DEPARTMENT OF STATISTICS

STATS 784 Data Mining

Assignment 3 2017

Due Friday September 1. Mail to Alan by 5pm.

Question 1. Using the books recommended in class, or any other source, write approximately one page about ROC curves (including the definition and how to interpret them.) Then, find a suitable R function and draw the ROC curve for the logistic regression predictor and the credit card default data discussed in class.
[10 marks]

Question 2. The data for this question consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples have been pre-processed by acoustic analysis and a series of summary measures recorded for each voice sample. The summary measures are

- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness
- kurt: kurtosis
- sp.ent: spectral entropy
- sfm: spectral flatness
- mode: mode frequency
- centroid: frequency centroid
- peakf: peak frequency
- meanfun: average of fundamental frequency
- minfun: minimum fundamental frequency
- maxfun: maximum fundamental frequency
- meandom: average of dominant frequency
- mindom: minimum of dominant frequency
- maxdom: maximum of dominant frequency
- dfrange: range of dominant frequency
- modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range

In addition, the gender of the speaker is recorded.

Your task is to develop a classification rule to predict the gender of a speaker from the summary measures. You should try all the techniques we have discussed in class. Which is best? The data are in the file **voices.csv**.
[10 marks]

Please include your name in all filenames.