# SDMML HT 2016 - MSc Problem Sheet 2

1. 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 $87 \times 2665$ 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)`).

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 $\sigma^2$ to be very small? How does the resulting algorithm as $\sigma^2 \to 0$ relate to K-means?

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

3. Consider two univariate normal distributions $\mathcal{N}(\mu, \sigma^2)$ with known parameters $\mu_A = 10$ and $\sigma_A = 5$ for class A and $\mu_B = 20$ and $\sigma_B = 5$ for class B. Suppose class A represents the random score $X$ of a medical test of normal patients and class B represents the score of patients with a certain disease. A priori there are 100 times more healthy patients than patients carrying the disease.

   (a) Find the optimal decision rule in terms of misclassification error (0-1 loss) for allocating a new observation $x$ to either class A or B.

(b) Repeat (a) if the cost of a false negative (allocating a sick patient to group A) is $\theta > 1$ times that of a false positive (allocating a healthy person to group B). Describe how the rule changes as $\theta$ increases. For which value of $\theta$ are 84.1% of all patients with disease correctly classified?

4. For a given loss function $L$, the risk $R$ is given by the expected loss

$$R(f) = \mathbb{E}\left[L(Y, f(X))\right],$$

where $f = f(X)$ is a function of the random predictor variable $X$.

(a) Consider a regression problem and the squared error loss

$$L(Y, f(X)) = (Y - f(X))^2.$$

Derive the expression of $f = f(X)$ minimizing the associated risk.

(b) What if we use the absolute ($L_1$) loss instead?

$$L(Y, f(X)) = |Y - f(X)|.$$

5. Download `wine.data` from
`http://www.stats.ox.ac.uk/~sejdinov/sdmml/data/`
and load it using `read.table("wine.data",sep=",")`. Description of the dataset is given at `https://archive.ics.uci.edu/ml/datasets/Wine`. The goal is to build a classifier for predicting the cultivars given in column 1. Train LDA classifier on a random subset of 50% of the data, and show the projections of the data vectors as well as the decision boundaries in the 2D LDA component space. Then predict the cultivars for the other 50% of the data and plot these in the LDA component space as well (using a different `pch`). How many errors did the classifier make (a) on the training set, (b) on the test set?

## Optional

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 $C_k$, 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?

7. In binary classification, suppose that $\mathbb{P}(Y = -1)$ is very small, so that the constant classifier $f(x) = +1, \forall x$, has a small risk under the 0/1 loss. Consider the following loss instead:

$$L_{\alpha,\beta}(Y, f(X)) = \begin{cases} \alpha & \text{if } Y = -1, f(X) = +1, \\ \beta & \text{if } Y = +1, f(X) = -1, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\alpha$ and $\beta$ that result in the following risk

$$R(f) = \mathbb{P}(f(X) = +1 | Y = -1) + \mathbb{P}(f(X) = -1 | Y = +1).$$