

Mini-project assignment
MS1b Statistical Data Mining

Deadline

Mini-project assignments for *MS1b Statistical Data Mining* are due to be handed in at the Examination Schools by **12 noon on Monday 26 April 2010 (week 1 Trinity Term)**.

The details of this mini-project are on pages 2–4 below.

Mini-project MS1b Statistical Data Mining, HT 2010

Gene expression data

The 6 relevant datasets for this mini-project contain gene expression data along with tumour class information from a certain number of patients¹. For each patient one knows the response variable, which is binary (in $\{0, 1\}$), where 1 indicates presence of a specific tumour (or subtype of one) and 0 absence of this tumour. Also available are for each patient several thousand predictor variables, namely the gene expression pattern for this patient on several thousand measured genes. Using the available samples, it is of interest to build a classifier that can provide a good classification of absence/presence of tumours for new patients.

You can download the predictor variables for each of the 6 datasets (in `X1.txt`, ..., `X6.txt`) from the website

<http://www.stats.ox.ac.uk/~meinshau/MS1/miniproject>

The associated outcomes are in the files `Y1.txt`, ..., `Y6.txt` on the same site.

You should produce a report that covers the six tasks below.

Take extra care to ensure that your report is clear on how you have performed the analysis, so that an interested reader could reproduce your results. For example, if a classifier needs a choice of tuning parameters, describe how you set these tuning parameters, whether you used cross-validation (and which form of cross-validation, if so), and so on. The report *should not* contain computer code. You may (and are encouraged to) include figures.

Task 1: Bayes Classifier Describe the Bayes classifier for general two-class problems (without using the data yet). Why do we not have access to the Bayes classifier in practice? Describe *plug-in classification* and give an example for plug-in classification (again, without using the data yet).

10 marks

Task 2: Dimensionality Reduction Reduce the dimensionality of the data from p to $\tilde{p} = 2$ by Principal Component Analysis or metric Multi-Dimensional Scaling and plot the training samples in the resulting two-dimensional plane for gene-expression datasets 2 and 5. Indicate the class of each sample by plotting symbols and/or colour (you should not use this class information in the dimension reduction itself). Based on this 2-dimensional projection, for which of the two datasets would you expect to achieve a better classification (in the sense of lower misclassification error)?

¹the patients are not identical across datasets

What is the minimum number of dimensions needed to preserve 90%, 95% and 99% of the variability in the data for each of the six datasets?

15 marks

Task 3: Clustering For each of the six datasets, perform a k -means clustering of the data with $k = 2$ clusters, without using the class labels.

What would the misclassification error rate be (on each given dataset) if you classified observations in one cluster as class 0 and observations in the second cluster as class 1? If you obtain misclassification errors larger than 50%, you should switch the labels. How accurate could you be with simple random guessing of class membership on average in comparison?

15 marks

Task 4: LDA analysis Use Linear Discriminant Analysis on the 5th dataset (`X5.txt` and response `Y5.txt`).

Include the first k genes in the list in the same order as they appear in the matrix `X5.txt`. Based on these k genes, compute the LDA prediction on the data and compute the error rate. For the LDA prediction, use 10-fold cross-validation.

Plot the cross-validated error rate (the misclassification error based on the cross-validated predictions) while varying the number of genes between $k = 2, 3, 4, 5, \dots, 50$ (always taking the first k , using the same order as in the data files). How many genes should be included in the LDA analysis? Why do you think the performance possibly deteriorates for very large values of k ?

15 marks

Task 5: Classifier comparison Fit single trees and Random Forests to all datasets and explore their classification performance. In each dataset, use only the first 200 genes.

Try to use 5-fold cross-validation, if possible, to estimate error rates. Produce Receiver-operating characteristic (ROC) curves for each classifier and interpret the results. Which of the two methods would you prefer if you wanted to get the most accurate predictions on these datasets (again using only the first 200 genes)?

How many of class 0 observations are misclassified with either method in all datasets, when at most 1/3 of all class 1 are misclassified?

Using either method for the 5th dataset, find some genes which seem to be important for a good classification.

25 marks

Task 6: Naive Bayes and nearest neighbour classification In LDA, the $p \times p$ -covariance matrix $\hat{\Sigma}$ of the p predictor variables/genes is estimated from the data by maximum likelihood estimation. For ‘Naive Bayes’, only the diagonal terms of the covariance matrix are kept and all non-diagonal terms are set to 0. Find a simple description of the ‘Naive Bayes’ classification rule (assuming that each variable is standardized to mean 0 and variance 1) under the assumption that prior probabilities are equal for each class. Compare briefly with a nearest neighbour classification method under Euclidean distance, giving a possible computational advantage of ‘Naive Bayes’. What is a possible

advantage of Naive Bayes in higher dimensions compared to LDA? Compute the Naive Bayes solution for the 5th dataset when including the first $k = 2, 3, \dots, 100$ genes and compare with the LDA result you obtained in Task 4.

20 marks

[End of mini-project assignment]