SMLDM HT 2014 - Part C Problem Sheet 2

1. In lectures we discussed using the Mahalanobis distance to measure distances in K-means:

$$\|x - y\|_M = \sqrt{(x-y)^\top M^{-1}(x-y)}$$

where M is a positive definite matrix. Explain why using this distance measure is equivalent to applying K-means with the original Euclidean distance on a transformed data set. What is choice of the M matrix leads to an equivalent algorithm to first whitening the data?

2. In lectures we derived the M step updates for a mixture of Gaussians, for the mixing proportions and cluster means, assuming the common covariance $\sigma^2 I$ is fixed and known. What happens to the algorithm if we set σ^2 to be very small? How does the resulting algorithm as $\sigma^2 \to 0$ relate to K-means?

3. In lectures we derived the M step updates for a mixture of Gaussians, for the mixing proportions and cluster means, assuming the common covariance $\sigma^2 I$ is fixed and known. If σ^2 is in fact not know and to be learnt as well, derive an M step update for σ^2.

4. Assume you are interested in clustering n binary images. Each binary image $x_i = (x_{i1}, \ldots, x_{ip})$ for each $i = 1, \ldots, n$ corresponds to a vector of p binary random variables; p being the number of pixels. We adopt a probabilistic approach and model the probability mass function of x_i using a finite mixture model with mixing proportions $\pi_1, \ldots, \pi_K$ satisfying $\pi_k \geq 0$ for each k and $\sum_{k=1}^K \pi_k = 1$, and each mixture component k is modelled as a product of p Bernoulli distributions with parameters $\phi_k = (\phi_{k1}, ..., \phi_{kp}) \in [0,1]^p$. In other words, mixture component k models each entry x_{ij} in the data vector as an independent Bernoulli with mean ϕ_{kj}, for $j = 1, \ldots, p$.

   (a) Write down the log likelihood explicitly as a function of the parameters $\theta = (\pi_k, \phi_k)_{k=1}^K$ given $x_{1:n}$.

   (b) We want to estimate the unknown parameters by maximizing the log likelihood using the EM algorithm. Let Z_i be the variable indicating which component x_i belongs to, z_i a value in $\{1, \ldots, K\}$, and $\mathbf{z} = (z_i)_{i=1}^n \in \{1, \ldots, K\}^n$. Write down explicitly the free energy $\mathcal{F}(\theta, q)$ lower bound on the log likelihood, as a function of $q(\mathbf{z})$ and of the parameters (π_k, ϕ_k).

   (c) Derive explicitly the EM update equations by setting derivatives of $\mathcal{F}$ to zero and solving.

5. Obtain `http://www.stats.ox.ac.uk/%7Eteh/teaching/smldmHT2014/cognate.txt` and load it using `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 agglomerative clusterings of the data. Try using both SMC and Jaccard, as well as single, average and complete linkage. Plotting the dendrograms with language labels on the leaves, which combination of distance and linkage algorithm seem to produce sensible results? Include your R code for generating your favourite dendrogram. You can use `D<-dist(X,method="binary")` to compute the Jaccard distances, and `D<-dist(X,method="manhattan")` for SMC, followed by `hclust(D,method=....)` or `agnes(D,method=...)` for linkage algorithms (agnes is part of the cluster library, so you have to load using `library(cluster)`).

6. Let $x_1, \ldots, x_n$ be a dataset of p-dimensional vectors and $C = \{C_1, C_2, \ldots, C_K\}$ a partition of $\{1, \ldots, n\}$. For each cluster C_k, define

$$\bar{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i \qquad \text{to be the within-cluster mean}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{to be the overall mean}$$

and let

$$T = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x})(x_i - \bar{x})^\top \qquad \text{to be the total deviance to the overall mean}$$

$$W = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k) \qquad \text{to be the within-cluster deviance to the cluster mean}$$

$$B = \sum_{k=1}^{K} \sum_{i \in C_k} (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top \qquad \text{to be the between cluster deviance}$$

where T, W and B are $p \times p$ matrices.

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

(b) Explain how the K-means objective is related to W.

(c) How does T change during the course of the K-means algorithm? How does B change?