Beijing-Dublin International College

SEMESTER 1 RE-SIT EXAMINATION - (2017/2018)

School of Computer Science

COMP3010J Machine Learning

Prof. Pádraig Cunningham
Dr. Vivek Nallur*

Time Allowed: 120 minutes

Instructions for Candidates:

Answer all questions concisely 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. What is the Receiver Operating Characteristic Curve? How does it help with evaluating a classification method? Explain with reference to the *reference line.* **(5)**

2. What is meant by *binning*? Describe two ways of binning data. Which one would you choose on a dataset relating age to income? Why? **(8)**

3. Explain the idea of *inductive bias* in machine learning. What is the inductive bias present in the *kNN* method of classification? **(4)**

4. Give the formula for Jaccard index. List two domains where you would recommend its usage and explain why? **(6)**

5. The ID3 algorithm recursively builds a decision tree. Under what conditions does it stop and construct a leaf node? **(6)**

6. Explain the difference(s) between *k-means* and *k nearest neighbour* algorithm **(3)**

7. Explain the difference between *Type I* and *Type II* errors. Give one example of each. **(4)**

8. Explain the F1-Measure and how it is calculated? What kinds of machine learning algorithm are they most useful to evaluate? In what kind of domains would you use it? **(6)**

9. In an *Analytics Base Table*, what is meant by a `data quality issue`? Give three examples of a data quality issue and two ways of handling these issues **(6)**

10. A metric that critically affects Decision Trees is the feature selection metric. List two ways of feature selection and their respective formula? **(6)**

11. Why is the Bayes' Classifier called Naïve? What is the Naïve assumption made by the classifier? **(1)**

(Sub-total for the section: 55)

Long Questions

1. A multivariate logistic regression model has been built to predict the propensity of shoppers to perform a repeat purchase of a free gift that they are given. The descriptive features used by the model are the age of the customer, the socio-economic band to

which the customer belongs (A, B, or C), the average amount of money the customer spends on each visit to the shop, and the average number of visits the customer makes to the shop per week. Figure 1 shows the weights of the trained model. Using the model given, make a prediction for each of the query instances given below: You can make the following assumptions:

- The positive level is yes
- The classification threshold is 0.65

Feature	Weight
Intercept ($\mathbf{w}[0]$)	-3.82398
Age	-0.02990
Socio Economic Band B	-0.09089
Socio Economic Band C	-0.19558
Shop Value	0.02999
Shop Frequency	0.74572

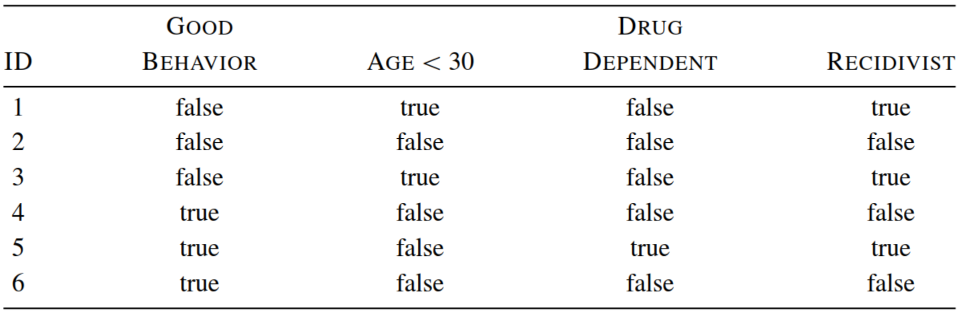
Figure 1: Table of weights of the trained model

		Socio Economic	Shop	Shop
ID	Age	Band	Frequency	Value
1	56	b	1.60	109.32
2	21	c	4.92	11.28
3	48	b	1.21	161.19
4	37	c	0.72	170.65
5	32	a	1.08	165.39

Figure 2: Query Instances

(15)

2. A convicted criminal who re-offends within two years of being released from prison is called a *recidivist*. The dataset shown in Figure 3 lists instances where prisoners were released, and whether they were recidivists or not. The dataset contains three descriptive features (`Good Behaviour`, `Age`$<$ 30, and `Drug Dependent`) and a target feature (`recidivist`). Using the dataset, construct a decision tree using the **ID3** algorithm, using entropy-based information gain. Show all steps involved in the process. Draw the final tree resulting from the computation.

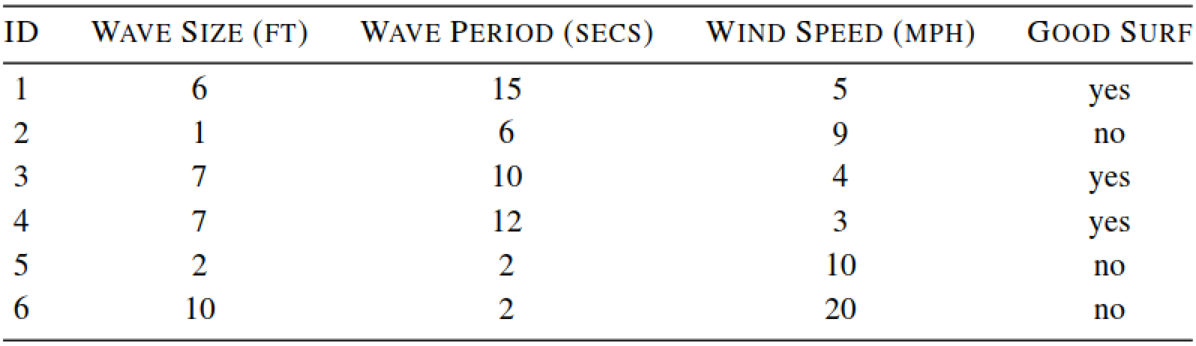


ID	GOOD	AGE < 30	DRUG	RECIDIVIST
	BEHAVIOR		DEPENDENT	
1	false	true	false	true
2	false	false	false	false
3	false	true	false	true
4	true	false	false	false
5	true	false	true	true
6	true	false	false	false

Figure 3: Recidivism Dataset

(15)

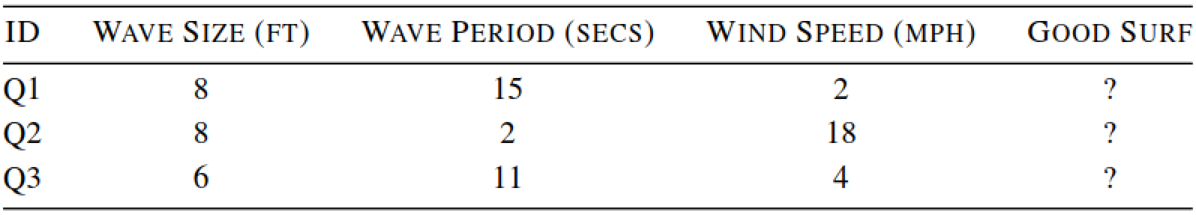
3. Surfing is a water-sport that is affected by multiple factors, such as wind speed, size of waves, etc. Figure 4 shows a dataset that was used to create a *k nearest neighbour* model. The model predicts whether it would be a good day to go surfing or not.



ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

Figure 4: Surfing Dataset

What will the model predict for each of the query instances shown in Figure 5? Assume that the model was built using Euclidean Distance. Show the distance calculations for your answers.



ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
Q1	8	15	2	?
Q2	8	2	18	?
Q3	6	11	4	?

Figure 5: Query Instances for Surfing

(15)

(Sub-total for the section: 45)

Total marks for the paper: 100