

The University of Melbourne

Semester 2 Assessment, 2006

Department: Mathematics and Statistics
Subject Title: 620-472 Data Mining
Exam Duration: Three hours
Reading Time: Fifteen minutes
This paper has six (6) pages (including cover sheet)

Authorised material:

Hand-held electronic calculators may be used, provided all memories and programs are cleared.

Instructions to Invigilators:

Closed book.

Instructions to Students:

All questions may be attempted.

The number of marks for each question is indicated; this reflects the relative weighting of the questions.

The total number of marks available in this examination is 90.

Working and/or reasoning must be shown.

1. Association Rules

Let S be the set of possible purchases, $|S| = n$, and let N be the number of transactions. Each transaction is a subset of S .

- (a) Let A and B be subsets of S . Define the *support*, *confidence* and *lift* of the rule

$$A \Rightarrow B.$$

- (b) A *strong* rule has support at least s_0 and confidence at least c_0 . For $s_0 = 0.6$ and $c_0 = 0.8$, from the data below, find all the strong rules of the form $\{x_1, x_2\} \Rightarrow \{y\}$ where $x_1, x_2, y \in S$, $x_1 \neq x_2 \neq y$.

Transaction	Item Set
1	$\{a, b, d, k\}$
2	$\{a, b, c, d, e\}$
3	$\{a, b, c, e\}$
4	$\{a, b, d\}$

- (c) The a-priori algorithm for finding frequent item sets is (in pseudo-code):

```
k = 1
L1 = {A ∈ S : |A| = 1, support(A) ≥ s0}
while Lk ≠ ∅
    k = k + 1
    Ck ← a-priori-gen(Lk-1)
    Lk ← {A ∈ Ck : support(A) ≥ s0}
end while
return ∪k Lk
```

- (i) Describe, using pseudo-code or otherwise, the function a-priori-gen.
(ii) What property of L_{k-1} is required to enable a-priori-gen(L_{k-1}) to execute quickly?
(iii) What is C_4 , given that

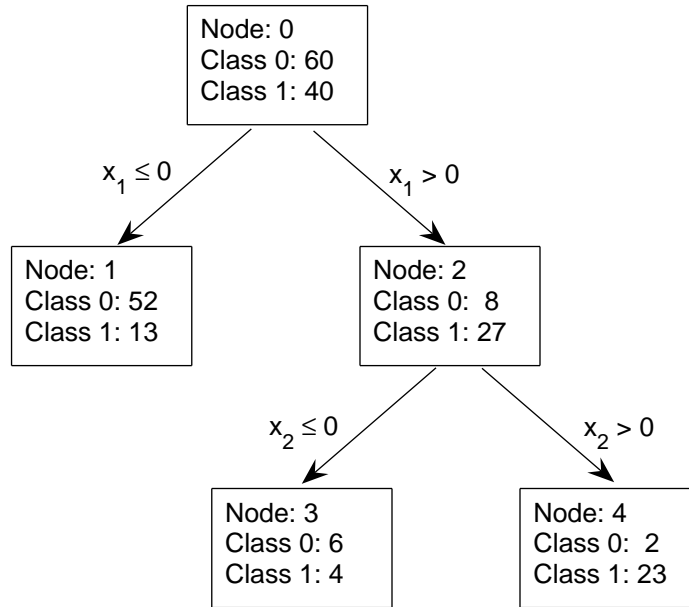
$$L_3 = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 3, 4\}, \\ \{1, 3, 5\}, \{1, 4, 5\}, \{2, 3, 4\}, \{2, 4, 5\}\}.$$

[3+3+6=12 marks]

2. Classification Trees

For $i = 1, \dots, N$ we have observations $(x_1(i), \dots, x_k(i), y(i))$ where the x_j are input variables and $y \in \{0, 1\}$ is the response.

- (a) Consider the following classification tree. For each node we indicate the number of observations with response 0 and the number with response 1.



- (i) For all possible threshold values $p \in [0, 1]$, classify each of the leaves of this tree as either class 1 or 0, according as the proportion of class 1 observations is $\geq p$ or not.
- (ii) Hence sketch the ROC curve for this decision tree. Be sure to label both axes.
- (b) Let $f(i, j)$ be the frequency ($\in [0, 1]$) of class j observations in node i and let $I(i)$ be the impurity of node i .
- (i) Define the Gini impurity (used by CART) and the Entropy impurity (used by C4.5).
- (ii) Let $s(i)$ be the size of node i . Define the *gain* in purity made by splitting node i into nodes i_0 and i_1 .
- (iii) For the classification tree given in part (a), calculate the gain made by splitting node 2 into nodes 3 and 4, using the Gini impurity.
- (c) (i) Define the *complexity* of a classification tree and explain how it is used to *prune* a tree. In particular explain the role of the *complexity parameter*.
- (ii) Explain how *cross-validation* can be used to choose the complexity parameter optimally.

[8+5+7=20 marks]

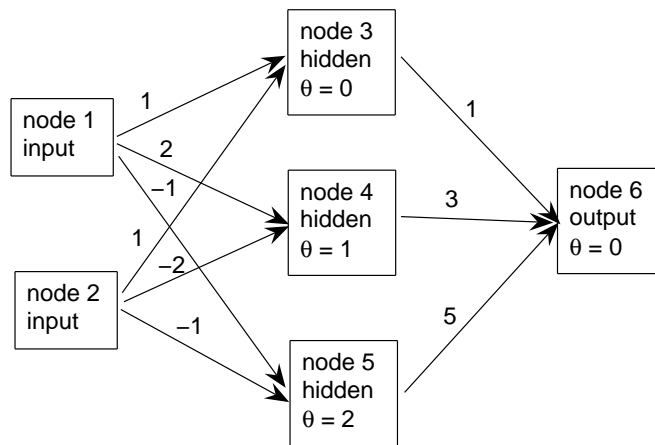
3. Neural Nets

- (a) Consider the following feed-forward neural network (FFNN) with edge weights and thresholds θ as indicated. The hidden nodes (3, 4 and 5) all have a stepwise activation function of the form

$$f(x, \theta) = \begin{cases} 0 & x < \theta \\ 1 & x \geq \theta \end{cases}$$

and the output node (6) has a linear activation function

$$f(x, \theta) = x - \theta.$$



- (i) If the values of the input nodes 1 and 2 are fixed at 1 and -1 respectively, what are the values at all the other nodes (after the network is updated)?
- (ii) Fix the value at node 1 at 1, then sketch a graph of the value y of the output node 6 against the value x of the input node 2, for $x \in [-2, 2]$.
- (b) (i) Explain what overfitting is.
- (ii) In the context of using the backpropagation algorithm to fit a feed-forward neural network, explain how you would use a *test data set* to detect overfitting. Your explanation should include a definition of the test data set.

[4+6=10 marks]

4. Logistic Regression

For $i = 1, \dots, N$ we have observations $(x_1(i), \dots, x_k(i), y(i))$ where the x_j are input variables and $y \in \{0, 1\}$ is the response. For simplicity of notation we assume that $x_1(i) = 1$ for all i . The logistic regression model is then

$$\mathbb{P}(Y(i) = 1 | \mathbf{X} = \mathbf{x}(i)) = \mu(\mathbf{x}(i) | \boldsymbol{\beta}) := \frac{\exp\{\sum_j \beta_j x_j(i)\}}{1 + \exp\{\sum_j \beta_j x_j(i)\}}.$$

- (a) Give an expression for the *likelihood* of the model parameters $\boldsymbol{\beta}$ given the observations.
- (b) We can find a local maximum of the log-likelihood iteratively using the rule

$$\boldsymbol{\beta}^{(n+1)} = (X^T W X)^{-1} X^T W (X \boldsymbol{\beta}^{(n)} + W^{-1} (Y - P)) \quad (1)$$

where $\boldsymbol{\beta}^{(n)}$ is the n -th approximate solution and

$$\begin{aligned} X(i, j) &= x_j(i) \\ Y(i) &= y(i) \\ P(i) &= \mu(\mathbf{x}(i) | \boldsymbol{\beta}^{(n)}) \end{aligned}$$

- (i) What is W ? (You may wish to complete part (ii) before answering this.)
 - (ii) Show that (1) can be obtained by applying the Newton-Raphson algorithm to the log-likelihood. (The Newton-Raphson algorithm has the form $\boldsymbol{\beta}^{(n+1)} = \boldsymbol{\beta}^{(n)} - H^{-1} \nabla f(\boldsymbol{\beta}^{(n)})$ where f is the objective function and H the Hamiltonian.)
- (c) (i) Explain how the log-likelihood ratio can be used to choose between two logistic regression models.
 - (ii) Why would we want to remove statistically insignificant variables from a logistic regression model?

[2+8+6=16 marks]

5. Additive Models

Suppose that $Y = f(\mathbf{X}) + E$ where $\mathbb{E}E = 0$ and $\text{Var } E = \sigma^2 < \infty$. We observe inputs $\mathbf{x}(1), \dots, \mathbf{x}(N) \in \mathbb{R}^k$ and responses $y(1), \dots, y(N) \in \mathbb{R}$.

- (a) Show that if we choose f to minimise $\mathbb{E}\|Y - f(\mathbf{X})\|^2$ then we get $f(\mathbf{X}) = \mathbb{E}(Y | \mathbf{X})$. (Note that $\|\mathbf{a} + \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + 2 \langle \mathbf{a}, \mathbf{b} \rangle + \|\mathbf{b}\|^2$.)
- (b) What is the general form of an additive regression model?
- (c) We use penalised least squares to fit additive models, rather than ordinary least squares. What is penalised least squares, and what is the effect of the penalty?
- (d) Explain the *curse of dimensionality* in the regression context. How does an additive regression model deal with the curse of dimensionality?

- (e) Using pseudo-code or otherwise, give the backfitting algorithm of Hastie and Tibshirani for fitting an additive regression model. You may assume that you are given a function *smooth* which takes as input $u(1), \dots, u(N)$ and $v(1), \dots, v(N)$ and returns a function f such that $f(u(i)) \approx v(i)$. (You do not need to explain how *smooth* works.)

[4+2+3+3+4=16 marks]

6. Clustering

Consider the K -means algorithm in \mathbb{R}^d , with data $\mathbf{x}(1), \dots, \mathbf{x}(N)$. Let $C(i)$ be the cluster containing $\mathbf{x}(i)$, \mathbf{m}_k the (current) centroid for cluster k and N_k the size of cluster k . The total squared error is

$$S := \sum_{k=1}^K N_k \sum_{i:C(i)=k} \|\mathbf{x}(i) - \mathbf{m}_k\|^2.$$

Now take the following set of points in \mathbb{R}^1 : $\{6, 12, 18, 24, 30, 42, 48\}$.

- (a) For each of the following sets of centroids, create two clusters by assigning each point to the nearest centroid, then calculate the total squared error S .
- (i) $\{18, 45\}$
 - (ii) $\{15, 40\}$

Which pair of clusters gives the smallest value of S ?

- (b) For each case in part (a), if you ran the K -means algorithm using this set of points as starting centroids, what clusters would you end up with? Comment on your result.
- (c) Using hierarchical clustering with Euclidean distance and single linkage, what sequence of clusters are obtained? Sketch the dendrogram.
- (d) (i) Using pseudo-code or otherwise, describe the K -means algorithm for assigning clusters.
- (ii) Let $\bar{\mathbf{x}}_k$ be the mean over cluster k , then prove that

$$\begin{aligned} & \min_{\text{clusters}} \sum_{k=1}^K N_k \sum_{i:C(i)=k} \|\mathbf{x}(i) - \bar{\mathbf{x}}_k\|^2 \\ &= \min_{\text{clusters}; \mathbf{m}_1, \dots, \mathbf{m}_K} \sum_{k=1}^K N_k \sum_{i:C(i)=k} \|\mathbf{x}(i) - \mathbf{m}_k\|^2. \end{aligned}$$

[3+4+4+5=16 marks]