MS1b: SDM - Problem Sheet 5

1. (Missing Features) Assume you have used some training data $D = \{x_i, y_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$ and $y_i \in \{1, 2, ..., K\}$ to learn a probabilistic classifier (say using maximum likelihood). We are interested in classifying a new input vector. However we have only been able to collect $p-1$ features, say you have only $\left(x_i^1, ..., x_i^{l-1}, x_i^{l+1}, ..., x_i^p\right)$ and x_i^l is missing. Explain whether or not it is possible to use your classifier to classify this incomplete input vector in the two following scenarios:

   (a) When we consider a naïve Bayes model where
   $$f\left(x|\,\phi_k\right) = \prod_{l=1}^{p} g\left(x^l \middle|\, \phi_k^l\right);$$
   i.e. conditional upon $Y = k$, you assume that the features are independent and feature x^l follows a distribution with density $g\left(x^l \middle|\, \phi_k^l\right)$.

   (b) When we consider a QDA model; i.e.
   $$f\left(x|\,\phi_k\right) = \mathcal{N}\left(x; \mu_k, \Sigma_k\right).$$

   (c) Generally speaking, which conditions on $f\left(x|\,\phi_k\right)$ are necessary to allow us to implement easily, that is without using any numerical integration scheme, a probabilistic classifier in presence of missing features?

2. (Bayesian classification) Consider some training data $D = \{(x_i, z_i), y_i\}_{i=1}^n$ where $(x_i, z_i) \in \mathbb{R}^p \times \mathbb{R}$ is the vector of inputs and $y_i \in \{0, 1\}$ the response. We adopt the following regression model for class k
   $$Z = \beta_k^{\mathrm{T}} X + \varepsilon$$
   where $\varepsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_k^2\right)$ if $Y = k$. Hence we have for the class conditional density $f_k\left(z|\,x\right) = \mathcal{N}\left(z; \beta_k^{\mathrm{T}} x, \sigma_k^2\right)$ so that the unconditional density of Z follows a so-called mixture of regressions model. Note that this model differs conceptually from the examples discussed in lectures as we do not model X. We adopt the notation $P\left(Y = k\right) = \pi_k$ and denote $\theta = \left(\pi_1, \beta_0, \beta_1, \sigma_0^2, \sigma_1^2\right)$ the set of unknown parameters.

   (a) Give an expression of the estimate $\widehat{\theta}$ of θ maximizing the conditional log-likelihood
   $$l\left(\theta\right) = \sum_{i=1}^{n} \log\, p\left(y_i, z_i|\, x_i, \theta\right).$$
   What happens when $n < p$?

   (b) Consider a Bayesian approach with $\pi_1 \sim$Beta(a, b) and
   $$p\left(\sigma_0^2, \beta_0, \sigma_1^2, \beta_1\right) = p\left(\sigma_0^2, \beta_0\right) p\left(\sigma_1^2, \beta_1\right)$$

where $p\left(\sigma_k^2, \beta_k\right)$ satisfies a normal inverse-Gamma distribution; e.g.

$$\begin{aligned} p\left(\sigma_k^2, \beta_k\right) &= p\left(\sigma_k^2\right) p\left(\beta_k \middle| \sigma_k^2\right) \\ &= \mathcal{IG}\left(\sigma_k^2; \frac{\nu}{2}, \frac{\kappa}{2}\right) \mathcal{N}\left(\beta_k; 0, \sigma_k^2 \Sigma\right) \end{aligned}$$

with some hyperparameters $(\nu, \kappa, \delta\Sigma)$ such that $\nu, \kappa > 0$ and Σ is a positive definite matrix. $\mathcal{IG}\left(\sigma^2; \frac{\nu}{2}, \frac{\kappa}{2}\right)$ denotes the inverse-Gamma density given by

$$\mathcal{IG}\left(\sigma^2; \frac{\nu}{2}, \frac{\kappa}{2}\right) = \frac{\left(\frac{\kappa}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} \left(\sigma^2\right)^{-\frac{\nu}{2}-1} \exp\left(-\frac{\kappa}{2\sigma^2}\right).$$

The posterior distribution $p\left(\theta \middle| D\right)$ satisfies

$$p\left(\theta \middle| D\right) = p\left(\pi_1 \middle| D\right) p\left(\sigma_0^2, \beta_0 \middle| D\right) p\left(\sigma_1^2, \beta_1 \middle| D\right).$$

Show that $p\left(\pi_1 \middle| D\right)$ is a Beta distribution and $p\left(\sigma_k^2, \beta_k \middle| D\right)$ a normal inverse-Gamma distribution.

(c) Given a new test data (x, z), establish the expression of $p\left(z \middle| D, x, y = k\right)$ and explain how you would use this expression to obtain a Bayesian classifier. What are the potential benefits of this approach over using $p\left(z \middle| \widehat{\theta}, x, y = k\right)$?

3. (Logistic Regression) Consider two-class data, (X, Y) with $Y \in \{-1, 1\}$ and $X = (B, Z)$ with $B \in \{0, 1\}$ and $Z \in \mathbb{R}^{p-1}$, so data vectors are made up of a binary variable and $p-1$ continuous variables. Training data (X_i, Y_i), $i = 1, \ldots, n$ are available. Consider logistic regression of the log posterior odds, $\log P(Y = 1|x) - \log P(Y = -1|x) = \alpha + \beta^T x$ with $\beta = (\beta_b, \beta_z) \in \mathbb{R} \times \mathbb{R}^{p-1}$, so that $\alpha + \beta^T x \equiv \alpha + \beta_b b + \beta_z^T z$.

   (a) Suppose there are no $B = 1$ outcomes in the class $Y = -1$ data (so $B_i = 0$ for all $i \in \{j : Y_j = -1\}$) but there are both $B = 0$ and $B = 1$ outcomes in the $Y = 1$ data. Show that the maximum likelihood estimate for β_b is $\hat{\beta}_b = \infty$.

   (b) Problems of this kind arise when there exists a separating-hyperplane. Show that the assumptions of 3a do not imply the existence of a separating hyperplane.

4. Load the Vanveer gene expression data used in a previous practical. Make use of the 20 'best' genes (according to a marginal t-test) by using the following commands.

```
load(url("http://www.stats.ox.ac.uk/%7Eteh/MS1b/PracticalObjects.RData"))
vanv<- vanveer.4000[,2:21]
prog<- vanveer.4000[,1]
```

Your X matrix is thus `vanv` and the response Y is `prog`. Split the data into a test and training set (of equal size). Using logistic regression, plot a ROC curve.