MS1b Statistical Data Mining – Specimen Paper TT 2013

Time: $1\frac{1}{2}$ hours

Rubric: *"You may attempt as many questions as you wish but* **only your two best answers will count***."*

[**MS1b 2008 Q2**]

1. Write brief and concise yet complete answers to the following questions (a few sentences in each case).

(a) What is the key difference between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) for binary classification? How many parameters need to be estimated for LDA and for QDA? Why is LDA sometimes preferable to QDA? Would this argument be more relevant for data with a small sample size or with a large sample size? Compare bias and variance qualitatively for LDA and QDA.

(b) Give one difference between Random Forests and bagged trees in terms of how the individual trees are grown.

(c) If we grow a classification tree with n samples and a p-dimensional predictor variable, over how many split-points do we have to search when looking for the best split at the root node in an exhaustive search? How many leaf nodes can a tree have at most? What is the minimal depth of a tree (depth is the maximal pathlength from root node to any of the leaf nodes) if its grown to full size as a function of the sample size n? What is the apparent error, or training error, if we grow the tree to full size?

(d) Explain briefly 5-fold cross-validation and give an advantage of 5-fold cross-validation compared to leave-one-out cross-validation.

(e) Give two advantages of ensembles of trees over single trees. Conversely, give one advantage of single trees over ensembles of trees.

(f) Does the variance of a tree increase, on average, if the size of the tree is getting larger or does it decrease? What about the bias? How does bagging affect these two quantities typically? [*No derivation is required.*]

[**MS1b 2008 Q3**]

2. (a) Given a data matrix $X \in \mathbb{R}^{n \times p}$, we are interested in using Exploratory Data Analysis techniques to view the data and achieve a dimensionality reduction. Give a short description of (i) Principal Components Analysis, (ii) Linear Discriminant Analysis, and (iii) Multi-Dimensional Scaling, and their underlying assumptions.

(b) Given a data matrix $X = (X_1, \ldots, X_p)$, explain in detail how to find the first two principal components of the data. Give not only details of the calculation but give also details of the underlying constrained optimisation problem.

(c) Suppose we project the data $X \in \mathbb{R}^{n \times p}$ into the k-dimensional space spanned by the first k principal components, where $k \leqslant p$ and $p \leqslant n$. Suppose also that n observations Y of some response variable are available. A logistic regression is performed, using the projected data of dimension k instead of the original data X of dimension p as predictor variables. Is the result going to be different to a logistic regression that uses the original data X as predictor variables? Justify your answer. What will be the effect of performing logistic regression with the projected data of dimension $k \leqslant p$ instead of the original data of dimension p on the bias and the variance of the logistic regression estimation?

(d) Does it make sense to search for Principal Components for a data matrix $X \in \mathbb{R}^{n \times p}$ if $p > n$? Suppose we have additionally n observations of a binary response variable Y. Does it make sense to apply Linear Discriminant Analysis to the data if $p > n$? Justify your answers.

[**MS1b 2008 Q4**]

3. (a) Suppose we observe $\mathcal{T} = \{x_{k,i}, \ldots, x_{k,n_k} : k = 1, 2\}$ where each $x_{k,i} \in \mathbb{R}$. Assuming $p_k(x) \sim N(\mu_k, \sigma_k)$ with class prior π_k for each class, find an expression for the Bayes Classifier $\widehat{c}(x)$ under 0-1 loss.

(b) Comparing classes i and j, show that the resulting decision boundaries of this classifier are quadratic in x.

(c) Find the plug-in estimates $\widehat{\mu}_k$ and $\widehat{\sigma}_k$ for $k = 1, 2$ by maximum likelihood estimation.

(d) Suppose the plug-in estimates are

$$\widehat{\mu}_1 = (1, 1)^T,\ \widehat{\mu}_2 = (-1, -1)^T,\ \widehat{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } \widehat{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and $\pi_1 = \pi_2 = 1/2$. What is the equation of the decision boundary separating the two classes?

(e) What is the decision boundary if all plug-in estimators are unchanged except for the covariance matrices, which are now assumed to be

$$\widehat{\Sigma}_1 = \widehat{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix},$$

for some value of $\sigma^2 > 0$? Find the asymptotic limit of the decision boundary for $\sigma^2 \to \infty$.