Statistical Machine Learning HT 2018 - Problem Sheet 2

1. **(Clustering, moved from PS1).** 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 the matrix W. Does it depend on all elements of W?

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

2. **(K-means, moved from PS1).** 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 is equivalent to applying K-means using the standard Euclidean distance on a transformed data set. What is the choice of the M matrix that leads to an algorithm which is equivalent to first whitening the data? [*Hint: Consider a linear transformation $x \mapsto Ax$.*]

3. **(Deviance minimization equivalence).** In the lectures, we considered *within-cluster deviance* $W(C_k, \mu_k) = \sum_{i \in C_k} \|x_i - \mu_k\|_2^2$. Show that

$$W(C_k, \mu_k) = W(C_k, \bar{x}_k) + |C_k| \cdot \|\mu_k - \bar{x}_k\|_2^2$$

where $\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$. Hence, conclude that $W(C_k, \mu_k)$ is minimized for $\mu_k = \bar{x}_k$ and show that

$$W(C_k, \bar{x}_k) = \frac{1}{2|C_k|} \sum_{i,j \in C_k} \|x_i - x_j\|_2^2.$$

4. **(Regression loss, squared and absolute).** 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. **(ESL: Train/test)**. Consider a linear regression model with p parameters, fit with unregularized linear regression (sometimes called "least squares" or "ordinary least squares" or even just OLS) to a set of training data $(x_i, y_i)_{1 \le i \le N}$ drawn at random from a population. Let $\hat{\beta}$ be the estimator. Suppose we have some test data $(\tilde{x}_i, \tilde{y}_i)_{1 \le i \le M}$ drawn at random from the same population as the training data.

If $R_{tr}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i^\top \beta\right)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_{i=1}^{M}\left(\tilde{y}_i - \tilde{x}_i^\top \beta\right)^2$, prove that

$$E[R_{tr}(\hat{\beta})] \le E[R_{te}(\hat{\beta})]$$

where the expectation is over all that is random in each expression.

6. **(Weighted loss).** 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 θ increases. For which value of θ are 84.1% of all patients with disease correctly classified?

7. **(*Coding, moved from PS1*: MDS dissimilarity measure, hierarchical clustering).** Download `cognate.txt` from `http://www.stats.ox.ac.uk/~palamara/teaching/SML18/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 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)`).

8. **(Optional: ESL Bayesian L2)**. (Assumes some knowledge of Bayesian statistics.) We are performing linear regression using the squared loss, but we place a Gaussian prior on the vector of coefficients we are trying to learn: $\beta_j \sim N(0, \sigma_\beta^2)$, for $j \in \{1, \ldots, p\}$. We assume that the observations y_i are sampled adding Gaussian noise with variance σ_ϵ^2 to points from the underlying linear model given by $\beta_0 + x_i^\top \beta$, so that $y_i \sim N(\beta_0 + x_i^\top \beta, \sigma_\beta^2)$, for $i \in \{1, \ldots, n\}$. Find the maximum-a-posteriori (which is also the posterior mean) for the vector of coefficients β. Show the equivalence between this estimator and the estimator obtained when performing L2-regularized (ridge) regression, and interpret the relationship between the regularization parameter λ and the parameters σ_β^2 and σ_ϵ^2. *[Hint: marginal and conditional distributions of multivariate Gaussians have standard expressions].*