
**DUBLIN INSTITUTE OF TECHNOLOGY**

# DT228A/1 MSc. in Computing
# DT228B/1 MSc. in Computing
# DT228B/2 MSc. in Computing

**SUMMER EXAMINATIONS 2015/2016**

## MACHINE LEARNING [SPEC9270]

DR. JOHN MCAULEY
DR. DEIRDRE LILLIS
DR. AISLING O'DRISCOLL

WEDNESDAY 18TH MAY 4.00 P.M. – 6.00 P.M.

TWO HOURS

Answer any **two** questions.
All questions carry **equal** (50) marks.

1. (a) You work for a large online retailer who has asked you to build a recommender system to improve their sales. You are given a large binary dataset that has been created showing customers purchases over the last ten years.

The table below shows an example of two customers A and B purchasing behaviour. 1 indicates that the customer has purchased the item while 0 indicates that the customer has not purchased the item.

| ID | Item 107 | Item 498 | Item 5645 | Item 7256 | Item 1762 | Item 28063 | Item 75328 |
|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| B | 1 | 0 | 0 | 0 | 0 | 1 | 1 |

This second table shows the purchasing behaviour for a third customer Q. Again, 1 indicates that the customer has purchased an item and 0 indicates that the customer has not purchased the item.

| ID | Item 107 | Item 498 | Item 5645 | Item 7256 | Item 1762 | Item 28063 | Item 75328 |
|---|---|---|---|---|---|---|---|
| Q | 1 | 0 | 1 | 1 | 0 | 0 | 0 |

(i) Given that there are over 100,000 items available in the store which of following models of similarity is most appropriate for this domain?

- Russell-Rao
- Sokal-Michener
- Jaccard

Give an explanation for your choice.

**[6 marks]**

(ii) Assuming that the system will recommend to person **Q** the items that the person most similar to person **Q** has already purchased but that person **Q** has not bought, which item or items will the system recommend to person **Q**? Support you answer by showing your calculations and explaining your analysis of the results. Assume that the recommender system uses the similarity metric you selected in Part (i) of this question.

**[8 marks]**

(b) Binning is a technique used to convert a continuous feature into discrete features. Two common approaches are equal-width and equal-frequency binning.

(i) What is the trade-off (difference) between using a very small number of bins and a very large number of bins?

**[5 marks]**

(ii) What is the difference (potential advantage or disadvantage) between equal-width and equal-frequency binning?

**[5 marks]**

(iii) Given the following dataset, sort each instance into a set of bins using equal-width and equal-frequency (3 bins), show your workings.

[2, 10, 22, 167, 90, 19, 26, 7, 107, 23, 16, 93, 139, 129, 3]

**[10 marks]**

**(c)** Often when working with very large datasets there is a requirement to sample the data when developing an ABT. Describe the difference between **top sampling**, **random sampling** and **stratified sampling**.

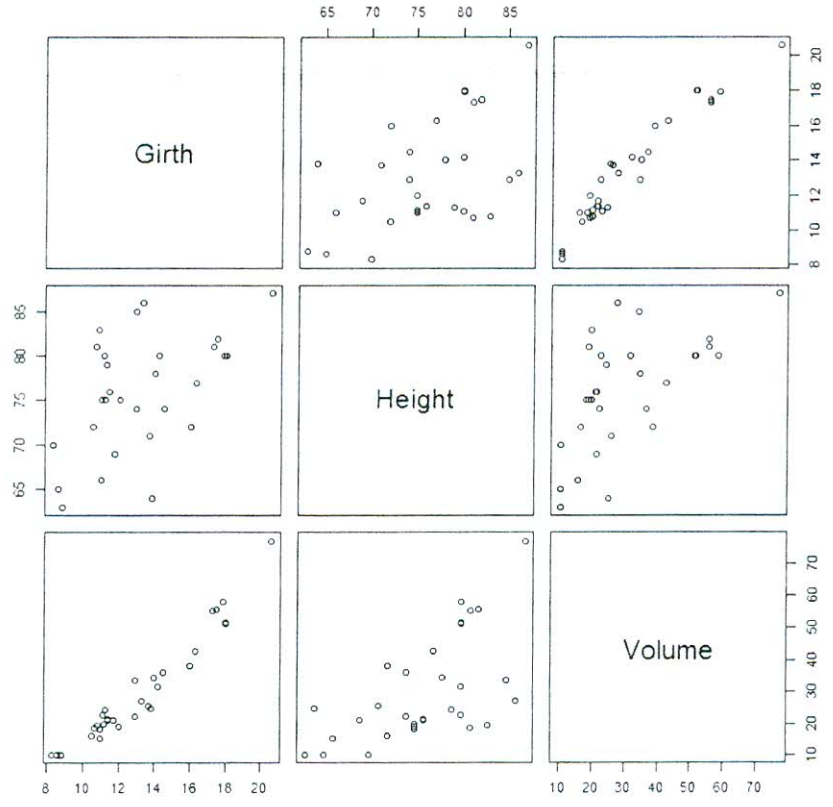
**[12 marks]**

**(d)** For the following dataset calculate the **mean**, **median**, **mode** and the **range**. Show your workings.

[2, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20]

**[4 marks]**

2. (a) The following **scatter plot matrix** (SPLOM) shows the relationships between three continuous features within a trees dataset that has 100 features. From a **data exploration** perspective, what can be learned from this visualisation? What feature engineering could you apply before training a model on this data?




**[10 marks]**

**(b)** You have been given the following dataset with 5 instances from a HR department:

**Age**: 22, 24, 26, 29, 30, 33
**Salary**: 20000, 22000, 20000, 26000, 95000, 95000

**(i)** Why could this dataset present a problem when developing a predictive model using similarity-based measure such as Euclidean distance?

**[4 marks]**

(ii) Using range normalisation, please normalise each instance between 0 and 1. Show your workings.

**[6 marks]**

(c) Inductive machine learning is often referred to as an **ill-posed problem**. Explain why this is the case and discuss the implications that follow from it. In your answer please consider the consistency of a model with a dataset, the generalisability of a model beyond the training dataset, how one model is selected over another model for a given dataset and how different machine learning algorithms make this selection.

**[20 marks]**

(d) What is the difference between supervised and unsupervised learning? Please give an example of each.

**[10 marks]**

3 (a) During a game of scrabble, you received the following 9 letters.


Please answer the following questions, each question carries equal marks.

(i) What is the entropy in bits of the letters in this set? Please show your workings.

**[5 marks]**

(ii) What would be the reduction in entropy (i.e. the information gain) in bits if we split these letters into two sets, one containing the vowels and the other containing the consonants?

**[10 marks]**

(b) A convicted criminal who reoffends after release is known as a *recidivist*. The table below lists a dataset that describes prisoners released on parole, and whether they reoffended within two years of release.

| ID | GOOD BEHAVIOR | AGE < 30 | DRUG DEPENDENT | RECIDIVIST |
|---|---|---|---|---|
| 1 | false | true | false | true |
| 2 | false | false | false | false |
| 3 | false | true | false | true |
| 4 | true | false | false | false |
| 5 | true | false | true | true |
| 6 | true | false | false | false |

This dataset lists six instances where prisoners were granted parole. Each of these instances are described in terms of three binary descriptive features (GOOD BEHAVIOR, AGE < 30, DRUG DEPENDENT) and a binary target feature, RECIDIVIST. The GOOD BEHAVIOR feature has a value of *true* if the prisoner had not committed any infringements during incarceration, the AGE < 30 has a value of *true* if the prisoner was under 30 years of age when granted parole, and the DRUG DEPENDENT feature is *true* if the prisoner had a drug addiction at the time of parole. The target feature, RECIDIVIST, has a *true* value if the prisoner was arrested within two years of being released; otherwise it has a value of *false*.

**(i)** What is the entropy for the entire dataset?

**[5 marks]**

**(ii)** What is the remaining entropy and information gain for each of the descriptive features?

**[15 marks]**

**(iii)** What would be the root node of a decision tree built using ID3?

**[5 marks]**

**(c)** Over fitting can be a problem with Decision Tree Induction. Describe **two techniques** that can be used to address over fitting.

**[10 marks]**

## Appendix A

| | |
|---|---|
| Euclidean distance | $d(x_1, x_2) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_1) - a_r(x_2)\right)^2}$ |
| Entropy | $H\left(P(v_1), \ldots, P(v_n)\right) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i)$ |
| Remaining Entropy | $rem(d, D) = \Sigma_{l \in levels\,(d)} \frac{\lvert D_{d=l} \rvert}{\lvert D \rvert} \; x \; H(t\,, D_{d=l})$ |

# Appendix B

## Table of Base 2 Logs for Different Fractions

| $\log_2(a/b)$ | | a | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| b | 1 | 0.00 | | | | | | | | | | | | | |
| | 2 | -1.00 | 0.00 | | | | | | | | | | | | |
| | 3 | -1.58 | -0.58 | 0.00 | | | | | | | | | | | |
| | 4 | -2.00 | -1.00 | -0.42 | 0.00 | | | | | | | | | | |
| | 5 | -2.32 | -1.32 | -0.74 | -0.32 | 0.00 | | | | | | | | | |
| | 6 | -2.58 | -1.58 | -1.00 | -0.58 | -0.26 | 0.00 | | | | | | | | |
| | 7 | -2.81 | -1.81 | -1.22 | -0.81 | -0.49 | -0.22 | 0.00 | | | | | | | |
| | 8 | -3.00 | -2.00 | -1.42 | -1.00 | -0.68 | -0.42 | -0.19 | 0.00 | | | | | | |
| | 9 | -3.17 | -2.17 | -1.58 | -1.17 | -0.85 | -0.58 | -0.36 | -0.17 | 0.00 | | | | | |
| | 10 | -3.32 | -2.32 | -1.74 | -1.32 | -1.00 | -0.74 | -0.51 | -0.32 | -0.15 | 0.00 | | | | |
| | 11 | -3.46 | -2.46 | -1.87 | -1.46 | -1.14 | -0.87 | -0.65 | -0.46 | -0.29 | -0.14 | 0.00 | | | |
| | 12 | -3.58 | -2.58 | -2.00 | -1.58 | -1.26 | -1.00 | -0.78 | -0.58 | -0.42 | -0.26 | -0.13 | 0.00 | | |
| | 13 | -3.70 | -2.70 | -2.12 | -1.70 | -1.38 | -1.12 | -0.89 | -0.70 | -0.53 | -0.38 | -0.24 | -0.12 | 0.00 | |
| | 14 | -3.81 | -2.81 | -2.22 | -1.81 | -1.49 | -1.22 | -1.00 | -0.81 | -0.64 | -0.49 | -0.35 | -0.22 | -0.11 | 0.00 |