MS1b: SDM - Problem Sheet 2

1. (K-means) Troubled by the fact that K-means gives a different answer each time it is run, lets try removing the randomness from K-means. Instead of starting with random prototypes, lets always starting with the first k data points as the prototypes. Explain why this is a bad idea with an example. What might be a better approach?

2. (K-means) Let X be an $N \times p$ data matrix and $C = \{C_1, C_2, \ldots, C_K\}$ a partition of its row-vectors. Label the row vectors $X_{i(j)},\ i = 1, \ldots, K,\ j = 1, 2, \ldots, n_i$ by their cluster so that $X_{i(j)}$ is the j^{th} data vector in the i^{th} cluster, where $X_{i(j)}$ is a $1 \times p$ data vector.

   For each cluster C_i, define

$$\begin{aligned}\bar{X}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i(j)} \text{ be the within-cluster mean}\\ \bar{X} &= \frac{1}{n}\sum_{i=1}^{K} n_i \bar{X}_i = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} X_{i(j)} \text{ be the overall mean}\end{aligned}$$

   and let

$$\begin{aligned}T &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{i(j)} - \bar{X})^T(X_{i(j)} - \bar{X}) \text{ be the total deviance to the overall mean}\\ W &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{i(j)} - \bar{X}_i)^T(X_{i(j)} - \bar{X}_i) \text{ be the within-cluster deviance to the cluster mean}\\ B &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^T(\bar{X}_i - \bar{X}) \text{ be the between cluster deviance}\end{aligned}$$

   so, T, W and B are $p \times p$ matrices

   (a) Verify that $T = W + B$.

   (b) Suppose we cluster to minimize a within-cluster variation tr(W), what happens to tr(B)? i.e. How does T change as we vary the assignment of data vectors to clusters to minimize tr(W)? Is this desirable?

3. (K-means) In the notation of the previous question, let $W_i = \sum_{j=1}^{n_i}(X_{i(j)} - \bar{X}_i)^T(X_{i(j)} - \bar{X}_i)$ denote the within-cluster variance of the i^{th} cluster.

   (a) Express tr(W_i) and det(W_i) in terms of the variances of the principal components of the data vectors $X_{i1}, \ldots, X_{in_i}$ in the i^{th} cluster.

   (b) With the aid of an example, or otherwise, explain qualitatively why clustering to minimize $\sum_i$ tr(W_i) tends to produce clusters which are more spherical in outline than those produced by minimizing $\sum_i$ det(W_i).

4. (EDA) Obtain `http://www.stats.ox.ac.uk/%7Eteh/teaching/datamining/cognate.txt` and load it using something like `X <- read.table("cognate.txt")`. It contains an 87×2665 matrix of observations on each of 87 Indo-European languages where the presence (1) or absence (0) of 2665 homologous traits has been recorded.

   Historical linguists have grouped these languages into clades. Most large scale groupings are contested, but something like

$$\{Indic, Iranian\}$$

$$\{Balto - Slav, (Germanic, Italic, Celtic)\}$$

is not too controversial. The position of the Armenian, Greek, Albanian, Tocharian and Hittite groups is in doubt (though not within the second of the above super-clade).

We would like to cluster the languages into groups on the basis of these data. It is also of interest to represent the languages in a planar map in order to visualise similarities between languages.

(a) These data are categorical. The **S**imple **M**atching **C**oefficient for two data vectors is the proportion of variables which are unequal. The Jaccard coefficient for two language data vectors is the proportion of variables with at least one present which are unequal (so 1100 and 1010 have SMC 2/4 and JC 2/3). Which dissimilarity measure is appropriate for these data?

(b) Compute an agglomerative clustering of the data, and plot a dendrogram with language labels on the leaves. You will need to specify a distance measure between clusters. Include your R code for generating the dendrogram.

(c) Compute K-means clustering with 10 groups. Include your R code.

(d) Using MDS (the Sammon map may be best), represent the languages in a 2D plot. Plot the clusters obtained in part 4c using different symbols, or colors and super-pose the language name. Can you see any geographical grouping in the layout?