Statistics 202
Fall 2011
Data Mining
Assignment #5
Due Tuesday December 13, 2011

Prof. J. Taylor

You may discuss homework problems with other students, but you have to prepare the written assignments yourself. Late homework will be penalized 10% per day.

Please combine all your answers, the computer code and the figures into one PDF file, we will not be accepting any other format for any of the homeworks. Name your file ``LastName_FirstInitial_HW5.pdf'' and send it to : `stats202-aut1112-staff`.

Grading scheme: 10 points per question, total of 40.

Q. 1) Use the similarity matrix below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which points are merged.

	V1	V2	V3	V4	V5
V1	1.00	0.10	0.42	0.54	0.35
V2	0.10	1.00	0.63	0.46	0.98
V3	0.42	0.63	1.00	0.41	0.85
V4	0.54	0.46	0.41	1.00	0.73
V5	0.35	0.98	0.85	0.73	1.00

Q. 2) The file http://stats202.stanford.edu/R/clust.R contains a function `generate` used to generate a dataset. We will use this function to generate some data to which we will apply some clustering techniques

(a) Describe, as explicitly as possible, the distribution of the data generated by this function with `extradim=0`.

(b) What does the `extradim` argument to the function do? Do you expect that it will be easier or harder to produce a good clustering with `extradim > 0`? Why or why not?

(c) Use this function to generate a dataset with the default arguments. Plot the points, coloring them by their "known label" which you can determine from reading the `R` code.

(d) Use `Mclust` to fit a model-based clustering to this data set. What should `Mclust` return in this case? Does it produce the expected result? (THIS IS OF COURSE RANDOM, SO WE SHOULD ASK "DOES IT PRODUCE THE EXPECTED RESULT MOST OF THE TIME?")

(e) Try increasing `extradim` to 10, 20, 30, 40, 50. Roughly when does `Mclust` stop producing the expected result most of the time?

(f) With `extradim=50`, try adjusting some of the other parameters to see if `Mclust` can more easily produce the correct result. What seems to happen with larger `n`? Smaller `sigma`? Larger `mu`?

Q. 3) We will again use the file http://stats202.stanford.edu/R/clust.R to generate data for clustering.

(a) Write a function that computes a "bounding box" for a data matrix of size $n \times p$ and generates an IID sample of size n uniformly on this bounding box. By bounding box, we mean a hyperectangle in $\mathbb{R}^p$ that contains all n points. We will use this function from (a) to compute the Gap statistic.

(b) Fix `extradim=0` and for $K = 1, \ldots, 9$ generate several datasets using the results of (a) and cluster these points using K-means.

(c) Make a plot of the clusterings, labelling each point by color or symbol, for each $K = 1, \ldots, 9$ to show what a typical clustering looks like on this bounding box.

(d) Produce the figure used for the Gap statistic. Do you see the expected gap at the "true" number of clusters?

(e) Repeat the above process for `extradim=10,20,30,40,50`. If there was a gap before, does it persist as `extradim` grows?

Q. 4) Use the red-wine quality data http://stats202.stanford.edu/data/winequality-red.csv for this question.

(a) Use `Mclust` to fit a model-based clustering mixture to this data. Which model does it choose as the best model?

(b) Fit an unconstrained model (`modelNames="VVV"`) to this data with the default settings for `G`.

(c) Use the EM algorithm described in class to fit this same model. You may also want to consult *Elements of Statistical Learning* to verify the equations.

(d) Compare your results to those of `Mclust`. You can find `Mclust`'s parameters in the output as `$parameters$mean` and `$parameters$variance$sigma`.