MS1b Statistical Data Mining – Practice Questions TT 2013

MS1b Statistical Data Mining has not had a written exam for the last few years. In addition to the questions on the TT 2013 Specimen Paper, the questions below are those that are most relevant for this year's version of MS1b. Many of these are MSc questions on the course from recent years.

[**MS1 2006 Q6, part (I)**]

1. (I) Suppose X is a p-dimensional vector of independent Bernoulli variables, $X = (X_1, X_2, ..., X_p)$. Let $X_j = 1$ with probability p_{kj} when X belongs to class $C = k$, $k = 1, ..., K$. Show that the Bayes-optimal classification rule leads to a linear discrimination function between pairs of classes.

(II) [. . .]

[**MS1 2006 Q8, parts (iii), (iv)**]

2. Consider an $n \times n$ matrix Δ of dissimilarity measurements. Entry $\Delta_{i,j}$ gives some measure of dissimilarity between objects corresponding to rows i and j. We wish to find n data vectors $x_1, x_2, ..., x_n$ with the property that $\Delta_{i,j} = |x_i - x_j|$ for each $i, j = 1, 2, ..., n$. The data vectors are row vectors $x_i \in \mathbb{R}^p$ $i = 1, 2, ..., n$ for some p to be determined.

(i) [. . .]

(ii) [. . .]

(iii) What is a biplot? Explain how the plotted points, and the plotted vectors, are computed from the data.

(iv) How do we interpret inner products between projections of the original variable axes (of the data) into the axes of the biplot? Justify the interpretation in terms of the singular value decomposition of X.

[**MS1 2007 Q6, parts (i)(a), (ii)(a), (ii)(b)**]

3. (i) Consider a partition of n data vectors into K clusters. Let y_{ij} denote the jth data vector in the ith cluster (so, y_{ij} is a $1 \times p$ data vector and y_{ij}^T is a column vector). For $i = 1, 2, \ldots, K$, let $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ and $\overline{y} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij}$. Denote by

$$W = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^T (y_{ij} - \overline{y}_i)$$

the within-cluster variation and by

$$B = \sum_{i=1}^{K} (\overline{y}_i - \overline{y})^T (\overline{y}_i - \overline{y})$$

the between cluster variation.

(a) Define the K-means clustering algorithm, give its objective function show that it stops, and show that it may stop at a clustering which is not globally optimal.

(b) [...]

(c) [...]

(ii) (a) Suppose $C \in \{1, 2\}$ with distribution $\pi = (\pi_1, \pi_2)$ and $X|C = c \sim N(\mu_c, \Sigma)$ with μ_c the known p-component mean for class c data, and Σ the known common covariance matrix. Show that the classification rule minimizing the total risk for the 0–1 loss rule has a linear discriminant function.

(b) Suppose this classifier is used to classify subsequent draws (X, C) from the joint distribution of measurements and class. Calculate the probability of misclassification (giving your answer in terms of the CDF of a standard normal random variable and the discriminant function).

[**MSc 2007 Q8**]

4. (a) Suppose we observe $\mathcal{T} = \{x_{k,i}, \ldots, x_{k,n_k} : k \in \{1,2\}\}$ where each $x_{k,i} \in \mathbb{R}$. Assuming $p_k(x) \sim N(\mu, \sigma)$ 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 $k = 1$ and $k = 2$, show that the resulting decision boundaries of this classifier are linear.

(c) Find the plug-in estimators $\widehat{\mu}_k$ and $\widehat{\sigma}_k$ by maximum-likelihood estimation and give details of the calculation.

(d) Suppose we generate 2-dimensional data from $p_k(x) \sim N(\mu_k, \Sigma)$ for $k \in \{1,2\}$ with $x = (x^{(1)}, x^{(2)}) \in \mathbb{R}^2$ and suppose the plug-in esimators are

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

(i) Find the decision boundary as a function for $\pi_1 = \pi_2 = 1/2$ and for $\pi_1/\pi_2 = \exp(-1)$. Would you classify a new observation at $x = (0, 1/3)$ as class $k = 1$ or class $k = 2$ if indeed $\pi_1/\pi_2 = \exp(-1)$? For which values of $\pi_1 = 1 - \pi_2$ would you classify the observation $x = (1,1)$ as belonging to class 2?

(ii) What is the decision boundary under the assumption $\pi_1 = \pi_2$ if the plug-in estimators $\widehat{\mu}_k$ are as above for $k = 1, 2$ but the covariance matrix is

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

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

[**MSc 2009 Q8, parts (a), (b), (c)**]

5. Suppose we observe data (X, Y), where $X = (X^{(1)}, X^{(2)})$ is a 2-dimensional predictor variable and $Y \in \{-1, 1\}$ is a binary class variable. Conditional on $Y = 1$, the predictor variable X has a $\mathcal{N}(\mu_1, \Sigma)$ distribution. Conditional on $Y = -1$, X has a $\mathcal{N}(\mu_{-1}, \Sigma)$ distribution. Assume that the a priori probabilities are equal, i.e. $P(Y = 1) = P(Y = -1) = 0.5$, and Σ is known,

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

Upon observing data (X_i, Y_i) for $i = 1, \ldots, n$, the empirical mean of all observations of class $Y = 1$ is $\widehat{\mu}_1 = (0, 0)$, whereas the empirical mean of all observations of class $Y = -1$ is $\widehat{\mu}_{-1} = (1, 0)$.

(a) What is the classification $\widehat{Y}(X)$ of Linear Discriminant Analysis (LDA) for $\rho = 0$, as a function of the predictor variable $X = (X^{(1)}, X^{(2)})$?

(b) Find the decision boundary and classification $\widehat{Y}(X)$ for LDA for general $\rho \in [0, 1]$.

(c) What is the misclassification error rate for observations of class $Y = 1$ under the optimal LDA classification if $\rho = 1$, i.e. what is $P(\widehat{Y}(X) = -1 | Y = 1)$ in this setting? (*Hint: think about what values X can take in the 2-dimensional plane for $Y = -1$ and for $Y = 1$.*)

(d) [. . .]

[**MSc 2009 Q9**]

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

(a) Explain the difference between single-linkage, average-linkage and complete-linkage for bottom-up hierarchical clustering of p variables. What impact has the choice of the linkage on the choice of the first pair of variables that are grouped?

(b) Suppose a logistic regression is fitted to n observations (X_i, Y_i), where X is a p-dimensional predictor variable and Y a binary response variable in $\{-1, 1\}$. Consider the conditional probabilities $P(Y = 1|X)$ and $P(Y = -1|X)$. Which function of these conditional probabilities is modelled as a linear function of X? Name another classification method that leads to a linear dependence on X for the same function of the posterior probabilities?

(c) Briefly explain leave-one-out cross-validation for regression trees, given data samples (X_1, Y_1),
$\ldots, (X_n, Y_n)$ with squared error loss and give the corresponding estimate for prediction error. What is a disadvantage of leave-one-out cross-validation compared to 5-fold cross validation?

(d) Explain briefly bagging for a general estimator predictor $\widehat{Y}$ of a response variable Y, in the context of regression trees. What is the typical effect of bagging on the variance and bias of a regression tree or classification tree? Do you expect more improvements through bagging when working with classification trees or when working with Linear Discriminant Analysis?

(e) Suppose training data (X_i, Y_i), $i = 1, \ldots, n$, are available, where Y is a binary response variable and $X = (X^{(1)}, \ldots, X^{(p)})$ a p-dimensional predictor variable. Logistic regression is fitted. Then, logistic regression is fitted again to the same observations, where the predictor variable is now

$$(X^{(1)}, \ldots, X^{(p)}, X^{(p+1)}),$$

that is an additional predictor variable $X^{(p+1)}$ has been added. How do the first and second logistic regression compare in terms of the training and test error?

[**MSc 2010 Q8**]

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

(a) What is the difference between K-means clustering and online vector quantization?

(b) Explain briefly the aim of metric multi-dimensional scaling (MDS) for n samples by writing down the stress function that MDS is trying to minimize.

(c) Explain the different settings in which vector quantization and learning vector quantization are used. What is one advantage of learning vector quantization over nearest neighbour classification?

(d) What is an ROC curve in the context of binary classification with classes 0 and 1?

(e) The n given samples are (X_i, Y_i) for $i = 1, ..., n$, where each X_i is a p-dimensional predictor variable and $Y_i \in \{0, 1\}$ a binary response variable, indicating the class of the i-th observation. Explain leave-one-out cross validation of misclassification error for a classification tree. Compare leave-one-out with 5-fold cross-validation. What is a drawback of leave-one-out cross-validation in practice?

[**MSc 2010 Q9**]

8. (a) Suppose $X = (X^{(1)}, X^{(2)})$ is a two-dimensional predictor variable with both $X^{(1)} \in \{-2, 2\}$ and $X^{(2)} \in \{-2, 2\}$. Suppose that $Y \in \{0, 1\}$ is a binary response variable and that, conditional on Y, $X^{(1)}$ and $X^{(2)}$ are independent. Assume that

$$\begin{aligned} P(X^{(1)} = 2|Y = 1) &= 0.1 \\ P(X^{(2)} = 2|Y = 1) &= 0.9 \\ P(X^{(1)} = 2|Y = 0) &= 0.7 \\ P(X^{(2)} = 2|Y = 0) &= 0.3 \end{aligned}$$

and the probabilities for $X^{(1)} = -2$ and $X^{(2)} = -2$, conditional on $Y = 1$ and $Y = 0$, are the corresponding complements.

(i) Derive the Bayes classifier $\widehat{Y} = \widehat{Y}(x^{(1)}, x^{(2)})$ for an observation $(x^{(1)}, x^{(2)}) = (2, 2)$ under $P(Y = 1) = P(Y = 0) = 0.5$.

(ii) At what value of $P(Y = 1)$, if any, is it optimal to classify as $\widehat{Y}(x^{(1)}, x^{(2)}) = 0$ for all possible predictor variables $(x^{(1)}, x^{(2)})$?

(b) Explain bagging in a few sentences. What is in general the qualitative effect of bagging on the variance of a regression estimator? Would you expect greater improvements in prediction accuracy when bagging regression trees with few or with many leaf nodes?

(c) Suppose k-nearest neighbour classification with $k = 1$ is applied to $n = 4$ data samples (x_i, y_i), $i = 1, \ldots, n$:

$$(1, 1), (2, 0), (3, 1), (4, 1)$$

where, for $i = 1, \ldots, n$, $x_i \in \mathbb{R}$ is a real-valued predictor variable and $y_i \in \{0, 1\}$ is the binary response variable with the class information.

(i) Let $\widehat{Y}$ be the nearest-neighbour prediction for a new observation with predictor variable x. Give $\widehat{Y}(x)$.

(ii) Let $\widehat{Y}_{bagged}(x)$ be the prediction if bagging is applied to the nearest neighbour regression and the majority vote is taking over all nearest neighbour classifications that are obtained under bootstrapping. What is the probability that the training sample $(2, 0)$ will *not* be contained in a given bootstrap sample? Using this result, what is the bagged nearest neighbour classification for an observation at $x = 2$?

[**MSc 2011 Q4**]

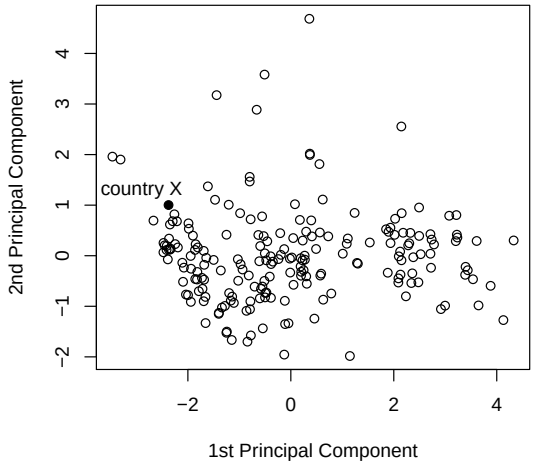
9. (a) Describe in a sentence or two the main differences between supervised and unsupervised learning.

(b) Name two different techniques for dimension reduction, and for each technique describe the criterion that is being used to define the "best" lower-dimensional representation.

(c) Data are available for 201 countries on five demographic and economic characteristics: the rate of population growth (`popgrwth`), GDP per capita in US dollars (`gdpcap`), births per 1000 persons/year (`birthrate`), proportion of workforce employed in agriculture (`agriculture`) and the rate of literacy (`literacy`). A Principal Component Analysis will be carried out to look for interesting structure in this data.

Why is it a good idea to first scale the data?

The figure below is a scatter plot of the first two principal components of the scaled data.



(i) Do you think there is correlation between the first and second principal components? Justify your answer.

(ii) Describe the interesting features of the figure.

(iii) Use the R output below to interpret the structures or features identified in (ii).

```
Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
popgrwth      0.364  0.744  0.385  0.301 -0.275
gdpcap       -0.402  0.523 -0.702  0.226  0.146
birthrate     0.520  0.128        -0.104  0.836
agriculture   0.452 -0.385 -0.351  0.706 -0.161
literacy     -0.481         0.481  0.591  0.423

Importance of components:
                       Comp.1  Comp.2  Comp.3  Comp.4  Comp.5
Standard deviation     1.788   0.916   0.689   0.566   0.408
Proportion of Variance 0.639   0.167   0.095   0.064   0.033
Cumulative Proportion  0.639   0.807   0.902   0.966   1.000
```

(iv) Do you think the point labelled `country X` in the figure is more likely to correspond to The United States of America or to Ethiopia? Explain your reasoning.

(v) Suppose that a biplot of these data were to be produced by adding the projections of the variable axes. Show with a sketch the direction of the vectors which would correspond to `popgrwth` and `birthrate`.

[**MSc 2012 Q7 (a "half-question")**]

10. In binary classification, the loss function we usually want to minimize is the risk associated with the 0/1 loss:

$$L\left(y, \widehat{y}\left(x\right)\right) = \mathbb{I}\left(\widehat{y}\left(x\right) \neq y\right)$$

where $\widehat{y}\left(x\right), y \in \{0, 1\}$. In this problem we will consider the effect of using an asymmetric loss function

$$L_{\alpha,\beta}\left(y, \widehat{y}\left(x\right)\right) = \alpha \mathbb{I}\left(y = 0, \widehat{y}\left(x\right) = 1\right) + \beta \mathbb{I}\left(y = 1, \widehat{y}\left(x\right) = 0\right)$$

where $\alpha, \beta > 0$.

(a) Determine the Bayes optimal classifier, i.e. the classifier that achieves the minimum risk for the loss function $L_{\alpha,\beta}$.

(b) Suppose $P\left(Y = 0\right)$ is very small. This means that the classifier $\widehat{y}\left(x\right) = 1$ for all x will have a small risk under the 0/1 loss function. We may try to put the two classes on even footing by considering the risk

$$R = P\left(\widehat{y}\left(X\right) = 1 \middle| Y = 0\right) + P\left(\widehat{y}\left(X\right) = 0 \middle| Y = 1\right)$$

Show that this risk can be rewritten as the expected loss function $L_{\alpha,\beta}$ for certain values of α, β.