21/5/2019

14.00 - 16.00pm

SPEC 9270 Machine Learning

Basement 1, Kevin Street

Programme Code: DT2228A, DT228B, DT9231
Module Code: SPEC 9270
CRN: 22812, 20253, 32043

TECHNOLOGICAL UNIVERSITY DUBLIN

KEVIN STREET CAMPUS

MSc. in Computing
(Full-Time)

Year 1

MSc. in Computing
(Part-Time)

Year 2

Postgraduate Certificate in Data Science

Year 1

SEMESTER 2 EXAMINATIONS 2018/19

Machine Learning

Dr. David Leonard
Professor Sarah Jane Delany
Dr. Deirdre Lillis
Dr. Barry Haycock

Please answer 2 questions

Each question is worth 50 marks

1. **(a) (i)** Why would you use Information Gain Ratio instead of Information Gain?

(4 marks)

(ii) What aspect of similarity does cosine similarity capture? Include in your answer a real-world example that illustrates the application of this measure.

(4 marks)

(iii) Using the data in the table below, calculate the error rate and false positive rate.

		Predicted Class: Yes	Predicted Class: No
Actual Class	Yes	80	40
	No	20	60

(4 marks)

(iv) In the context of Linear regression what is the meaning of the following terms?

- polynomial regression
- high leverage points

(4 marks)

(v) Regularisation is a common technique adopted during the training of Machine Learning models. Explain the purpose of regularisation, including in your answer an example that clearly illustrates the concept.

(4 marks)

(vi) What are the characteristics of Stacked Generalisation (Stacking) that distinguish it from other ensemble techniques such as Bagging.

(4 marks)

(b) (i) The table below shows the weights of a multivariate logistic regression model that has been learned to predict the probability that shoppers perform a repeat purchase of a free gift that they are given.

The descriptive features are the **Age** of the customer, the **Socio Economic Band** to which the customer belongs (a, b, or c) and the average **Spend** of the customer on each visit to the shop

Feature	Intercept	Age	Socio Economic Band b	Socio Economic Band c	Spend
Weight	-4	-0.2	-0.9	-0.2	0.2

Use this model to make a prediction for a query instance with the following values: Age = 30, Socio Economic Band = a, Spend = 60

(6 marks)

Question 1. continues overleaf

(ii) The training process for both artificial neural networks and multi-variate logistic regression models can be viewed as a gradient decent over an error surface.

Explain the above statement. (8 marks)

Describe one significant issue that arises when gradient descent is used, and suggest an approach to handling this issue.

(2 marks)

(c) (i) Specify the details of the key steps involved in the **stochastic gradient descent** version of the BackPropagation algorithm for feedforward neural networks containing two layers of sigmoid units (precise formulae are **not** necessary).

(10 marks)

2. **(a)** **(i)** Show the different options at a decision tree node for creating a binary split of a nominal attribute with 4 distinct values.

(4 marks)

(ii) Given 8 instances with class labels and values for a numeric attribute as shown in table below:

Class Label	No	Yes	No	Yes	Yes	No	Yes	No
Attribute Value	88	72	85	57	83	89	39	2

Pick an attribute value to create a binary split at a decision tree node that best separates the instances into purer subsets according to the entropy measure. Calculate the value of the Information Gain for the chosen split.

(12 marks)

(iii) Suppose that a new instance is missing some attribute values, explain the process by which a decision tree may be used to predict its class label.

(6 marks)

(b) **(i)** Consider the task of classifying online user generated comments on news articles as positive or negative. Explain the steps involved in generating a representation of the text to create an ABT.

(3 marks)

For each step identify two options available and justify with examples an option you would use for your task.

(6 marks)

(ii) One of the challenges with dealing with text data is high dimensionality. Explain why with an example.

(2 marks)

(iii) Feature selection can help with the challenge of high dimensionality. Explain the difference between filter and wrapper feature selection in classification.

(4 marks)

(iv) Select a feature selection technique that you would use for the task in **(b)** **(i)** above. Give justifications for your answer.

(3 marks)

(c) **(i)** Explain the purpose of a Receiver Operating Characteristic (ROC) Curve.

(5 marks)

(ii) Outline a procedure that can be used to construct a ROC curve.

(5 marks)

3. **(a)** **(i)** Calculate the probability of a model ensemble that uses simple majority voting making an incorrect prediction on a binary classification problem in the following scenario:

- The ensemble contains 3 independent models, all of which have an error rate of 0.3

(6 marks)

(ii) Compare and contrast the Rotation Forest and AdaBoost algorithms for creating ensembles.

(12 marks)

(b) **(i)** Specify an algorithm that can be used to perform basic Agglomerative Hierarchical Clustering (AHC).

(6 marks)

(ii) Distinguish between the **single link**, **complete link** and **group average** techniques for calculating the proximity between two clusters in AHC.

(6 marks)

(iii) When using K-means clustering, give examples of 3 practical issues to be aware of and the workarounds that can be applied to address them.

(6 marks)

(iv) Outline three of the main ideas behind the pre-processing technique of Principal Component Analysis.

(6 marks)

(c) **(i)** Suppose that you are involved in devising a solution to a Machine Learning task. Assuming that this involves hyper-parameter tuning and model selection. Give details of the key steps and important factors to take into consideration in the development and evaluation of a final model.

(8 marks)

Appendix

Useful Formulae

$$Entropy = -\sum_{i=1}^{m} p_i(log_2\, p_i)$$

$$Information\ Gain = \Delta = Entropy(parent) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

$$Sigmoid(z) = \frac{1}{1+e^{-z}}$$

$$Binomial\ Distribution\ = Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

Binomial Coefficient Values

$$\binom{3}{0} = 1, \quad \binom{3}{1} = 3, \quad \binom{3}{2} = 3, \quad \binom{3}{3} = 1$$