

The University of Melbourne

Semester 2 Assessment, 2007

Department: Mathematics and Statistics

Subject Title: 620-472 Data Mining

Exam Duration: Three hours

Reading Time: Fifteen minutes

This paper has five (5) 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 85.

Working and/or reasoning must be shown.

1. **Association Rules** Let S be a finite set and Ω a collection of transaction records (each record is a subset of S).

- (a) Explain what an *association rule* is, then define the *support*, *confidence* and *lift* of an association rule.
- (b) By defining a suitable probability measure on Ω , give a probabilistic interpretation of the support, confidence and lift of an association rule.

[3 + 3 = 6 marks]

2. **Classification Trees**

- (a) Name some advantages and disadvantages of classification trees as classifiers?
- (b) Does pruning a classification tree increase or decrease performance sometimes/always/on-average for
 - (i) the training data set, and
 - (ii) the test data set?
- (c) The following data consists of inputs x_1 and x_2 and output y :

x_1	x_2	y
red	5.1	0
red	0.8	1
red	6.6	0
red	7.7	1
red	1.3	1
blue	4.6	1
blue	6.0	1
blue	4.6	0
yellow	7.4	0
yellow	5.9	0

Suppose that the first split is on x_1 . Using *gini* impurity, calculate the *gain* for each possible split of x_1 , and hence identify which split you should make.

- (d) Explain the trick for optimally splitting a continuous variable with feasible computation.

[3 + 2 + 6 + 3 = 14 marks]

3. **Logistic Regression** Suppose that we have inputs $\mathbf{x}(i) \in \mathbb{R}^d$ and outputs $y(i) \in \{0, 1\}$, for $i = 1, \dots, n$.

- (a) What is a logistic regression model for y ?
- (b) How do you fit a logistic regression model? Express the fitting as an optimisation problem, then suggest a numerical technique for solving it.
- (c)
 - (i) Explain what overfitting is and why it is a problem.
 - (ii) Define the Akaike Information Criteria (AIC) and explain how it is used to perform parameter selection for a logistic regression model. Include a description of how you would search the parameter space.
- (d) Explain how each of the following techniques could be used to avoid overfitting by a logistic regression model:
 - (i) Test data set;
 - (ii) Cross validation;
 - (iii) Penalised maximum likelihood.

[2 + 3 + 6 + 8 = 19 marks]

4. **Optimisation** Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$.

- (a) Define the gradient-descent method for finding a local minimum of f .
- (b)
 - (i) Define Newton's method for finding a local minimum of f .
 - (ii) Give a derivation of Newton's method.
- (c) Suppose that $\mathbf{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ are dependent random variables. Show that the function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ which minimises $\mathbb{E}(Y - g(\mathbf{X}))^2$ is $g(\mathbf{x}) = \mathbb{E}(Y | \mathbf{X} = \mathbf{x})$.

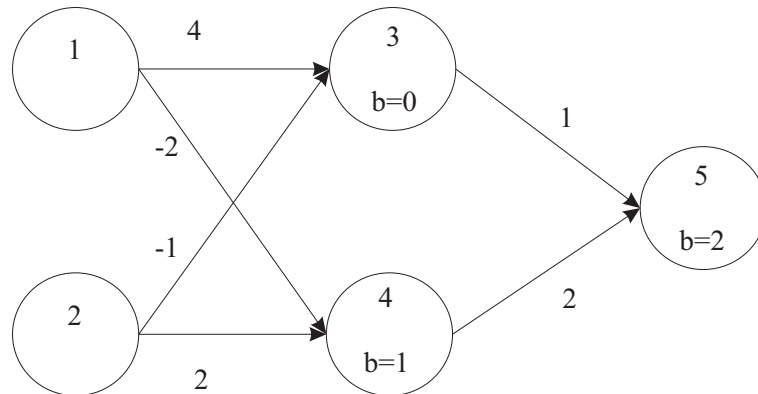
[2 + 5 + 4 = 11 marks]

5. Neural Nets

- (a) The neural net below uses the linear activation function at node 5 and the following activation function at nodes 3 and 4:

$$f(x) = \begin{cases} 1, & 1 < x \\ x, & -1 < x \leq 1 \\ -1, & x \leq -1 \end{cases}$$

Biases are given by b .



Suppose that the input at node 1 is x and at node 2 is y , for $x, y \in \mathbb{R}$.

- (i) What is the range of output values produced by this neural net?
 - (ii) What values of x and y achieve the maximum output value?
- (b) (i) Specify the architecture, activation functions and loss/error function for a neural net to be equivalent to linear regression.
- (ii) Specify the architecture, activation functions and loss/error function for a neural net to be equivalent to logistic regression.

[6 + 6 = 12 marks]

6. Regression

- (a) What is the curse of dimensionality, and how does a Generalised Additive Model (GAM) avoid it?
- (b) A Multivariate Adaptive Regression Spline (MARS) fits a regression function from a class \mathcal{S} of functions.
 - (i) Specify \mathcal{S} , then explain how the MARS algorithm selects an element from it. That is, explain how the MARS algorithm works.
 - (ii) How does the MARS algorithm avoid overfitting?

[4 + 10 = 14 marks]

7. **Clustering** Suppose $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$.

- (a) Give the k -means algorithm.
- (b) What value of \mathbf{a} minimises $\sum_{i=1}^n \|\mathbf{x}_i - \mathbf{a}\|^2$? Justify your answer.
- (c) Show that for clusters K_1, \dots, K_k

$$\min_{K_1, \dots, K_k} \sum_{i=1}^k \sum_{j \in K_i} \|\mathbf{x}_j - \bar{\mathbf{x}}_i\|^2 = \max_{K_1, \dots, K_k} \sum_{i=1}^k N_i \|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}\|^2,$$

where N_i is the size of cluster i , $\bar{\mathbf{x}}_i$ is the centroid of cluster i and $\bar{\mathbf{x}}$ is the centroid of the whole sample.

[3 + 3 + 3 = 9 marks]