Beijing-Dublin International College

SEMESTER 1 FINAL EXAMINATION - (2018/2019)

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

2. Give the formula for *Laplacian Smoothing*. Explain why it is used **(2)**

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

4. 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)**

5. Give the formula for Jaccard index. List two domains where you would recommend its usage and explain why? **(3)**

6. The accuracy of a new diagnostic test for a particular rare disease is 90%. The incidence of disease in the population is 0.1% A patient, J, has tested positive for the disease. Give the formula for Bayes' Theorem. What is the probability that J actually has the disease? **(5)**

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

8. What is meant by *binning*? Describe two ways of binning data. Which one would you choose on a dataset relating age to income? Why? **(5)**

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

10. 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? **(3)**

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

(Sub-total for the section: 38)

Calculation Questions

1. Take a look at the data given in Figure 1. It gives the predictions made by a model for a categorical target feature. Assume that the target level *true* is the positive level. Based on the data, calculate the following evaluation measures:

   (a) Show the confusion matrix and calculate the misclassification rate **(3)**

   (b) The precision, recall and the F_1 measure **(4)**

ID	Target	Prediction	ID	Target	Prediction	ID	Target	Prediction
1	false	false	8	true	true	15	false	false
2	false	false	9	false	false	16	false	false
3	false	false	10	false	false	17	true	false
4	false	false	11	false	false	18	true	true
5	true	true	12	true	true	19	true	true
6	false	false	13	false	false	20	true	true
7	true	true	14	true	true			

Figure 1: Predictions made by a model

2. Figure 2 shows the macroeconomic variables of countries from publicly available data. The target feature is **CPI** or *corruption perception index*. CPI measures the perceived levels of corruption among public sector companies in the country, with 0 regarded as highly corrupt and 100 regarded as very clean.

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.30	4.44	0.40	1.5171
Haiti	45.00	47.67	73.10	0.09	3.40	1.7999
Nigeria	51.30	38.23	82.60	1.07	4.10	2.4493
Egypt	70.48	26.58	19.60	1.86	5.30	2.8622
Argentina	75.77	32.30	13.30	0.76	10.10	2.9961
China	74.87	29.98	13.70	1.95	6.40	3.6356
Brazil	73.12	42.93	14.50	1.43	7.20	3.7741
Israel	81.30	28.80	3.60	6.77	12.50	5.8069
U.S.A	78.51	29.85	6.30	4.72	13.70	7.1357
Ireland	80.15	27.23	3.50	0.60	11.50	7.5360
U.K.	80.09	28.49	4.40	2.59	13.00	7.7751
Germany	80.24	22.07	3.50	1.31	12.00	8.0461
Canada	80.99	24.79	4.90	1.42	14.20	8.6725
Australia	82.09	25.40	4.20	1.86	11.50	8.8442
Sweden	81.43	22.18	2.40	1.27	12.80	9.2985
New Zealand	80.67	27.81	4.90	1.13	12.30	9.4627

Figure 2: Dataset showing macroeconomic variables

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Russia	67.62	31.68	10.00	3.87	12.90	?

Figure 3: Macroeconomic variables of the Query Instance

Take a look at the query instance, given in Figure 3. What value would a **weighted-kNN** prediction model return for range-normalized data of Russia? Use k= 16 (the full dataset) and a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbour and the query. Note since the units of each variable are different, you should first range-normalize the macroeconomic variables.

(15)

(Sub-total for the section: 22)

Total marks for the paper: 60