SMLDM HT 2014 - Part C Problem Sheet 3

1. 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?

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

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

   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?

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

3. Consider applying LDA to a two-class dataset. We will verify some of the claims in the lectures. We use the notation in the lectures.

   (a) Show that $\Sigma^{-1}(\mu_1 - \mu_2)$ spans the one-dimensional discriminant subspace. What is the corresponding generalized eigenvalue?

   (b) Explain why to predict the class of a data vector x it is sufficient to look at its projection onto the subspace spanned by $\Sigma^{-1}(\mu_1 - \mu_2)$.

   (c) In the case where the within-class covariance is $\Sigma = I$, explain the geometry of the decision rule of LDA with the help of a diagram.

4. 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

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

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

5. 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.

6. (Challenging; submit this in Friday Week 5)

   Implement the EM algorithm from Problem Sheet 2 Question 4 on a dataset of handwritten digits of '0', to '9'.

   You can obtain the dataset from
   `http://www.stats.ox.ac.uk/\%7Eteh/teaching/smldmHT2014/usps.txt`
   Load the dataset using
   `X <- as.matrix(read.table("usps.txt"))`
   The dataset is binary and has 1000 rows and 64 columns. Each row gives a 8×8 binary image, and you can visualize the images using a command like
   `image(matrix(X[i,],8,8))`
   where $i \in \{1, \ldots, 1000\}$.

   In the E step of the algorithm, you will need to compute the the posterior probabilities $q(z_i = k)$ in a numerically stable manner. Specifically, the unnormalized posterior probabilities are:

   $$\pi_k \prod_{j=1}^{p} (\phi_{kj})^{x_{ij}} (1 - \phi_{kj})^{1 - x_{ij}}$$

   We compute the logarithms instead:

   $$S_{ik} = \log \pi_k + \sum_{j=1}^{p} x_{ij} \log \phi_{kj} + (1 - x_{ij}) \log(1 - \phi_{kj})$$

   Then the posterior probabilities are computed as:

   $$q(z_i = k) = \frac{e^{S_{ik} - \max_{k'} S_{ik'}}}{\sum_l e^{S_{il} - \max_{k'} S_{ik'}}}$$

while the marginal log probability is computed as:

$$\log p(x_i) = \log \sum_k e^{S_{ik}} = \left(\max_{k'} S_{ik'} \right) + \log \sum_k e^{S_{ik} - \max_{k'} S_{ik'}}$$

This is called the "log-sum-exp" trick.

In the M step, to prevent overfitting, where the estimated probabilities approach 0 or 1, we simply add small constants to the formulas:

$$\phi_{kj} = \frac{\alpha + \sum_{i=1}^n q(z_i = k) x_{ij}}{2\alpha + \sum_{i=1}^n q(z_i = k)} \qquad \pi_k = \frac{\beta + \sum_{i=1}^n q(z_i = k)}{K\beta + n}$$

You can try setting $\alpha = \beta = 1$. (We will understand this technique, called Laplace smoothing, in the coming lectures.)

(a) Run the EM algorithm, with 20 mixture components for 50 iterations. Plot the log likelihood as it increases over iterations of the algorithm. Has the algorithm converged? Display the learned means of each mixture component.

(b) Do the clusters generally correspond to the classes of handwritten digits?

(c) Do you get the same solution if you run the EM algorithm from different initial starting configurations?

(d) If you increase the number of components do the resulting log likelihood objective function generally increase or decrease?

(e) How many components do you think is sensible?