The people on the forum helped me get through this course given some of the challenges with the assignments. Between Coursera, the google search function and youtube, I am getting an excellent free education.

Thanks to the professor, the TA's and all of my classmates on the forum. I am having fun with this.

Sushil Bhattacharjee 21 June 2014

I am glad someone started this thread. I would also like to thank Prof. Mirkin for the wonderful course. I was already familiar with most of the topics, but I found Prof. Mirkin's perspective on the various topics very insightful. I hope the Prof. Mirkin will consider giving other courses on related subjects via Coursera.

divya subramanian 22 June 2014

I too would like to thank Prof.Mirkin,TAs and all the fellow coursemates. All this while I was wondering on how to understand concepts in data analysis and this course proved be very interesting starting point. I learnt lot of new concepts , the quizzes were very challenging which kept my enthusiasm high and ofcourse the discussion forum was a great help which shifted my understanding to different perspectives. Please do come up wiith a next level for this course.

Richard DeVos 23 June 2014

Professor Mirkin and Staff,

Thank you for the great course. Your efforts in preparation were obvious. As a result, I have gained a renewed appreciation for linear algebra, an understanding of the powerful ability to extract meaning from data that is much too large for us to understand or sift through.

It is thoroughly enjoyable to have developed an understanding where none existed.

Keep up the good work

Rick

Krsto Prorokovic 24 June 2014

This is one of the best (if not the best) courses I've taken so far.

Lectures are interesting and some advanced material not mentioned in many textbooks is covered. Homeworks are great - we are not restricted to particular programming language and they are somewhat practical; you give us the data, we give you the answer. I really enjoyed this course.

Thank you prof. Mirkin and Ekaterina.

Robert Brookes· 2 months ago

Definitely not a 'first' course in stats (or in R etc.), but I really appreciate the different perspective. This is one of the more enjoyable courses I have taken so far on Coursera. I also think it is a good course to take alongside the data specialisation courses as it provides a deeper theoretical (and dare I say at times philosophical) foundation, while retaining practical elements. I have also been following along in the book, which helps.

4votes received.

Gustavo Esquinca Ledesma 27 June 2014

Thanks for the concept, every course I take on coursera give me motivation to keep learning. The math and programming was a little difficult for me, but thanks to the discussion forums, I could get through the test and assigments. I hope to take this course againg in a future and see if I'm able to undestanding everything by myself. But however the concepts and the K-cluster assigments was superb.

Thanks a lot!

zhitao.zhang· 1 July 2014

Great course! let me know enhance some knowledge not only just apply, but also why. I like the Professor Mirkin standing in front of me...enjoy!

Matt Taylor 4 July 2014

Excellent job on the course! Thank you Professor Mirkin and staff. I appreciated the varied format of the homework in particular. The quizzes, programming assignments, and essay questions provided a variety of stimulating tasks through which the content of the course could be exercised.

I was originally looking for a Coursera course on pure statistics. I reluctantly signed up for this course since it was close, but not exactly, what I was looking for. The course definitely exceeded my expectations and I think I learned many things I would never have been exposed to in a regular stat class. Thanks again.

Marc Chaplin 9 July 2014

Thank you to Prof Mirkin and Ekaterina Chernyak.

The course offered a very interesting slant on some great topics, especially Quetelet and the entire PCA section. I really felt that the peer review exercises also contributed a lot towards the learning process - which is not the case with many other online courses - and hope others experienced likewise.

It was also very refreshing to see the high level of involvement of staff on the forums - great customer service!

Zacharias Voulgaris· 2 months ago

Maybe I'm a bit biased because I'm in love with this subject, but I really find this course really good. It manages to clarify the main concepts of data analysis without getting too focused on the programming side. Even though I'm quite well versed in this field, I find that there are things to learn (e.g. this SVD analysis which I never fully grasped) and have fun doing so.

Kudos to the professor and the staff!

14votes received.·

Diego Mastrogiuseppe· 2 months ago

I really like this course too. I devoured the two first weeks of lectures in a couple of hours, so I'll be impatiently waiting for the upcoming videos.

It also motivated me to start improving my basic knowledge in Phyton and R.

Ashish Singh· a month ago

Personally, I think this course is very underrated. A lot to be earned here than just a certificate !

8votes received.·

Anonymous· a month ago

Yes I agree. I have taken many data science courses none of the courses have taught the perspective that Prof. Mirkin has taken. It has been great so far. As well its nice to see how people in Russia teach/learn. We can only wish our russian was a good as their english is.

Finally I attend a lot of big data/data science meetups here in silicon valley and none of them EVER talk about anything that professors talk about. They are mostly very practical about SQL and noSQL coding skills. So all you folks getting aggravated, keep the larger perspective.

8votes received.

Размышления по поводу опыта, полученного при преподавании курса анализа данных на сайте бесплатных он-лайн курсов Курсера

Борис Григорьевич Миркин
Профессор, Научно-Исследовательский университет Высшая школа экономики, Москва РФ
Почетный профессор, Биркбек колледж, Лондонский университет, Лондон СК

Оглавление Стр.

1. Общее описание курса

Я – российский научный работник, специализирующийся на разработке и применении методов кластерного анализа данных, представленных в различных форматах.

В далекие 60е я получил высшее образование, а затем и кандидатскую степень в области компьютерных наук в одном из провинциальных российских университетов.

Затем я стал заниматься разработкой методов анализа данных, которые в те годы считались частью математической статистики. Мои научные разработки складывались под влиянием двух факторов: (1) очень фрагментарными сведениями о зарубежных разработках в этой области и (2) взгляд на данные с позиций информатики, а не классической математической статистики. Поэтому я и мои коллеги в России развивали подходы к анализу данных, несколько отличные от тех, что развивались за рубежом. Я думаю, что наши разработки полезны в прикладных исследованиях.

Я включаю сюда и свои представления о целях анализа данных, а также и его структуре с точки зрения преподавания. Мой стиль преподавания во многом определяется опытом, приобретённым во время преподавания подобного курса в магистерской программе в области компьютерных наук в Биркбек колледже Лондонского университета. Многие студенты в этой программе имеют хорошую подготовку как программисты, но их подготовка в области математики обычно оставляет желать много лучшего. Поэтому они нуждаются в более поверхностных, чем обычно, математических построениях. Мои представления об анализе данных и накопленный опыт преподавания отражены в моем учебнике «Ключевые понятия анализа данных», опубликованной на англ. языке в 2011 г. издательством Шпрингер. (Русский перевод его начальных глав в модернизированном варианте составил основное содержание моего учебника-практимума для бакалавриата «Введение в анализ данных», Юрайт, 2014, 176 с.)

Я считаю, что мои представления об анализе данных, хотя и не совсем вписываются в современную международную систему, но и не противоречат ей, а скорее дополняют ее в направлении повышения эффективности разработок по анализу данных. Поэтому я с удовольствием воспринял предложение нашей Школы включиться в разработку бесплатного он-лайн курса для дистанционного обучения на одной из популярнейших страниц, Курсера, в числе первых российских предложений в этой области. Используя опыт преподавания, накопленный в Лондоне и Москве, я разработал совокупность 8 лекций, ориентированных на студента информатики, несколько «плавающего» в математике. Составление таких слайдов, которые бы могли одновременно служить заменой учебника и в то же время инструктивным материалом по обработке конкретных данных, оказалось довольно трудным делом. Пришлось существенно уточнить некоторые «ходы», которые подразумевались сами собой в статьях и монографиях. Как бы то ни было, к началу марта 2014 слайды были готовы. Мне помогала Екатерина Черняк, моя аспирантка и правая рука в нескольких других проектах, которая была учебным ассистентом данного курса. В этом курсе мы столкнулись с новой совершенно неожиданной проблемой. А именно, оказалось, что привычные методы оценки знаний на основе взаимодействия преподавателя со студентом в заочном курсе на Курсере неприменимы. Ожидаются сотни и тысячи студентов – методы оценки знаний должны быть изменены соответственно, путём полной автоматизации. Мы решили использовать все три формы оценки знаний, предусмотренные в Курсере. А именно, мы разработали встроенные в лекции вопросники с ответами в закрытой форме, а также вычислительные задания и эссе для взаимной оценки студентов студентами. Вычислительное задание по применению того или иного метода предъявляло студенту случайно выбранное подмножество некоторой таблицы данных и требовало сообщить параметры ответа, которые мы могли сравнить с параметрами нашего собственного решения. К сожалению, Курсера сама по себе не может обеспечить передачу данных каждому отдельному студенту. Надо было найти какой-то другой способ. К счастью, Николай Вяххи, опытный преподаватель онлайн курсов, в том числе на Курсере, предложил нам использовать для этого платформу СТЕПИК, разработанную ранее с его участием для обслуживания курсов биоинформатики. Мы с благодарностью воспользовались этим предложением, что решило проблему.

Издательство Шпрингер помогло нам тем, что предоставило бесплатный доступ к файлам тех частей моего учебника (см. выше), которые использовались в данном курсе, а также предложило студентам 20% скидку на эту книгу. Аналогично, компания Матворкс, разработчик системы Матлаб, бесплатно предоставила подписчикам запрошенную мной студенческую версию Матлаб и разместила качественные учебные материалы на сайте. Необходимо отметить, что помощь была получена не по нашей инициативе.

Представитель издательства Шпрингер А. Бирюков и представитель компании Матворкс Ч. Читэйл сами обратились ко мне с вопросом о том, как они могут помочь. Конечно, работа соответствующего департамента ВШЭ в лице Евгении Кулик и Александра Мазурова была решающей как в целом, так и в деталях.

К моему удивлению, на курс, запланированный на апрель-июнь 2014, подписалось более 57 000 человек, неслыханная для меня цифра. Два главных источника студентов – США (28%) и Индия (16%). Три другие страны – Китай, Россия, Канада – дали по 4%. Профессиональный уровень студентов был очень высок, например, 9% имели степень PhD. Большинство составили работающие на полную ставку (60%) и студенты (23%). Распределение по полу оказалось довольно смещённым – только 23% женщин, тогда как на Курсера их общая доля составляет 40%. Около 22 тысяч из 57 ни разу не открыли сайт курса, а другие 22 тысячи покинули его сразу же по появлении. Вероятно, я недооценивал требуемый уровень математики и программирования в своем объявлении о курсе.

Количество студентов, просматривавших лекции, постепенно уменьшалось: с 15 000 на первой неделе до менее 4000 на последней неделе. Число студентов, присылавших ответы на тесты, было еще меньше, сократившись до 1000 к концу третьей недели. Однако студенты активно участвовали в форумах, особенно в связи с неясными или ошибочными утверждениями в лекциях и домашних заданиях, а также выражали мнения по другим вопросам. Число студентов, сдавших все задания, включая финальный тест около 400, из них около двух третьих – с отличием. Конечно, это малая доля от числа подписавшихся на курс, но для меня лично – это огромное достижение. Вероятно, это число больше общего числа очных студентов, прослушавших мой курс за все десять лет.

К сожалению, я не смотрел в студенческие форумы в то время, когда курс проводился. Согласно протоколу, наблюдение за форумами возлагается на учебного ассистента. Екатерина Черняк беспокоила меня очень редко, только если речь шла об ошибках в лекциях. Теперь же мне удалось ознакомиться со студенческими форумами и очень впечатлен разнообразием обсуждавшихся проблем. Дальнейшее содержание этого материала – обзор тех из них, которые относятся к курсу.

2. Общий взгляд на студенческие обсуждения курса

Ощутимое множество замечаний получено от тех, кто подписался на курс по ошибке, не ожидая такого объема мыслительной и вычислительной работы. В это число я включаю тех, кто декларировал другую причину, например, ссылался на мой плохой английский, не позволявший им понять смысл задаваемых в тестах вопросов. Некоторые вполне корректно сформулированные, но необычные задания, как например задача оценки уровня точности регрессионного уравнения также иногда рассматривались как результат неадекватного перевода. Некоторые ссылались на использование «коммерческой» системы Матлаб вместо имеющихся бесплатных аналогов, хотя система предлагалась слушателям бесплатно. Более того, специально оговаривалось, что оценки никак не зависят от используемых вычислительных средств.

Много критики шло по поводу ошибок и описок, допущенных в первых лекциях (как правило, не в связи с основным содержанием курса). Описки были найдены и в моем учебнике «Ключевые понятия анализа данных», изданном издательством Шпрингер в 2011 г. Эти замечания я принимаю с благодарностью и прошу извинить за эти описки и ошибки. Я исправил соответствующие части слайдов, так что эти ошибки больше появиться не должны.

Много замечаний и дискуссий, особенно поначалу, было связаны с чисто техническими проблемами, начиная с вопросов о том, как получить Матлаб или учебник и сколько это будет стоить – хотя эти вопросы, включая бесплатность доступа, чётко объяснялись в соответствующих местах – и кончая вопросами правильного форматирования ответов на тесты и домашние задания с тем, чтобы они были приняты системой при сдаче.

Оживлённая дискуссия развернулась относительно использования взаимных оценок студенческих эссе. Курсера допускает три типа контрольных заданий: вопросы, задания по программированию и задания для взаимооценки. Чтобы избавиться от проблем, связанных с спецификацией вычислительных устройств и выбором языка программирования, мы решили сделать все задания по программированию чисто

вычислительными так, чтобы нами проверялась не программа, а результат ее работы. Эта особенность получила признание в дискуссиях. Использование взаимооценки, т.е. оценки каждого эссе специально назначенными студентами – отличительная особенность Курсеры. Чтобы уменьшить субъективность при взаимооценке, мы решили ограничить этот тип заданий заданиями по написанию эссе о ключевых понятиях, рассматриваемых в курсе. При этом каждый студент-оценщик получает указания, что эссе оценивается в зависимости от того, упоминаются ли в нём конкретные одно-два свойства рассматриваемого понятия. Кроме того, общий вклад оценки этого вида в финальную оценку был фиксирован на относительно низком уровне 10%. В Приложении 1 помещены письма с мнениями об этом методе оценки. Можно видеть, как первоначальные худшие ожидания постепенно сменяются более благоприятными высказываниями, признающими полезность этого метода в процессе обучения.

Несколько потоков посланий содержали похвалы и/или благодарности. Уровень и разнообразие положительных отзывов дают мне основания полагать, что в целом мне удалось донести материал до слушателей. Репрезентативная выборка таких писем, в основном, содержащих и информативную часть, размещена в Приложении 3. Несколько студентов, как я понимаю, более высокого уровня выразили желание участвовать в следующих «пробегах» в качестве учебных ассистентов. Это замечено и воспринято с благодарностью.

3. Полезные замечания и сделанные выводы

Чтобы не слишком удлинять этот текст, в данном разделе я остановлюсь только на тех аспектах моих лекций, которые я считаю необходимым подвергнуть изменению в свете полученных от студентов комментариев. Задача - когда и если случится повторение курса, эти аспекты должны быть выправлены.

A. Роль описок и оговорок

У меня, как автора большого количества статей и книг, сложилось впечатление, что опечатки – это неизбежное зло, от которого невозможно избавиться даже после многократной вычитки текста. Более того, при проведении обычных очных занятий случайная ошибка или описка может оказаться полезным инструментом в поднятии активности студентов. Это происходит, когда один-два студента замечают описку или оговорку и начинают задавать вопросы. Они активно включаются в процесс исправления ошибки, что побуждает и других студентов к активизации. Заочный же студент, встретив одну-две ошибки в учебных материалах, начинает подозревать в ошибочности и любое другое непонятное место – вместо того, чтобы попытаться разобраться в сути дела. В этом плане случайные описки в заочном курсе – это реальное зло, замедляющие и усложняющие процесс обучения. Следовательно, я должен избегать их – любыми средствами.

Остаётся только надеяться, что теперь, после того, как учтены все замечания из студенческих писем, в моих слайдах нет – или почти нет – описок, а теперь следует соответственно поправить и соответствующие места в ведеозаписях лекций.

B. Примеры и указания в тестовых заданиях

Многие жалобы были связаны с недостаточной гибкостью в форматах данных, принимаемых Курсерой и СТЕПИКом, что заставляло студентов тратить много лишних усилий при подаче своих ответов на задания. Некоторым не хватало 5 минут, разрешённых для единократной передачи ответов. Многие жаловались, что не понимали задаваемых вопросов, я думаю, поскольку не понимали изучаемый материал, хотя некоторые считали, что смысл потерян при переводе с русского на английский. Думаю, что после многих лет преподавания этого курса в Биркбек колледже Лондонского университета мой английский не настолько плох, чтобы допускать бессмыслицу. Более того, в Лондоне студенты говорили мне, что предпочитают мои лекции (по сравнению с коллегой англичанином), потому что я говорю медленно и громко. Возможно, стоит упомянуть, что почему-то на веб странице Курсеры оказалось невозможно отразить мой международный опыт.

Я думаю, что подобные жалобы можно учесть, если более подробно и с примерами инструктировать студентов о том, как выполнять задания.

C. Математическая корректность и позиционирование курса

Образ «идеального» для данного курса студента в моём сознании определяется моим опытом преподавания в Департаменте компьютерных наук Биркбек колледжа Лондонского университета (2000-2011). Биркбек колледж – это бренд вечернего обучения в Соединенном Королевстве. Соответственно, мои студенты в основном были программисты, поступившие к нам для получения диплома о высшем образовании. Эти студенты обычно были слабо знакомы с математикой; иногда даже попадались такие выпускники английских колледжей, которые никогда не слыхали о понятии производной, основной в математическом анализе. В то же время все или почти все хорошо разбирались в программировании. Мне приходилось приспосабливать свои лекции и семинары таким образом, чтобы студенты могли понимать материал даже и не очень-то понимая в математике. В связи с такой переработкой, исходя из неявного постулата, что серьёзный анализ данных невозможен без вычислений, я представлял и курс, и учебник, как не требующие глубокой математической подготовки, забывая упомянуть о том, что изучение курса невозможно без вычислений. Некоторые записались на курс, не имея должной подготовки, но быстро поняли, что не могут следовать за лекциями и заданиями, и ушли. Таким образом, при описании курса и пре-реквизита к нему мне следует быть более специфичным и явно указать, что от студента ожидается готовность и способность выполнить несколько вычислительных проектов.

Были комментарии и с противоположного конца – о том, что курсу не хватает математической точности. Такого рода комментарии вполне ожидаемы от бывалых студентов, уже прослушавших курс или два по математической статистике. Такой курс состоит из фрагментов, в каждом из которых сначала определяется некая математическая структура, затем устанавливаются ее свойства, сформулированные, а подчас и доказанные, математически. Затем эта структура применяется к некоторому числу задач статистического оценивания или проверки статистических гипотез, после чего даются примеры иллюстрирующие такие применения. Мой курс построен по-другому. Это курс об основных понятиях анализа данных и их использовании в решении реальных задач коррелирования или суммаризации. Математика при этом не более чем инструмент. Например, понятие линейной регрессии y по x в математической статистике вводится через двумерное вероятностное распределение пар (y,x), для которого определяется понятие условного математического ожидания y по x; доказывается, что оно минимизирует квадратичную ошибку; после чего выводится соответствующее разложение безусловной дисперсии y; и устанавливается, что условное математическое ожидание y есть линейная функция от x при условии, что исходное двумерное распределение – Гауссово. Параметр этого распределения, называемый коэффициентом корреляции, явно связан как с наклоном линейной функции, так и квадратичной ошибкой. Эти результаты прилагаются к проверке статистических гипотез о свойствах модели, например, гипотезы о том, что наклон и коэффициент корреляции равны нулю. Всё это очень элегантно, приятная математика, всё аккуратно доказано – никаких вопросов! Можно двигаться к изучению следующей главы. В контексте же анализа данных я стараюсь рассмотреть, что же можно вывести, если не использовать модель двумерного распределения. Я прихожу к понятию коэффициента корреляции, основанному на имеющихся данных (т.е. выборочной оценке теоретического коэффициента корреляции в терминах математико-статистической модели) без какого-либо использования понятия распределения и, тем более, Гауссова распределения. Затем я смотрю на значения этого коэффициента корреляции на конкретных примерах данных и вижу, что эти значения могут быть нулевыми, скрывая имеющуюся корреляцию просто потому, что выборка неоднородна, или наоборот, могут быть сколь угодно большими из-за наличия выбросов, даже если корреляция равна нулю в основной выборке. Примеры показывают справедливость известного афоризма о том, что существуют три уровня лжи: простая ложь, наглая ложь и статистика. Эти проблемы пропускаются в курсе математической статистике – там обучают математике, но не в моём – я обучаю понятиям.

Нижеследующая дискуссия, представленная здесь тремя письмами (перевод мой) и продолженная в Приложении 2, на самом деле как раз касается различий между моим курсом анализа данных и обычными курсами математической статистики, правда, выраженных другими словами.

Ганс Тунгаяя (Hans Tunggajaya):

Я считаю, что этому курсу не хватает математической точности. Я не говорю о настоящей формальной математике, а скорее об уровне курса Нг по Машинному обучению, в котором интуиция и формализмы хорошо сбалансированы. (3 голоса за)

Аноним:

Факт номер один: я участвую в этом курсе и получаю высокие баллы, так что не обижен и не возмущен.

Факт номер два: Я участвовал в большом количестве курсов машинного обучения и статистики как в очном, так и заочном форматах, так что хорошо знаком с предметом.

Моё мнение: Этот курс значительно более неупорядочен и запутан, чем любой из тех курсов, в которых я участвовал ранее. Я по-настоящему не понимаю, почему профессор избегает более формального математического рассмотрения на слайдах и почему он не тратит никакого времени на объяснение формул. В частности, я считаю, что все переменные на слайдах должны быть указаны и объяснены, по крайней мере, упомянуты. Иначе их присутствие бессмысленно (эту фразу примите со «щепоткой соли»). Такое впечатление, что он думает, что принятый им неформальный, нематематический тон поможет простому человеку; на самом же деле такое намеренное избегание математики лишь делает всё еще более запутанным, чем на самом деле. Если кто-то хочет использовать анализ данных, а я думаю, что записавшиеся люди этого действительно хотят, то логично ожидать от такого человека, что он должен быть готов к использованию более серьёзной математики. В противовес взглядам профессора я уверен, что люди будут благодарны за использование более формального подхода для уточнения тех туманных понятий, которые представлены в курсе. Данный предмет – в высшей мере количественный, и его надо представлять соответствующим образом. Если время ограничивает, как это происходит здесь, то лучше иметь, например, 5 понятий, хорошо объяснённых, чем 10 понятий, слегка затронутых. Далее, о еще более серьёзном предмете. Тот факт, что профессор использует свои собственные обозначения и свои нестандартные взгляды для некоторых методов в таком вводном курсе, как этот, следует признать неудачным. В вводном курсе следует использовать общепринятые понятия, чтобы дать людям устойчивую базу, то знание, которое можно легко транслировать в другие курсы и литературу. В то же время более продвинутые курсы могут позволить себе использование нестандартных, так сказать, хипповых интерпретаций. Я глубоко благодарен профессору и его ассистентам за этот курс и их усилия, но никак не могу искренне заявить, что мне нравится этот курс. Мне просто нужно получить ещё один сертификат, который можно упомянуть в моем резюме. П.С. Взаимные оценки просто смех, в самом деле. Написание одного-двух предложений ничему не помогает. Эссе, да, они очень полезны, но коротенькие ответы на них – ну никак. Моё прямое, честное мнение.

(4 голоса за)

Вилко Дийкхвис (Wilko Dijkhuis):

Вах, этот курс начинает мне очень нравиться.

Однако как преподаватель элементарной статистики я должен выразить одно сомнение. Описание курса намекает, что его содержание доступно новичкам в статистике. Некоторые могут подумать, что это мягкое, не математическое, введение. Тем, кто попал на этот курс по подобной причине, я, профессиональный преподаватель мягкого введения в статистику, даю срочный совет – и не надейтесь.

Это – мастер-класс для тех, кто уже знаком с предметом, но хочет получить другой – с точки зрения анализа данных – взгляд на основания данной области.

Да, это – не математический курс. Но здесь это означает две вещи:

1) Ведущую роль практики анализа реальных данных. Цель – не представить новые интересные математические доказательства, а использовать существующую математику, чтобы делать интересный анализ данных.

2) Необходимая математика грубо обозначена, но не разработана формально (к ужасу настоящих математиков; в этом смысле – подход «анти-математический», а не «не математический»).

П.С. Для тех, кто хочет именно мягкое введение, я могу рекомендовать:
http://exexstats.tumblr.com/post/58597571228/tips-for-www-learning-statistics-intro-to (14 голосов за).

В приложении 2 помещена копия письма Анонима и последующая дискуссия, опровергающая все основные утверждения этого письма и объясняющие, как следует воспринимать материал курса.

Вывод: мне следует существенно улучшить объяснение основных задач ключевого анализа данных, а также и моего курса. Как сформулировал Вилко Дийкхвис, цель – не дать интересные математические доказательства, а использовать математику для проведения интересного анализа данных. При этом я должен ясно указывать, что определённая математическая подготовка необходима. Вероятно, знание компьютеров – тоже.

D. Нетрадиционные представления

Поскольку модели анализа данных используют сами значения признаков, а не их вероятностные распределения, как классическая статистика, это может внести необычные интерпретации рассматриваемых понятий. Мне нравится показывать такие интерпретации, особенно когда это математически несложно, как, например, когда речь идёт о смысле такого центрального понятия как коэффициент хи-квадрат Пирсона: я показываю, что это не просто критерий для выявления статистической независимости, как это обычно утверждается, но и коэффициент статистической связи между значениями признаков. Однако менее очевидные следствия остаются незамеченными. Одно из них оказалось предметом обсуждения на студенческом форуме.

Речь идёт о многомерном анализе номинальных признаков, предмете одного из наших вопросов. Мне казалось, что я ясно указал способ их количественной обработки путём перевода номинальных категорий в 1/0 бинарные признаки. Этот совет, однако, оказался пропущенным подавляющим большинством студентов, вероятно потому, что был дан в связи с анализом конкретных данных, а не в общем виде. Посмотрев в интернет, они также не обнаружили никакого совета. Возможно, мне стоит ввести еще один раздел, в котором бы объяснялось, как применять рассматриваемые методы к номинальным и смешанным номинально-количественным данным.

Конечно, задача объяснения и обоснования других необычных утверждений является приоритетной. Это особенно важно, поскольку многие, встретив новое для них или неясное понятие, используют Гугл или другой поисковик, чтобы посмотреть, как это понятие трактуется на других страницах. К счастью, глубокая связь между анализом главных компонент и сингулярным разложением матриц, лежащая в основе моего подхода к этому методу, недавно стала отражаться и на других страницах. Однако другие полезные инструменты, как например, коэффициент Кетле для измерения связи между категориями, пока ещё не сделались популярными, что может обескуражить некоторых. В этой связи я хотел бы сделать так, чтобы студенты обращали больше внимания на лекционные материалы прежде, чем искать где-то на стороне. Например, я привёл примеры неправильных значений коэффициента корреляции, возникших из-за неоднородности или наличия выбросов, и рассматривал их как примеры ситуаций, в которых статистика – самый большой источник «лжи». Однако, студент, объяснявший этот парадокс другому студенту, предпочёл обратиться за примерами на другой веб-сайт вместо того, чтоб использовать эти.

E. Юмор

Многие студенты отметили, что мои лекции было интересно слушать из-за того, что в них постоянно возникали юмористические элементы.

Дипанжан Саркар (Dipanjan Sarkar)

Я согласен – лекции короткие, но они очень информативны, и профессор делает их не слишком занудными, добавляя крупицы юмора в разных местах. Интересные темы обещаны в остающиеся недели курса.

Дэн Нортруп (Dan Northrup), фрагмент персонального послания от 7 июля

Уважаемый Д-р Миркин,

Мне по-настоящему понравился Ваш курс и особенно то, что Вы постоянно шутите, читая лекции. Пожалуйста, продолжайте читать подобные курсы; они очень хорошо воспринимаются Вашими студентами.

Такая реакция оказалась для меня совершенно неожиданной, так как я старался читать эти лекции в формальной манере. Однако, просмотрев видеозапись лекций, я действительно обнаружил немало забавных моментов, как, например, когда я привожу примеры степенного закона, упоминая веб-сайты Гугла и Амазона как наиболее популярные, а мой собственный – как наименее популярный. В другой раз я использовал броское название «столкновение цивилизаций», рассказывая о такой малости, с точки зрения вопросов жизни и смерти, как разница в представлениях о сущности анализа данных в сообществах специалистов по статистике и по майнингу данных. Возможно, что моё острое восприятие различий в масштабах действительно ведёт иногда к подчёркиванию оных в форме некоего гротеска.

ВШЭ провела анкетирование студентов до и после курса (К.А. Кузьминых). Основные итоги представлены в Дополнении как два файла ответов на вопросы о том, что понравилось и что не понравилось в курсе. Там же – референтные графы, визуализирующие эти файлы, построенные Е.Черняк и мною на основе ранее предложенного нами метода анализа текстов. Эти результаты подтверждают и детализируют мнения, высказанные в данном отчете.

4. Заключение

В целом, несмотря на накопившуюся усталость от тяжёлого труда по подготовке слайдов и продумыванию их содержания, я испытываю чувство глубокого удовлетворения курсом и намерен улучшить его в направлениях, упомянутых выше. Конкретно, мне надо обратить специальное внимание на то, чтобы:

- Свести на нет возможные описки и ошибки;
- Дать конкретные примеры и указания по форматированию ответов на задания;
- Адекватно позиционировать курс;
- Лучше объяснить, какие части курса – математические, а какие - нет;
- Лучше определять ключевые элементы анализа данных и цели данного курса;
- Представить более широкий список нестандартных подходов, вытекающих из нацеленности курса на данные, а не распределения;
- Продолжить юмористические тенденции.

Кроме того, мне следует посмотреть, нельзя ли сделать какие-то части подлиннее. Возможно, я слишком серьёзно относился к указанию, чтобы отдельные части не превышали 10 минут длительности.

Я подумываю также о разработке математического обеспечения, покрывающего рассматриваемые методы не только так, чтобы получать результаты от применения их к конкретным данным, но и так, чтобы студент мог видеть и промежуточные результаты, получаемые на отдельных шагах применяемого алгоритма.

Конечно, некая линия обратной связи, позволяющая студентам контактировать со мной в процессе занятий, была бы полезной в разрешении текущих вопросов. Например, это мог бы быть форум для связи лично со мной, на котором бы я, скажем, раз в неделю отвечал бы на 10-15 вопросов. Вообще, следует подумать над тем, как усилить линию оценки заданий и контактов со студентами; возможно, существует такой способ организации добровольных помощников учебного ассистента, при котором они выполняют часть этой работы. Хотя пока что я не очень оптимистичен по вопросу установления обратной связи со студентами, так как, судя по комментариям студентов, другие курсы ещё хуже справляются с этим.

Следует также отметить, что обязанности учебного ассистента в течение 9-10 недель курса (8 недель обучения с последующим окончательным тестом и его проверкой) при тысячах и десятках тысяч слушателей требуют по настоящему интенсивной работы в течение двух месяцев без какого-либо перерыва. Следует подумать о том, как структурировать эти обязанности так, чтобы можно было использовать участие не одного, а нескольких ассистентов. Тогда, вероятно, можно будет использовать помощь тех из бывших слушателей, которые выразили такое желание.

Supplement. Survey of student opinions of the class summarized in two reference graphs

HSE undertook a couple of surveys among participants to the class. Here results of the last, after-class, survey are presented in the format of two collections. Collection 1 is a set of 215 unedited answers to the question of what the respondents did not like most in the class. Collection 2 is a set of 237 unedited answers to a symmetric question of what in the class the respondents did like most. Both collections are presented further on in this appendix.

Although sometimes informative, sometimes instructive, the hundreds of opinions are rather diverse. And it is really difficult to make any sense of them as a whole, the more so that one may find some features on both collections, "the bad and the good" ones, whereas a few opinions seem not related in here at all.

To summarize and visualize the opinions, the assistant instructor in the course Ekaterina Chernyak and myself applied a technique that we had proposed earlier in our paper B. Mirkin, E. Chernyak, O. Chugunova (2012) "The method of annotated suffix tree for evaluation of relevance between a phrase and a text", Business Informatics, n. 3, 31-41 (in Russian).

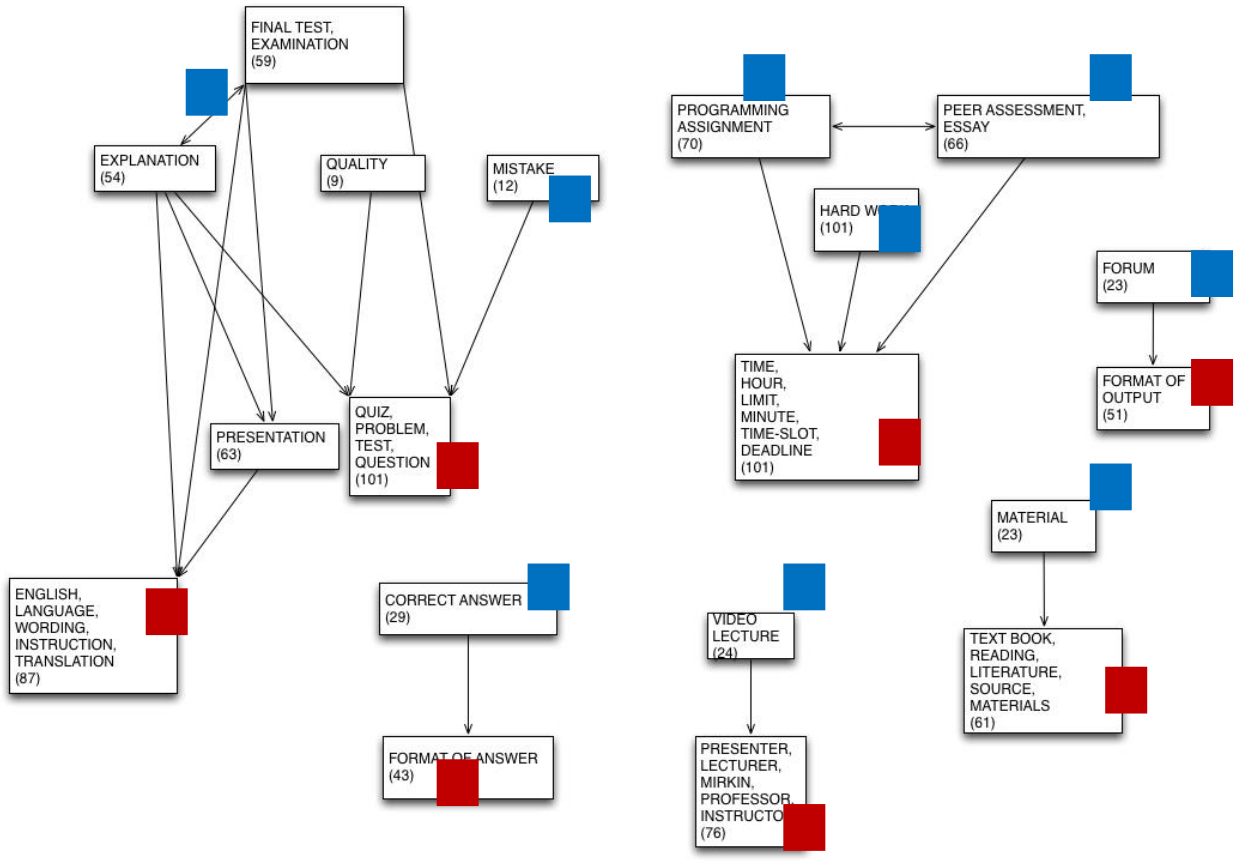

Figure 1. Reference graph for key-phrases in Collection 1 of opinions about not likeable features of the class. Each node is accompanied with a figure expressing the number of opinions in Collection 1 containing one or more key phrases assigned to the node. Input nodes are tagged by blue boxes and output nodes by red boxes.

This technique uses a set of key-phrases that are relevant to (some of) texts from a collection under consideration. It builds a directed graph which is referred to here as a reference graph. Its nodes are annotated by the key-phrases one-by-one, and an arrow goes from node A to node B if whenever a text contains the key phrase corresponding to A, in most cases it contains the key phrase corresponding to B as well. In this sense, the key-phrase A refers to the key phrase B or B is a reference for A, which explains the name "reference graph". We assume that two key-phrases are synonymous with respect to the text collection under consideration if they largely refer to similar key-phrases. The synonymy relation allows us to aggregate synonymous key phrases in one node of an aggregate reference graph. The figure assigned in a node box on Fig.1 and Fig.2 gives the number of opinions in which at least one of the corresponding key phrases occurs.

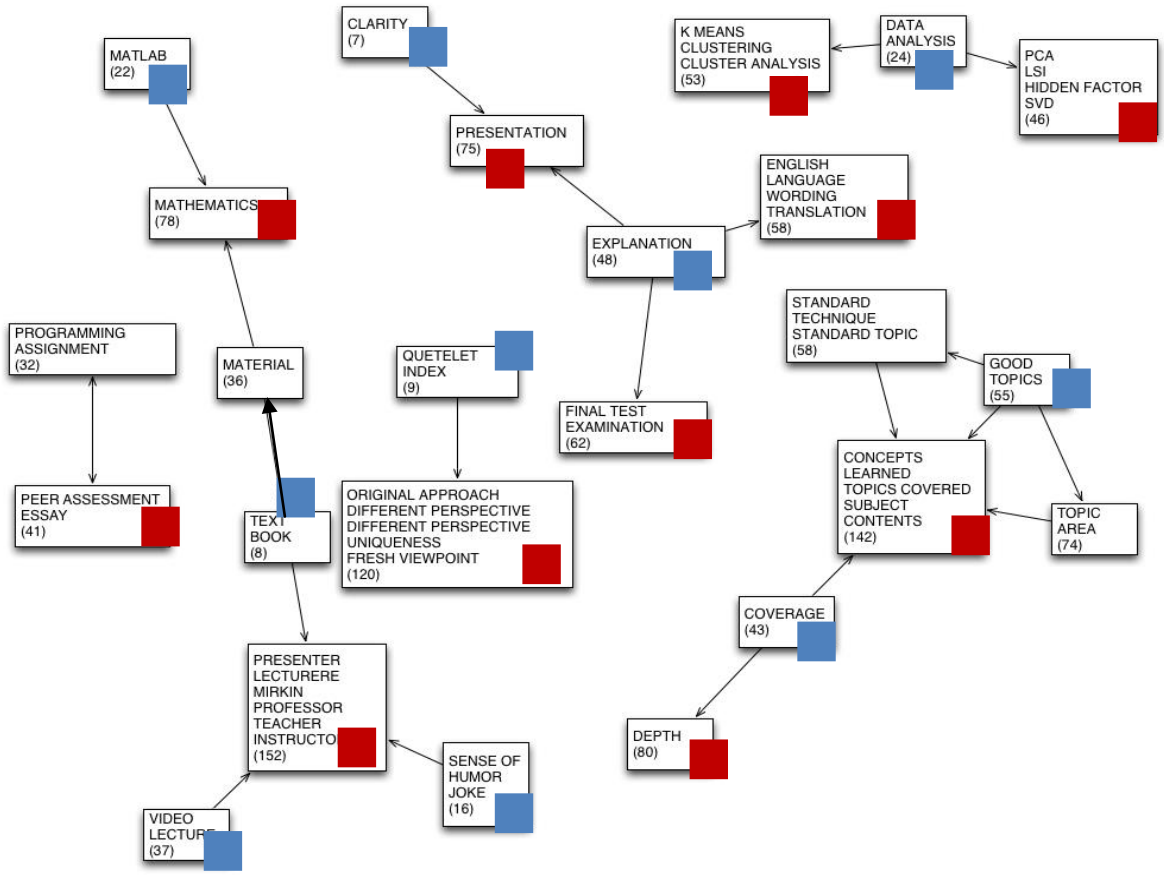

Figure 2. Reference graph for key-phrases in Collection 1 of opinions about good features of the class. Each node is accompanied with a figure expressing the number of opinions in Collection 1 containing one or more key phrases assigned to the node. Input nodes are tagged by blue boxes and output nodes by red boxes.

A reference graph for Collection 1 ("do not like") is on Fig. 1 and for Collection 2 ("do like") on Fig. 2. One can see that the mainstream bad features, those in terminal nodes, are those tagged red. First of all, they relate to the way of presenting things. Three red tagged boxes relate to comments on: (a) language (87 comments), (b) presenter (76 comments), and (c) teaching materials (61). The other red tagged boxes relate to tests: first of all, the way the Coursera accepted answers, quite demanding in terms of formatting indeed (two boxes point to this, 43 on format and 51 on answer comments), and then, to tests themselves and time limits (101 comments on each of the two). The reasons for the latter are three blue-tagged boxes: programming assignments, peer assessments and hard work needed. According to this graph, tests are flawed because of

poor explanation/presentation. Some (12 comments) pointed to mistakes which did occur more than once, to my shame.

A similar graph, but for comments of accolade for the class, is on Figure 2. One can see more nodes and greater variation in their frequencies than on Figure 1. Perhaps this is because the positive comments on average are as twice shorter as those negative ones. The most popular subjects within the terminal, red-tagged, nodes are about: (a) presenter (152)[2], (b) originality of the approach (120), and concepts learned (142).

A subset of red-tag nodes praise features that have been condemned on Figure 1: presentation (75), language (56), final test (62), essay/peer-assessment (41); they are supplemented by depth (80) and mathematics (78).

Two more final nodes refer to specific subjects learned – clustering (53) and principal component analysis (46). One more subject - specific node (Quetelet index, 8) referring to the original approach/fresh viewpoint (120) of which it is a representative example. On the level of input nodes, one can note clarity (7) and explanation (48) at one branch, coverage (43) and good topics (55), on another. Among the rarer inputs one may notice Matlab (22) and sense of humor (18).

Altogether these two reference graphs represent a rather informative summary of the opinions of both criticism and credit expressed by the respondents.

[2] The number of comments on the presenter in "bad" graph on Figure 1 is 76 only, so that I can see no need in amending my looks.

Collection 1.

What did not you like most in the class:

1. Some of the tasks formulated incorrectly
2. Words of the instructor were sometimes unintelligible.
3. Wish course went into more careful detail of PCA. I also wish the course would make a little more effort to bring the various techniques together..
4. We may have needed a few more examples in videos. Problems were very difficult, which I do not mind, but we may have needed a few clues or more background to assist preparation for quizzes.
5. Was not able to fully participate due to time constraints so can't fully comment.
6. video lectures, which were very generic in nature
7. Unclear questions on quizzes.
8. unclear job descriptions
9. Too much to delineate.
10. Too fast-paced. The language is too technical.
11. To urge my mind to think.
12. To much focus on the theory / calculations and not many working examples to show the theory in practice. Ex: PCA is VERY difficult to understand - so a simple working example all the way to the conclusion would made it much easier to understand.
13. Though the course was easy, it was slight harder for me to grasp. The course got progressively harder after PCA. The chapter of K-means was all mathematics and the coding portion of it was difficult for me to grasp. I did earlier programming question, but later it was getting too hard for me. :(
14. This isn't particular to this course, this is a general trend in scientific programming: very poor code quality (from a programming standpoint). By poor I mean: bad formatting, poorly selected variables and function names. Programs lack coherent structure and proper documentation.
15. This is a subject that I already had a fairly good grasp of but the lectures were very difficult to follow. If I hadn't already known a lot of the material I would have been completely lost. The presenter really needs to do a better job of planning the course and maybe practice the lectures some more.
16. There wasn't much feedback on the tests, such as explaining why a certain answer was wrong. Also, while I could follow and learn the 'math' needed to apply a certain analysis method I often asked myself what the numbers I was calculating really meant and why would they be useful in extracting information from data. The Quetelet Indices make a good example of that
17. there no explanation of Matlab, further in the presentation you say that any basis of statistics and matlab was necessary, but in the course and the development was necessary high knowledge.
18. The ways to explain the materials were too complex.
19. The ways the tests were made, and the huge amount of time spent not on the core of the course but on the format of the answers
20. The way things were graded and peer assessments
21. The way the programming assignments don't seem to be related to what was covered in the lecturers. The lectures weren't very clear and the notations don't seem to follow the norm, for example in PCA, the symbols used are [Z, mu, c] compared to [u,d,v] and the lecturer didn't explain why he uses that notation. I was rather disappointed with the course.

22. The way the ideas was explained
23. The typos in the pdfs
24. The teaching
25. The slides were to dense and I was not able to replicate properly the examples given in them
26. The seeming expectation to participate in group activities and the unavailability of R for exercises
27. The quality control on the quizzes and assignments. Many students lost many hours trying to figure out why they were not getting the correct answers when the error was in the question itself.
28. The programming exercises were not very challenging.
29. The programming assignment tools seems to be very stiff, only 5 mins for completing the exercise. Please explore a more flexible framework for programming assignments.
30. The professor and the quality of his explanations
31. The personal agenda of the presenter. It was not teaching the material, but rather involving unwitting students in a personal crusade to support the presenters view if the world against the academic establishment.
32. The peer-assessed assignments. Also, the 5-minute time-slots for submitting the programming assignments on the external platform was too tight. I feel that 10 minutes would be better.
33. The peer review
34. The peer assessments were useless - asking non-experts in a topic whether an answer 'matches what they think it should be' is far too subjective.
35. The peer assessments allowed a lot of nit-picking for minor issues.
36. The peer assessment s, where the question were not very clearly framed, as what was e expected, little confusing. Final test, errors in the test questions.
37. The peer assessment process and instructions
38. The pace of going thru PCA is too quick. I did studied matrix algebra in University, however that was 8-9 years ago which I feel quite challenging to apply this forgotten knowledge on PCA. Hard to imagine other students which may have no exposure to matrix algebra before. Recommend extend PCA for 1 more week on lectures and exercises. Also does not quite enjoy using Matlab.
39. The only trouble with the course was sometimes difficulty with understanding the quiz questions, but it was helpful later to have a practice set with correct answers for practicing the coding questions.
40. the load was not uniform
41. The lectures were boring and not engaging at all.
42. The lecturer
43. The lack of intuition behind the methods
44. The lack of detailed information and the link with applying the methods to data. I think exercises should be more in depth to allow people not familiar with Matlab to complete the programming assignments
45. The instructors presentation skills
46. The instructor is not well versed in English which requires a great deal of effort to decipher. Most test questions were badly formulated (prose) hence hard to understand. The noticeable bugs in certain tests/assignments almost made me quit the whole course.
47. The instructions were not always clear and I was sometimes lucky to have had discussion forums to help clarify.

48. The in-video quizzes were disconnected and the correct answers were not available for checking them after watching the videos.
49. The hurried nature of the presentation of the content
50. The graders on quizzes and particularly the final were horrible, poorly implemented, in many instances incorrect altogether.
51. the git part
52. the gap between easy intro and too challenging midparts of each module. One minor thing; Maybe you could reduce the length of the 'HSE' standard intro?
53. The download data for the assignments had to undergo some transformation before I could process it in Matlab (or other programs). This meant that I always ran out of time during my first attempt. This was not really a problem since you had a fair number of attempts but it also meant that you could not apply what you learned directly but had to 'clean' and 'transform' the data first which is probably an attempt to make it more real life.
54. The discussion of principal components analysis. It took me an hour just to *copy down* all the equations that were presented in a 17-minute lecture--which means they went by MUCH too fast--and the math was much too advanced for the background that was asked of class participants.
55. The difficulty!! But this is a problem of mine and not of the course!!
56. The course was VERY frustrating in that the questions were poorly constructed. What should have taken me 15-minutes to do took 4-6 hours because I was reading the forums trying to figure out what the question was really asking. This greatly interfered with my learning. I understand that the prof's first language in not English and I am very comfortable working with broken English. However, the slides and especially the test/assignment questions need to be proofread by a native English speaker so that they make sense. The major challenge in the course was not the course material but figuring out the broken English. It is obvious that the prof is not comfortable speaking English and I think that it is great that he taught the course in English. However, the written material slides/test questions have to be improved. The supplied textbook was well written. I went and purchased a copy of that.
57. The course left a lot to be desired in terms of organisation, management and support. It was rushed, poorly administered and with little support provided by the only (and yet excellent Teaching Assistant). The quiz questions and the final exams were riddled with errors, imprecisions, issues problems and omissions that made me lose faith in the quality of the overall assessment process. I found myself having to interpret what Prof. Mirking could have possibly meant with a question and what wrong answer was closest to his possible idea of a correct one. The final was just appallingly full of mistakes, especially on the KNN questions. This completely spoiled one of the best courses I attended so far (and I mean not courses in statistics but all university and professional courses I ever attended). I'm frustrated and saddened that lack of care and diligence amounted in such a poor outcome and I had to keep motivating myself through the realisation that the content was brilliant so I had to see through the repeating, grave and frankly not acceptable shortcomings of the course.
58. the consult for the future
59. the confusion explanations
60. The assesment patterns. Highly unprofessional formatting errors in the assessments. However the quality of questions asked was good.
61. The accent but still I'm able to live with it.

62. The 50% penalty for deadlines was unreasonable and based on an assumption that people were able to commit unlimited free time to the programme
63. that the queries where very vague and needed often interpretation, that the formating of the expected input was often unclear, that there where faults (identified by the students) in the expected answers especially in the final test
64. That I missed the first 3 weeks, lost points to deadlines, used the few late days we had (should have mentioned that in what to improve: give more!) and worked like crazy in the beginning to catch-up. Regret getting my lowest grade, by far, out of 13 completed courses on Coursera: 83.3 But it was worth it. This course was important to me: complements nicely other courses I have taken.
65. That I couldn't complete it due to problems with my computer and internet access. Not your problem, if course.
66. Textbook 'Core Concepts in Data Analysis and Summarization, Corellation and Visualization' needs to be edited for clarity and mistakes.
67. Text book was useless. Lack of depth of individual topics. Assignment questions were unclear, perhaps it was the wording or the language uses. Lack of support from academic team. There was just one person answering queries and she was unable to do a good job.
68. Test online
69. technicality
70. technical iisues in tests, hard deadlines for programming assesments
71. Teacher doesn?t explain deeply the matter of the course
72. Stepic, automatic grader. Difficult to please and hard to get reasonable feedback. Thank god for Elena and forums.
73. Spending more time trying to format my assignment submission properly than learning about data analysis.
74. sometimes the videos went too quickly over a topic. Even with repeated watching I was not clear. I eventually bought the book, but was too far behind to catch up. Will try again next offering.
75. sometimes the subtitles state 'inaudable ...' whilst i can perfectly understand what's being said. Doesn't do the presenter right.
76. Sometimes the questions of the assignments, especially those included in the videos, were not very clear.
77. Sometimes the automatic grading seemed to be a little bit flaky. Also, it would be good to have a consistent format for input of answers (for example, always separate by ONLY 1 space).
78. Sometimes content was difficult to understand. And there was so much math.
79. Sometime difficult to get credit for correct answer due to formatting issues.
80. Some statements in the quizzes and the final exam were not very clear by an English language point of view. Try to set question in the most undestandable (without possible misundestanding) way.
81. Some of the peer assessments. But this is due to my poor English writing skills.
82. Some of the excercise acommplished were not completily understand though.
83. Some complex topics (PCA, cluster analysis) not much developed in depth
84. Slow, boring presentation Too many peer reviewed essays
85. Should have given R as the core software since it is more open
86. References to Matlab instead of R

87. Quizzes and peer assignment: the expected ouput was not clear and this lead to a lot of waste of time (entering the results in the wrong format...)
88. Quite some errors on assignment, e.g.,format. I understand this is the first time offer of the class. I believe it will become better later on.
89. Quetelet index
90. Quality of English in problem formulations, somewhat unclear
91. pronounciation
92. Programming Assignments for 5 mn
93. Programming Assignments difficulty
94. programming assignment are time limited and very difficult if not used the recommended statistic program. I have received only 2 out of 12.
95. Professor had a convoluted way of explaining material. He spoke as if we were already experts on the topic instead of newcomers. Awful, awful course. You should feel bad.
96. Problems with programming assignment. Too much time on histograms and easy things.
97. Presenter's English
98. Poorly worded questions. Tests with incorrect options as answers. No basic computer programs available for the course. If the course has a sequel I would highly recommend this. For example, there is no code easily available for intelligent k-means. If a follow-up course has this topic it would be better to have students explore data sets with code provided rather than having them do programming exercises.
99. Poorly worded assignments. Because the lecturer used his own specific calculation methods (sometimes hard to reconstruct!), all standard methods for e.g. PCA were useless.
100. Peer review could be improved
101. Peer graded projects.
102. Peer assignments. The same kind of knowledge could have been tested through quizes and it would save time for participants. Also, the timing in programming assignments was too short. I would have been near the end and the time was just elapsed and I had to everything again.
103. Peer assignments
104. Peer assignments
105. Peer assignments
106. Peer assignment. The deadline of task
107. Peer assignm
108. Peer assignements
109. Peer assessments were very difficult to perform and grade the way they were presented. Also, most examples were from the same data set. Would have liked to have seen more real-life examples.
110. Peer assessments seemed like a waste of time.
111. Peer Assessments seemed least related to the video lecture material.
112. peer assessments -----> useless
113. Peer Assessments
114. Peer Assessments
115. peer assessments

116. Peer assessment.
117. Peer assessment seemed a little bit random. Sometimes the tests came up with questions out of the blue.
118. peer assessment
119. Peer assesments require free writing where individual research and sometimes judgement shapes the answers, but the evaluation criteria is totally rigid. In the case of programming assignments, the lack of feedback was very frustrating when trying to correct a problem.
120. peer assesment
121. PCA
122. Pace was too fast and not very much in depth, more applications on data analysis problems must be done to explain the theory learnt in detail
123. Pace
124. Organization and depth to which covered. Could have been more application focused
125. Nothing, I just didn't have the time to do the complete course.
126. Nothing in special
127. Nothing
128. nothing
129. Notation been used in presentation slides and peer assessments.
130. not being able to see the final score and feedback immediately after the exam, confusing questions, having to guess the answer format
131. not being able to access the data set for the first assignment, not getting help to do so. Then requirement to learn Matlab or R on the fly was just too much.
132. Not applicable.
133. Non-standard data analysis language. Confusing lectures and readings. The quizzes were very oddly worded. Instructions for many of the quizzes were confusing and not clear.
134. No complaints.
135. no
136. new for me
137. n/a
138. My being unable to keep track of steps in a simple mathematical equation.
139. My apologies to the lecturer, who is probably very skilled and knowledgeable. But the lecture videos (I got through the first 5-10 videos) were very boring. I have taken several other data science courses on Coursera and none have been this bad. Again I apologize to the lecturer for my comments - since he is very likely a brilliant scientist, but not a skilled communicator (in my opinion).
140. Mistakes, TA reaction
141. Mistakes in tests
142. Maybe my English is poor, I found that I have some difficult when listen or read book. I think that's my problem. I just hope this book (Core Concepts in Data Analysis) can be published in Chinese language version in future. Although I almost have read it all. But I can recommend it to my friends.
143. May be examples in Matlab, it's a good system for sure, but looks like R could be a better choice
144. Maths

145. maths
146. MATHLAB
147. Math got intense. Hard to work thru, but I did learn from some of the math.
148. listening to russian-english pronunciation wasn't easy on the ear
149. Language problems and peer assessments.
150. lack of time , in some parts lack of explanations of topic
151. lack of presenation skills, video lectures hard to follow, Prof. Mirkin explains things 'his own way' (i.e. other literature does not cover the topic, uses different names or explains it differently altogether making it impossible to use other sources for reference), peer assessments completely useless mostly due to fuzzy assignment description
152. I would have found useful more guidance on Matlab.
153. It would have been nice to do more exercises and practical applications.
154. It would be good to always give a test sample, and more feedback in programs not working
155. It was too idiosyncratic: it had different opinions to other courses, presented different techniques to other courses, but did not justify those techniques / opinions. Then it tested if we had the same opinion. For example one the peer assessments said 'Given that the data features are all nominal; does it make any sense to apply PCA to the data? Answer the question in your own words.' This is a hard question - see http://stats.stackexchange.com/questions/5774/can-principal-component-analysis-be-applied-to-datasets-containing-a-mix-of-cont So I would argue you could answer it either way because although techniques do exist for this it's not clear if they should be regarded as PCA or not. But the answer was reduced to yes or no rather than considering the discussion! So the level was hard, but then the answer was too opinionated. This happened a few times on the course. Luckily, the discussion forums were good so most times I was able to resolve these issues by referring to the discussion forum.
156. it should have had more assignments/quizzes or a project, because we learn by doing...
157. it is somewhat complicated in simple things
158. It felt as we were left on our own, without guidance from the staff (at least in comparison with other Coursera courses I've taken). It is promoted as a beginners' course, but there are so many aspects that are at intermediate level that are confusing for beginners. There were mistakes on slides that were also causing confusion. There were unexplained things on the slides i.e. in the video lectures that were confusing. I gave up during the second week. It was just taking too much of my time without adding value.
159. Innovative approaches
160. In a lot of cases questions were not clear at first sight. It took a lot of puzzling, reading the forums, and guessing, to understand the question. I think a course should be about thinking about the contents, and not so much about thinking what a question means.
161. important concepts lost in translation (literally!)
162. I'm not fond of equitations, empirical formula etc. It looks always complicated an I would need a very detailed explanation of each factor. I need the understanding for my work in research and so I'll try to learn and collect knowledge bit by bit by using different kinds of lectures. This was another try and I could not follow the explanations about the equitations as I was missing some simple explanations about each factor in it.

163. I'll reiterate what I've already said: this course was let-down by the instructor's poor presentation skills and messy slides. Additionally, the instructor's speaking pace was very slow and laboured, as if it was difficult for him to communicate his ideas in English. If these are improved, I will certainly take this class again in the future. I did not complete this class. The only reason that I did not unenroll is that I wanted an opportunity to complete the post-course survey -- this survey -- because I'm hoping that my feedback will lead to improvements in this class. I want to take this class again, but when the kinks have been ironed out. This could be a great class with improvements in presentation.
164. I would prefere a real life project covered as course project.
165. I would have liked to have a lot more working examples (like in the textbook) during the video lectures
166. I wanted to try to save the homework assignments and try them once I've finished other courses I've paid for. I want more practice on these concepts from the course, so I'll probably take it again once I finish the others (and brush up on some of the math that drives these concepts) I also think the course should list a math prerequisite, because some things were hard to follow. I'm still replaying the eigenvector video trying to get those down solidly before I move to the PCA concepts.
167. I try and avoid any course with peer assessments. I think of it as the blind leading the blind, and a waste of time.
168. I think the quizes/exams were more difficult then they should have been. Not because of the material, but because of the finickiness of the grader.
169. i think some things need more explanations. All these things can be found in course discussions. Some pictures are pour. f.e. i hardly understand those in Week 6 presentation - p. 28, p. 58
170. I should appreciate much more examples and variety of them to use the techniques we learnt. So did not like the little amount of examples. Other thing I should appreciate is much more feedback on the exercises and homework.
171. I really cannot tell if this course is going to be helpful to anyone. Lots of notations and examples are meaningless outside of academe. The first five weeks classes need to be changed. The hidden variables and clustering parts deserve an extension with more practical examples. And the format of questions really needs a re-work.
172. I like all the contents
173. I had done well on the first three quizzes, but I gave up on Quiz #4 because I wasn't sure what the questions were asking. I had to make a decision about investing more time to do the quizzes, and I decided to wait for the next offering of the course. I'm assuming that the quizzes will be improved.
174. I felt the lack of feedback for some of the programming work frustrating. By the time I reached the final examination I think the sense of uncertainty about how results were assessed affected my commitment to doing my best.
175. I dislike the peer essays. I would like that the topics would be covered more in-depth
176. I did not have the time to give the course the time needed. However, I hope to revisit the lectures as time permits and perhaps re-take the course for a certificate.
177. I can not remember anything. I had less time that I believed, so, probably deadlines could be more flexible

178. I believe the peer assessments did not add much in the format presented. There was also some confusion on interpretation in some of the quizzes which I believe they were due to language. I compliment the hard work, time, and dedication the TA put into the material and forums. I am sure these material can only improve in the future.
179. I am very busy and I could not keep up with the deadlines due to work, and my other studies at my own university.
180. I actually would have preferred for there to have been some instruction in R. I feel this is more applicable to working professionals.
181. for a person who is not familiar with topic, math was too hard to grasp without reffering to other (doing additiona research on Internet) sources. Finally it become like course materials shows the path what to study, but to learn materials I had to find other sources.
182. Flow of material could be improved
183. figuring out in which format to put the answer (percentage or not, how many decimals, etc.); it distracted from the real purpose of the quizzes, however towards the end of the course the explanation with the questions improved.
184. fast paced
185. Explanations are not clear enough, I feel that there is a language barrier that makes it harder to learn. I was frustrated by that and it made me work more hours, and look for explanations in the internet, and yet I'm not satisfied by the level of knowledge I've achieved
186. Explanation of knowledge
187. explanation about real applications
188. Essays
189. Error in quizzes. Lack of clarity in quizzes. Grader for programming assignment quirky and unintegrated. Peer assignments.
190. doing peer assessment
191. does not apply
192. difficulty of presentation of the content
193. didn't understand the questions sometimes; I sometimes missed the deeper understanding of the material; the video presentation could be less power point stylish
194. didactics
195. Decision Trees not explained well. Latent Semantic Analysis could use a programming assignment.
196. deadlines
197. Convenience
198. Compared to other Coursera classes I've taken, the course was less polished. The assignments were poorly written or translated. They lacked important details or were sometimes incorrect making them very difficult to complete. The TAs were very responsive, which I appreciated. I felt that the video lectures lacked the depth and richness of explanation that I've seen in some of the other coursera classes. I'm glad I took the class for an overview, but I don't feel well enough versed in the material yet. I wish, however, there were more classes covering this material. Perhaps a single course devoted to PCA, for example.
199. communication in certain situations

200. code exclusively provided for MATLAB
201. But with English being a second language for the professor or the teaching staff, the entire course lacked Clarity, couldnt get the right point across, terrible typos, the intent of questions is baseless (test on the content.. not on tricking the students... we arent highschool kids that are taking the course). the Questions in quizzes or programming assignments or the peer assignments were very ambiguous, and didnt make any sense until we got through the discussion forums. May be try giving some examples and really explain what you are looking for. Over all due to the communication skills both verbal and written, this course was extremely stressful. I wouldnt recommend any one to take this course again from this specific teaching staff.
202. Boring details in trivial issues.
203. Basic part
204. Bad user experience
205. Assigments grading.
206. Assessments not working properly.
207. Answers for quizzes and programming assignments not posted
208. ambugity on many explanation and quiz questions
209. All the mistakes that made the tests a bit frustrating
210. A bit computer science oriented. But I still learned a lot.
211. The course materials (pdfs) often lacked clarity, coherence and fluidity. 2) There was a wide divergence between details mentioned in course materials and assessment questions, leaving too much room to speculation. A good course is one where the material is lucidly taught and grasped, pre-requisites clearly explained and assessments made difficult but answerable within the framework of the course materials. This was not the case with this course. If people have to resort to guesswork or discussing answers in forums, then obviously there are gaps in the teaching process. 3) Some questions in the peer assessments were open ended, but the marking criteria too rigid. Hence some good answers received zero marks. 4) The programming assessment software initially did not seem agnostic to the computing platform (e.g. MATLAB or R or Python) used, but it did improve a bit as the course progressed. 5) Overall communication left much to be desired.
212. Short videos, less explanation 2. No guide to assignments
213. It is absolutely frustrating if the answer is correct but is not marked as correct due to technical problems or if problems do not state in which format answers should be given. However, this was a permanent issue in this course. 2. Throughout the course I had the feeling that I wasn't taught a general approach to cluster analysis and principal component analysis but merely Prof. Mirkin's interpretation of these statistical concepts.
214. Need better explanation for solution to quizzes and peer assignment (2) 5 minutes limit for programming is too short. Programming assignment is very importance. Suggest a 1 week limit

215. - No general idea, messy content - No real examples of data analysis, just artificial datasets - Poor programming assignments - not interesting, buggy, stupid time limit, reliance on expensive commercial software despite the presence of better free alternatives - Peer assessment should be excluded - too much hassle, too little output In general, the course provided no help in doing real data analysis.

Collection 2

Answers to the question: What did you like most.

1. you make mathematics clear for me.
2. Working in R environment
3. Videos
4. Video presentation. Thanks for the course.
5. video lectures, discussion forum, help from the course staff
6. video lectures look appealing optically at fist glance, grading system for programming assignment is well thought out
7. Video lectures and book
8. Video lectures
9. Video content was very useful. I love it.
10. Very original approach to data analysis. Prof Mirkin's approach is grounded in geometry and sound mathematics and yet very finely tuned for practical usage. Overall, one of the best combination of theory and practice
11. Very interesting content
12. very interesting and useful topics
13. Very important and useful subject is covered.
14. Very different way of introducing the concepts. Free software.
15. Very different perspective/content than comparable 'Western' course. I don't know though, if such was ultimately beneficial. To wit, Quetelet index (per this course's context) is absent from top Google results.
16. Variety of data analysis techniques covered
17. Unusual viewpoints on standard topics
18. Uniqueness in the approach of the topic.
19. Topics covered, great discusson orum and support staff
20. Topics covered and depth of coverage.
21. Topics covered
22. topics and explanations, assignments and quizzes
23. Topics and a different view the prof has wrt the topics
24. Topics
25. topic
26. This was my first introduction to PCA and latent semantic analysis. Very clear introduction, liked it very much.
27. Theoretical background underlying standard techniques
28. The will to introduce reproducible examples to illustrate theory
29. The videos.
30. the videos
31. The video-lectures, and the programming assignments. I liked Prof. Mirkin's presentations of the various topics.
32. The unification of approaches to data analysis
33. the unconventional approach

34. The topics were presented in a very clear and usefull way.
35. The topics The materials (videos)
36. The topics of the course
37. The topics covered, the obvious knowledge of the instructor in the area.
38. The topics covered
39. The topics covered
40. The topics and professor's new approaches.
41. The text book
42. The teacher's courses explaination
43. The teacher: he has a subtle sense of humor and a novel approach to some subjects.
44. The subjects covered, the programming assignments
45. The subject.
46. The subject matter. I've studied statistics over a period of some decades but found some of Professor Mirkin's approaches thought provoking and very useful in my educational measurement work.
47. the subject
48. The study of PCA from a basic level. Not many do that.
49. The sense of genuine practical research and thorough understanding of the material by the lecturer. I wish I had a hundredth of his knowledge.
50. The Quetelet index and hidden factor models.
51. The programming assignments made me learn Mathematica, which was my programming environment
52. the proffesor
53. The Professors's lecture knowledge and style were a refreshing change. I was very impressed with the Professor.
54. the professor's knowledge on math and depth in explaining some concepts.
55. The professor.
56. The professor was great. He did a wonderful job explaning a complex topic.
57. The Professor is not only an expert in the field, but thinks for himself, and shares his insights and opinions!
58. The professor and the content.
59. The presentation by the professor was good. The discussion forum was helpful.
60. The practical applications of some interesting toipics
61. The possibility of implementing in MATLAB the content of the classes.
62. The pace and the assignments.
63. The overall topic
64. The originality of the approach
65. The original approach on some subjects and the tricks learned from practice on the field
66. The new ideas and clarity of representation. The videos were excellent.
67. the new concepts
68. The material is fascinating and Mirkin makes me laugh :)
69. the lectures
70. The lecturer and his way to place the small stories and jokes in the lessons.
71. The lector, and the quality of video presentation.
72. The latest variations introduced in the K-means clustering for selection of initial clusters. Especially, the one done by Prof. Merkins and his group himself.

73. The instructor's demonstration that mean, median, and range were all just different solutions to the same format of equation with different exponents. I've been working with these metrics for a decade and had never been told that information!
74. The Instructor Presentaion clarity and content
75. The instructor appeared to be an affable, pleasant and knowledgable man whom I immediately liked. I liked his friendly, good-humoured presentation style. It is a shame that it was poorly executed. There is a lot of potential in this course that would be unlocked if only the instructor improved his communication skills and polished his slides. I really wanted to like this course, but it bacame just too hard to follow what the instructor was trying to explain. This was a great pity.
76. The insight into the subject material by Dr Mirkin
77. The initialization on of K means
78. The fresh viewpoints on certain topics.
79. The explanation of the topics of Data Analysis
80. The difficulty is just right
81. the different perspective on the classical machine learning topics
82. The different approach taken. It is also a good compliment to other courses that did also cover some parts of the material but did not go into detail.
83. The depth of the material covered and the way in which topics were spread out and segmented through out the course. The way of presentation and the descriptive examples were good.
84. The depth of mathematics
85. The covered topics, the Matlab license and the availability of Mirkin's book chapters.
86. The covered topics were interesting.
87. The course material was very interesting and timely. Difficult to find this material anywhere else. Judging from the postings on the forum, the other participants also appreciated the uniqueness of the course offering. I liked being able to get instruction from a lecturer on the other side of the world from me. I liked having the Matlab access and tutorials since I've not had a chance to use it before. Of course, it was nice that this class was offered for free.
88. The course is well thought out and I like the way Prof. Mirkin presents video lectures, especially his British accent :)
89. The course covered different material to other data analysis courses.
90. The contents of the course. So the topics covered.
91. The contents (including the free book chapters).
92. the Content is good. And idea of delivering it through the video lectures and in video quizzes followed by quizzes is good.
93. The content is absolutely fantastic and relevant in the applied domain.
94. The content
95. The content
96. The concepts and utility of the mathematical contents have been explained together with the methods.
97. The clarity of explanations
98. The challenge in the homeworks
99. The areas covered
100. The applied samples usin matlab
101. The alternative insights prof. Mirkin presented for many subjects.

102. The alternative (not so commonly used) statistical methods
103. Te name of the course
104. Subject
105. Some of the quizzes required a deeper level of understanding and analysis than expected.
106. Some new ideas on methods of data analysis
107. Some insight into new research directions
108. Simplicity of teacher presentations. Efforts deployed in simplifying some complex concepts. Use of illustrating examples from real life.
109. Simple explanations given by the professor.to difficult topics. Helped me understand the material much better.
110. Selection of topics
111. right amount of math; PCA
112. quizzes and programming assignments
113. Quetlet index, PCA, k-means
114. programming assingnments
115. Programming Assignments
116. Programming assignments
117. Programming assignements, lessons and all the content in general
118. Programming
119. programming
120. Programing assignments
121. Professor's very pragmatic attitude. He would share what he thought was useful and what was not useful.
122. Professor Mirkin's view on the topic.
123. Professor Mirkin's simplification of some advanced concepts. TA were very helpful.
124. Professor Mirkin
125. Prof. Mirkin's depth of knowledge and unique approach to the subject of data analysis.
126. Prof. Mirkin was great - funny, knowledgeable, effective.
127. Prof. Mirkin is a very good lecturer, really enjoyed his very impressive lectures. Many topics were covered and if you will use recommended book you will be able to catch as many details as you what - so course is a good opportunity to dive into data analysis and statistics .
128. Prof MIrkin's lectures are excellent and delivered with a subtle sense of humor. They are also creative as the material in the lectures and in his book are his own or his own way of looking at it. His lectures are also a refreshing and practical way of looking at the topics. Definitely a jewel to keep. I wish they covered more topics but I imagine I can do that by reading the chapter of his book not covered in the course. Thank you for an excellent and memorable course!
129. Prfessor's consize and clear presentations
130. Presentor did a good job. Learned just enough to have an idea on the concepts, what i was looking for.
131. presentations, topic, depth
132. Presentation of concepts not found elsewhere. Very refreshing and different perspective on the matter.
133. Preciseness of the content, Experiential learning, Very Knowledge-able instructor!
134. PCA presentation

135. PCA and SVD covered well.
136. Overall I liked it very much. All the material was provided and was very interesting and useful.
137. New concepts learned e.g. data analysis, k means, pca
138. New approaches e.g Quetelet
139. Most Coursera courses gloss over the math and dumb down what is actually happening in an analysis. This course covered the material by working from the fundamentals and that was great!! We learnt not only how to do something but why we were doing it. I have taken about 10 data science course on coursera and this one was by far the best in terms of the depth and complexity of the material. I would like to see more courses like this.
140. math
141. Material was very well presented. Instructor and staff were very responsive.
142. manner how everything was taught
143. make to work in out somewhere in this world
144. Make introductions to data analysis.
145. Loved the topic, the coverage, the different look at the topic than other similar classes.
146. Lectures, material selection
147. Learning about data analysis.
148. learn the material such as PCA and such
149. K-means and PCA material. And Dr. Boris conceptualization and explanation of Data Analysis
150. K-Mean principle
151. Jeff Leek's lecture (brilliant!!)
152. It was very different from other data science courses and gave me a view from a completely different perspective.
153. it was a completely new topic for me, but i was comfortable with the pace and depth.
154. It took some time to get accustomed to the lecturer, but in a peculiar way he is funny. Interesting views on the topic.
155. It opened my eyes to Data Analysis Techniques I was not aware of. The use of Matlab was excellent, as the course forced me to learn it to do the assignments as well, so i kind of received 2 key areas of learning.
156. It gives core, not just sequences
157. It gave me knowledge which I was able to apply in my current work right away.
158. It gave me an overview of a subject in which I had interest, and prepared me for deeper studies.
159. It gave me a different perspective on data analysis and presented concepts that I had never seen; for example, the Minkowski criterion. It is a good complement to courses on mathematical statistics.
160. It covered some of the topics i was interested in. Those giving the course understood the subject matter very well.
161. Interesting and new topics for me
162. Instructor Boris Mirkin
163. in-front presentation of Prof. Birkin; presentation of his own work and opinion
164. i've enjoyed lecturer, his depth of knowledge, his humour and there were good tasks for assessment

165. I think the lectures were very well prepared and presented.
166. I really appreciate the feedback of Ekaterina on the discussion forums and all the work she did.
167. I love, love, love that I can take this class whenever I have time. I work full time as a database administrator and have 2 kids, so I was able to download the classes to my ipad and watch them while the kids were doing dance lessons and horse riding lessons. This is the best way for me to continue education. I also thought the instructor laid out the material very logically, and I've bought the book from Amazon.com to try to dig a little deeper. I also appreciated his wry sense of humor about some of the topics.
168. I liked the instructors different outlook on application of the subject matter especially the material on PCA.
169. I liked programming assingments. It was interesting, because you need to know how everything works, and than, if you have a problem, you need to know where that can be. You can not find it if you do not understand the matter
170. I like the topics covered and the assignments.
171.I like the depth of the material and the challenges of the assignments.
172. I like everything about this course. There is nothing like a free gift of knowledge. THANK YOU
173. I enjoyed the lectures and would have liked to follow through with the full experience. Still learning about retirement 'demands', which competes with coursework.
174. I dealt during the past with traditional approaches of PCA and clustering. I was very interested in exploring some new techniques of analysis of the subjects above.
175. I could plod through it without watching the lectures and still got the statement of accomplishment. While the instructions for the many assignments and quizzes were lacking in clarity, the TAs for the course were responsive to clarifying them. Late assignments were more clear.
176. I am over all very satisfied!
177. I actually liked the speed of the course. I was just not able to prioritize it high enough along with my other responsibilities. I also liked the interaction among the various students.
178. How to use PCA and clyustering
179. How to solve the problem of life safely?.
180. Hated it
181. Great video lectures.
182. Good topics covered. Programming assignments were fun but suffered from language, wording and technical issues.
183. Good topics
184. Good material, nice tests and programming assignements content
185. Good Information
186. Good explanation and Topic Breadth
187. good content.presentation
188. gives a broad spectrum to data analysis, and clues to not only take the 'well-known' road
189. Explanations and details, peer assessment
190. Examples, explained with experiwnce.
191. Examples are clear, instantly discussing the programming techniques needed for exercises, matlab as a programming language

192. Everything. The professor is excellent, clear, accurate. Perfect. I loved the course. Unfortunately I had not the time enough for finishing the course and doing all the exercises. I also need a more basic course on data analysis before I will come again to this one. But I will come for sure!
193. every things
194. Ekaterinas replies and clarifications
195. Ekaterina
196. Discovering there was a formal subject between the study of statistics and the topic known as big data.
197. Differs from american approach. Emphasis on SVD.
198. different views on subject matters compared to other teachers
199. Different perspectives at topics from what I was taught at my college.
200. Different approaches, frequent summarizing
201. Different approach to many concepts I had learned in previous courses.
202. Detailed explanations
203. Depth of understanding by the professor who presented his strong personal views and interpretations on the topics covered.
204. Depth of the topic
205. depth of lecturer knowledge
206. Depth of explanation, Boris was very lucid in explaining the basics as well as the difficult topics like principal component analysis. A great course.
207. Deep coverage
208. Covers a wide variety of techniques. Lectures where good.
209. Covered Themes
210. course depth and explanation; accompanying materials
211. Course coverage was an excellent primer to many different techniques.
212. Course content - breadth of content Depth of content
213. Contents, programming assignments.
214. contents
215. Content, video presentations
216. Content and examples. Real examples
217. Content
218. content
219. content
220. content
221. Conciseness and Preciseness
222. Clarity of the explanation, the comments of the real usage of the techniques;
223. Choice of topics
224. Boris Is EXCELLENT!
225. Bayesian statistics
226. assignments
227. Anomalous cluster, pca hidden factor
228. All but the evaluation information system.
229. A wide variety of tasks (quizes, programm assignments, peer assignments)
230. A dedicated course to the absolutly core features of analysis. Perfect with the focus on the stuff that all analysts need to understand

1) Presented a slightly different viewpoint to data science. 2) Forced students to go through course materials more carefully 3) Forced students to think in alternative ways. 4) Deadlines were reasonable.

231. I got a basic understanding of principal component analysis and cluster analysis. 2. I got the chance to apply my basic R knowledge to practical problems. 3.The discussions with students were very helpful.

232. 1. Connecting pure mathematics & statistics to analytics 2. Broad spectrum of topics: 1: The TAs timely support our questions online. Big thanks TAs. 2: Have teaching text book can read online or can purchase of print book. 3: This course cover many important and interesting topics. 4: The 'Programming Assessments' are very useful. 5: Have many real cases and MatLab examples in text book.

233. The technical content (2) Present refreshing alternate approach to many topics. Example SVD approach for PCA instead of the traditional correlation matrix approach

234. 'Hands on' experience of able to reproduce and solve real problems. Also the insight and meaning behind the numbers.

235. Filling in some topics which are the specialty of Boris Mirkin like intelligent kMean or Quetelet, I haven't heard of before - Concentration on the understanding instead of the math

236. similar subject material as in other courses but a different perspective

237. concepts like quetelet index,metrics of accuracy, PCA,CLUSTERING etc