SDMML HT 2016 - 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

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

   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?

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

3. Suppose we have a two-class setup with classes -1 and 1, i.e., $\mathcal{Y} = \{-1, 1\}$, 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 estimated prior class probabilities $\hat{\pi}_1$ and $\hat{\pi}_{-1}$ 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.

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

   (a) Show that between-class covariance B is equal to $B = \frac{n_1 n_2}{n^2}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$ and thus has rank 1. Show that $v_1 = \Sigma^{-1}(\mu_1 - \mu_2)$ spans the one-dimensional discriminant subspace, i.e., that $u_1 = \Sigma^{\frac{1}{2}} v_1$ is an eigenvector of $B^\bullet = \Sigma^{-\frac{1}{2}} B \Sigma^{-\frac{1}{2}}$. What is the corresponding eigenvalue?

   (b) Explain why it is sufficient to look at the projection of a data vector x onto the discriminant subspace, i.e. 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.

5. Show that under a Naïve Bayes model with binary predictors from the lectures, the Bayes classifier $f_{\text{Bayes}}(x)$ minimizing the total risk for the $0-1$ loss has a linear discriminant function of the form

$$f_{\text{Bayes}}(x) = \arg\max_{k=1,2} \; a_k + b_k^\top x.$$

   and find the values of a_k, b_k.

6. Data in the table gives attributes of 14 customers of an electronics shop and a class label on whether or not they purchased a computer. Using a Naïve Bayes classifier, predict whether a new customer $x = (\texttt{age=youth}, \texttt{income=medium}, \texttt{student=yes}, \texttt{credit=fair})$ will buy a computer?

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no