UCT MA Psychology Multivariate Statistics

Assignment 7 2017

==========

(On reflection, part A of this assignment is difficult, so you can choose to answer either A or B)

A. In this assignment we will try to apply our statistical learning modeling skills to the problem of detecting fake hotel reviews (see the following link for a warning about fake hotel reviews by Tripadvisor (http://abcnews.go.com/Technology/story?id=8094231).

In fact, the hard work has been done for us to some degree by Myle Ott and colleagues (http://myleott.com/op_spamACL2011.pdf) – they built a web application which is a real time detector of fake reviews, using a method similar to those we have learnt about in the course (http://reviewskeptic.com/). Take a look at that web site – pretend that you have stayed in the Four Seasons Hotel in Paris, or the Savoy in London, or the Regis Hotel in New York, or some other swanky R 10 000 a night place, and see whether the fake review detector works!

If you want to know about the specific method Ott used, it is covered in our text, but we won't be covering it in the seminars (support vector machines, chapter 9).

1. Ott trained his model on a set of text reviews that were known to contain both genuine and fake reviews. These are available here: https://www.dropbox.com/sh/stj33r9xjijkcm3/AADJXhobtqJNWTppBllyQyn6a?dl=0
2. The text reviews need to be read into a data frame and processed in some way that will make them amenable to analysis. This requires us to learn a bit about Natural Language Processing – but just enough for the moment to turn the reviews into some form of data that can be used to predict veracity.
3. As a measure of veracity, we can code truthful reviews as 1 and fake reviews as 0.
4. The simplest predictors in NLP approaches are so-called n-grams (http://text-analytics101.rxnlp.com/2014/11/what-are-n-grams.html). We could take a very simple approach to the problem, like Ott did, and find the list of all 2-grams (bigrams) in the text corpus. Then we need to code each review for the presence of the bigram in it (i.e. if the bigram occurs in the review, we code 1 and if it does not occur we code 0)
5. This will lead to an unusual data structure, a little bit like the spreadsheet shown below:

Review_num	bigram_1_the_hotel	bigram_2_hotel_was	bigram_3_was_nice
1	0	0	0
2	0	0	0
3	1	1	1
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	1
9	0	0	0

6. This is a data structure known as a sparse matrix. We can use the set of predictor variables in the matrix to predict the outcome (in this case the binary variable coding veracity of review)

The most difficult part of the project is to get the hotel reviews into a set of predictor variables like those shown in the example above. There is a very useful package for R, though, called tidytext, by Julia Silge and colleagues, that makes it quite manageable. See the free web book http://tidytextmining.com/ that describes how to use the package, if you are interested.

Q1 Download the R script "tidytext deception script.R". This has all the steps required to get the data from the files into a form ready for analysis. Edit the script so that it works on your computer. Since the script contains no comments at all, you should put comments into the script so that each step is clearly explained for the next user. Improve the script, where appropriate, to make it clearer, or more efficient, or tidier (in the sense that Wickham means!) **(25 marks)**

Q2 Build a predictive model of review veracity. You are free to use any of the methods you have learnt about in the course. Remember to think about issues like validation sets, cross-validation, and so on. **(75 marks)**

B.

The data sets for this assignment concern wines grown in Italy and in Portugal. In the case of the Italian wines we want to correctly classify wine region and vintage on the basis of chemical properties of the wines. In the case of Portuguese wines we want to predict rated quality of wine according to chemical properties.

The data sets:

italianwines.csv The chemical properties are as declared in the names in the csv file. "hue" refers to the hue of the wine, it is not clear how this was measured. The column 'wine' refers to the geographic origin of the wine; each wine is made of a different combination of cultivars. Note that the row names have the year of production embedded in them (the last two digits).

winedata.xlsx (Portuguese wines). The chemical properties are as declared in the Excel row header. Quality is measured on a 10 point scale, subjectively i.e. by wine tasters

1a Attempt to predict the Italian wine types (Barolo, etc.) according to their physical properties. Use a classification tree approach, followed by bagging, random forests, and boosted classification trees. You may want to compare this to a simpler method, such as LDA. Report appropriate measures of error, and produce a final chart that compares the methods in their classification ability. **40**

1b Attempt to predict the vintage of Barolo in the data set, using the physical properties. Choose two methods only, and compare them. **10**

2 Attempt to predict the quality of Portuguese wines from their physical properties. Use a variety of regression methods in order to do so, but especially regression trees, bagged trees, random forests, and boosted trees. Report an appropriate comparison. **50**