SMLDM HT 2014 - MSc Problem Sheet 2

1. 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?

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. If σ^2 is in fact not know and to be learnt as well, derive an M step update for σ^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 θ increases. For which value of θ 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

$$R(\hat{Y}) = E(L(Y, \hat{Y}(X))),$$

   where $\hat{Y} = \hat{Y}(X)$ is a function of the random predictor variable X.

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

$$L(Y, \hat{Y}(X)) = (Y - \hat{Y}(X))^2.$$

   Derive the expression of $\hat{Y} = \hat{Y}(X)$ minimizing the associated risk.

   (b) What if we use the ℓ_1 loss instead?

$$L(Y, \hat{Y}(X)) = |Y - \hat{Y}(X)|.$$

5. Show that under a Naïve Bayes model, the Bayes classifier $\hat{Y}(x)$ minimizing the total risk for the $0-1$ loss function has a linear discriminant function of the form

$$\hat{Y}(x) = \arg\max_{k=1,2} \quad \alpha_k + \beta_k^\top x.$$

   and find the values of α_k, β_k. (Use notation from lecture slides).

6. Suppose we have a two-class setup with classes -1 and 1, that is $\mathcal{Y} = \{-1, 1\}$ and a 2-dimensional predictor variable X. We find that the means of the two groups are at $\hat{\mu}_{-1} = (-1, -1)^\top$ and $\hat{\mu}_1 = (1, 1)^\top$ respectively. The a priori probabilities are equal.

   (a) Applying LDA, the covariance matrix is estimated to be, for some value of $0 \leq \rho \leq 1$,

$$\hat{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

   Find the decision boundary as a function of ρ.

(b) Suppose instead that, we model each class with its own covariance matrix. We estimate the covariance matrices for group -1 as

$$\hat{\Sigma}_{-1} = \begin{pmatrix} 5 & 0 \\ 0 & 1/5 \end{pmatrix},$$

and for group 1 as

$$\hat{\Sigma}_{1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 5 \end{pmatrix}.$$

Describe the decision rule and draw a sketch of it in the two-dimensional plane.