SDMML HT 2016 - Part C Problem Sheet 2

1. 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, denote $n_k = |C_k|$ and 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

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

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

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

where T, W and B are all $p \times p$ 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?

2. In lectures, we derived the M-step updates for fitting Gaussian mixtures with EM algorithm, for the mixing proportions and for the cluster means, assuming the common covariance $\sigma^2 I$ is fixed and known.

   (a) 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?

   (b) If σ^2 is in fact not known and is a parameter to be inferred as well, derive an M-step update for σ^2.

3. Assume you are interested in clustering n binary images. Binary images are modelled as i.i.d. samples $\{x_i\}_{i=1}^n$ for each $i = 1, \ldots, n$ from a random vector $X_i = (X_{i1}, \ldots, X_{ip})$ of p binary random variables (p being the number of pixels). Probability mass function of X_i is a mixture 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 independent Bernoulli variables with parameters $\phi_k = (\phi_{k1}, ..., \phi_{kp}) \in [0, 1]^p$. In other words, $Z_i \sim$ Discrete $(\pi_1, \ldots, \pi_K)$ is the variable on $\{1, \ldots, K\}$ indicating which component X_i belongs to, and $X_{ij} | Z_i = k \sim$ Bernoulli (ϕ_{kj}) independently for $j = 1, \ldots, p$.

   (a) Having observed dataset $\{x_i\}_{i=1}^n$, write down the log-likelihood explicitly as a function of the parameters $\theta = (\pi_k, \phi_k)_{k=1}^K$.

(b) We want to estimate the unknown parameters by maximizing the log-likelihood using the EM algorithm. Denote by z_i a value in $\{1, \dots, K\}$, and $\mathbf{z} = (z_i)_{i=1}^n \in \{1, \dots, K\}^n$. Write down explicitly the variational free energy $\mathcal{F}(\theta, q)$ 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 $\mathcal{F}$ w.r.t. q, π_k and ϕ_{kj} to zero and solving.

4. Verify that in the probabilistic PCA model from the lectures, E-step of the EM algorithm at iteration $t+1$ can be written as

$$q^{(t+1)}(y_i) = \mathcal{N}\left(y_i; b_i^{(t)}, R^{(t)}\right)$$

where

$$\begin{aligned} b_i^{(t)} &= \left((L^{(t)})^\top L^{(t)} + (\sigma^2)^{(t)} I\right)^{-1} (L^{(t)})^\top x_i, && (1) \\ R^{(t)} &= (\sigma^2)^{(t)} \left((L^{(t)})^\top L^{(t)} + (\sigma^2)^{(t)} I\right)^{-1}. && (2) \end{aligned}$$

Optional (using R)

5. Download `cognate.txt` from `http://www.stats.ox.ac.uk/~sejdinov/sdmml/data/` 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 and why?

(b) Run MDS with Sammon mapping using both SMC and Jaccard distance on these data. You can use
`D<-dist(X,method="binary")` to compute the Jaccard distances, and
`D<-dist(X,method="manhattan")` for SMC.

(c) Compute agglomerative clustering of the data using Jaccard with single, average and complete linkage. Plotting the dendrograms with language labels on the leaves, which linkage algorithm seems to produce sensible results? You can use `hclust(D,method=....)` or `agnes(D,method=...)` for various choices of linkage (agnes is part of the cluster library, so you have to load using `library(cluster)`).