

MONASH UNIVERSITY

Faculty of Information Technology

CSE5230 Data Mining

Semester 2, 2004

Preliminary Quiz - Basic Mathematics

Data Mining is a multidisciplinary field which brings together a wide variety of techniques from areas of research and development with longer histories: machine learning, pattern recognition, statistics, databases and visualisation. The aim is to extract *knowledge* from the raw information stored in large databases, with the aim of better describing or understanding the existing data, or predicting how new data will be generated in the future.

Every data mining technique has a mathematical basis. At the end of this unit you should be able to describe the algorithms underlying the most common state-of-the-art data mining tools, and make an informed choice of data mining tool for a given problem. Although you will not be required to work mathematical problems in the assessment of this unit, a grasp of the basic mathematics underlying the data mining techniques discussed in lectures is essential.

This quiz is designed to allow you to assess your own chances of success in this unit. It does *not* form part of the assessment for the unit. If you cannot answer most of the questions on this quiz, you will find this unit very challenging. Solutions to this quiz will be discussed in tutorials in week 2.

Questions

- Let N be the number of items bought by a shopper during a visit to a supermarket. Ten shoppers visit the supermarket, and the number of items bought by each shopper is recorded: $\{4, 3, 4, 10, 2, 23, 15, 6, 4, 22\}$.
 - What is the sample **mean** of N ?
 - What is the sample **variance** of N ?
- Consider two random events A and B , which occur with probabilities $P(A) = 0.5$ and $P(B) = 0.8$.
 - Assuming that A and B are **independent** events, what is the probability that both A and B occur, $P(A \cap B)$?
 - Now consider the case where A and B are not independent, and $P(A \cap B) = 0.2$. What is the probability that A will occur given that B has occurred, $P(A|B)$?

3. Consider the following three points in a two-dimensional space: $X_1 = (1, 1)$, $X_2 = (4, 5)$ and $X_3 = (6, 3)$.
 - (a) What is the Euclidean distance (L_2 norm) between points X_1 and X_2 ?
 - (b) Consider a fourth point $X_4 = (2, 3)$. To which of the first three points is X_4 closest in Euclidean distance?

4. Consider the following two points in a four dimensional space: $Y_1 = (2, 5, 3, 6)$ and $Y_2 = (6, 4, 7, 8)$.
 - (a) What is the Euclidean distance between points Y_1 and Y_2 ?
 - (b) What is the Manhattan distance (L_1 norm) between points Y_1 and Y_2 ?

5. Consider two random variables A and B . Ten measurements are made of A and B , the results of each measurement being written as a tuple (value of A , value of B). The observed values are: $\{(0.4, 0.2), (0.8, 0.9), (0.1, 0.1), (0.4, 0.3), (0.9, 1.1), (0.4, 0.2), (0.5, 0.4), (0.7, 0.6), (0.6, 0.7), (0.2, 0.3)\}$.
 - (a) What is the **expected value** of A , $E(A)$?
 - (b) What is the expected value of B , $E(B)$?
 - (c) What is the expected value of $A + B$, $E(A + B)$?
 - (d) What is the **covariance** between A and B , $\text{Cov}(A, B)$?
 - (e) Are A and B **correlated**? If so, are they positively or negatively correlated? (Hint: plotting a graph of A against B may be helpful here.)
 - (f) What is the **correlation coefficient** of A and B , $\rho_{A,B}$?

6. Consider the equation

$$y = x^3 + 2x + 4.$$

What is the derivative of y with respect to x , $\frac{dy}{dx}$?

7. Consider the function

$$f(x, y) = ax^2 + by^3 + cxy^2 + dx^2y + ex + fy + gxy + h.$$

(a) What is the **partial derivative** of f with respect to x , $\frac{\partial f}{\partial x}$?

(b) What is the partial derivative of f with respect to y , $\frac{\partial f}{\partial y}$?

8. Consider the equation

$$y_j = \sum_{i=0}^N w_{ij}x_i.$$

For a given k , $0 \leq k \leq N$, what is the partial derivative of y_j with respect to x_k , $\frac{\partial y_j}{\partial x_k}$?