Beijing-Dublin International College

SEMESTER 1 FINAL EXAMINATION - (2017/2018)

School of Computer Science

COMP3010J Machine Learning

Prof. Pádraig Cunningham
Dr. Vivek Nallur*

Time Allowed: 120 minutes

Instructions for Candidates:

Answer succinctly and to the point

BJUT Student ID:_____ **UCD Student ID:**_____

I have read and clearly understand the Examination Rules of both Beijing University of Technology and University College Dublin. I am aware of the Punishment for Violating the Rules of Beijing University of Technology and/or University College Dublin. I hereby promise to abide by the relevant rules and regulations by not giving or receiving any help during the exam. If caught violating the rules, I accept the punishment thereof.

Honesty Pledge:_____ **(Signature)**

Instructions for Invigilators

Non-programmable calculators are permitted.
No rough-work paper is to be provided for candidates.

Short Questions

1. The k-Means clustering technique is very sensitive to the value of k. How can we convince ourselves that the chosen value of k is good? What other clustering method would you recommend if the user continually wants to change the levels of granularity at which clusters are shown? **(4)**

2. You have a medical dataset containing symptoms for which patients have tested either positive or negative (binary values). While trying to find similar patients, what metric would you use. Why? Give the formula used for the metric that you recommend. **(4)**

3. Why does the kNN classification encounter difficulties when the input data has an extremely high number of features? **(3)**

4. Why is **conditional independence** an important assumption to classification using Bayes' theorem? What would happen if we make no assumptions regarding independence? **(3)**

5. The ID3 algorithm recursively builds a decision tree. Under what conditions does it stop and construct a leaf node? **(3)**

6. Why is Machine-Learning considered an *ill-posed* problem? **(2)**

7. What is the **Analytics Base Table**? Describe its structure. **(2)**

8. What is a *data quality issue*, in the context of an ABT? List three common data quality issues and give strategies for dealing with any two? **(5)**

9. Precision gives us the proportion of *True Positives* among *True* and *False Positives*. Why do we then need to calculate *Recall*? Give an example **(4)**

10. **Hold-out** sampling is easy to understand and makes intuitive sense. However, it is not perfect. Give two reasons which may cause hold-out sampling to result in a poor evaluation of your machine learning technique. Suggest a technique that tries to address these issues. **(3)**

(Sub-total for the section: 33)

Long Questions

1. Consider the following three data items (d_1, d_2, and d_3). Give the formula for calculating the cosine similarity between the vectors representing the items, and calculate the cosine similarity between each pair. State which pair of items are closest.

$$d_1 = \langle Age = 2, Height = 142, Weight = 47, Income = 0 \rangle$$
$$d_2 = \langle Age = 52, Height = 185, Weight = 98, Income = 50 \rangle$$
$$d_3 = \langle Age = 37, Height = 175, Weight = 68, Income = 44 \rangle$$

(7)

2. Consider the following description of the causal relationship between burglars, cats, storms and house alarms.

> Stormy nights are rare. Burglary is also rare, and if it is a stormy night, burglars are likely to stay at home. Cats don't like storms either, and therefore if there is a storm, cats like to go inside. House alarms are set to be triggered if a burglar breaks into a house, but sometimes they are set off by cats coming into the house. Also, sometimes house alarms are not triggered even if a burglar breaks in (the alarm could be faulty or the burglar could be very skilful).

(a) Draw a Bayesian network that encodes these causal relationships **(2)**

(b) Using the data given (Figure 1), create the conditional probability tables for the network that you created in the previous sub-question. **(6)**

ID	STORM	BURGLAR	CAT	ALARM
1	false	false	false	false
2	false	false	false	false
3	false	false	false	false
4	false	false	false	false
5	false	false	false	true
6	false	false	true	false
7	false	true	false	false
8	false	true	false	true
9	false	true	true	true
10	true	false	true	true
11	true	false	true	false
11	true	false	true	false
13	true	true	false	true

Figure 1: Burglars, cats, and house alarms dataset

(c) What will the Bayesian network predict for the ALARM given that there is both, a burglar and a cat in the house, but there is no storm? **(2)**

3. Consider the dataset that shows the Oxygen consumption amongst astronauts, while performing five minutes of intense physical activity. A regression model has been built that predicts the amount of oxygen consumed, given the descriptive features: *age* and *heart rate*.

ID	OXYCON	AGE	HEART RATE	ID	OXYCON	AGE	HEART RATE
1	37.99	41	138	7	44.72	43	158
2	47.34	42	153	8	36.42	46	143
3	44.38	37	151	9	31.21	37	138
4	28.17	46	133	10	54.85	38	158
5	27.07	48	126	11	39.84	43	143
6	37.85	44	145	12	30.83	43	138

Figure 2: Dataset showing Oxygen consumption amongst astronauts along with their age and heart rate

The regression model is given by equation 1:

$$OxyCon = w[0] + w[1] * Age + w[2] * HeartRate \quad (1)$$

Assuming that the current weights in the regression model are as follows: $w[0] = -59.50$, $w[1] = -0.15$, $w[2] = 0.60$, the prediction for each training instance is as follows:

ID	OXYCON	AGE	HEART RATE	Prediction
1	37.99	41	138	17.15
2	47.34	42	153	26.00
3	44.38	37	151	25.55
4	28.17	46	133	13.40
5	27.07	48	126	8.90
6	37.85	44	145	20.90
7	44.72	43	158	28.85
8	36.42	46	143	19.40
9	31.21	37	138	17.75
10	54.85	38	158	29.60
11	39.84	43	143	19.85
12	30.83	43	138	16.85

Figure 3: Prediction using the current weights of the regression model

(a) Calculate the sum of squared errors and the `errorDelta` for each weight, from the set of predictions given in Figure 3. **(4)**

(b) Assuming a learning rate of 0.000002, calculate the weights at the next iteration of the gradient descent algorithm. **(3)**

(c) Calculate the new sum of squared errors for the predictions generated using the new weights calculated. **(3)**

(Sub-total for the section: 27)

Total marks for the paper: 60