08.01.20

09.30 - 11.30am

CMPU 4011 Mach Learng for Data Analytics

Basement 3, Kevin Street

Programme Code: DT211C, DT228, DT282
Module Code: CMPU 4011
CRN: 25772, 22422, 31092

TECHNOLOGICAL UNIVERSITY DUBLIN

KEVIN STREET CAMPUS

BSc. (Honours) Degree in Computer Science (Infrastructure)

BSc. (Honours) Degree in Computer Science

BSc. (Honours) Degree in Computer Science (International)

Year 4

SEMESTER 1 EXAMINATIONS 2019/20

Machine Learning for Data Analytics

Dr. Svetlana Hensman
Dr. Deirdre Lillis
Dr. David Malone – DT211C
Dr. Martin Crane – DT228/DT282

Two Hours

Question 1 is compulsory.

Answer **Question 1** (40 marks) and **any 2 other** questions (30 marks each).

1. QUESTION 1 (TOTAL MARKS 40)

(a) Distinguish between **supervised** and **unsupervised** machine learning algorithms.

(5 marks)

(b) Explain what is involved in the **data exploration** process.

(10 marks)

(c) In the context of machine learning discuss what is meant by **inductive bias**. Explain the inductive bias of the **ID3** algorithm.

(10 marks)

(d) Table 1 below shows the predictions made by a model of a spam classifier for a categorical target feature (the predicition is spam=true or spam=false).
Create the **confusion matrix** for the results listed in Table 1.

(5 marks)

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

Table 1: The predictions made by a spam classification model for a categorical target on a test set of 20 instances

(e) In the context of machine learning, explain the concept of **cross validation**, and distinguish between **hold out**, **k-fold cross validation** and **leave-one-out cross validation**.

(10 marks)

2. QUESTION 2 (TOTAL MARKS 30)

Table 2 on the next page lists a sample of data from a census. There are four descriptive features in this dataset (AGE, EDUCATION, MARITAL STATUS, OCCUPATION) and the target feature ANNUAL INCOME has 3 levels ($<25K$, $25K$–$50K$, $>50K$).

Note, Table 4 at the end of the paper lists some equations that you may find useful for this question.

(a) Calculate the ENTROPY for this dataset.

(5 marks)

(b) When building a decision tree, the easiest way to handle a continuous feature is to define a threshold around which splits will be made. What would be the optimal threshold to split the continuous AGE feature (use information gain based on entropy as the feature selection measure)?

(20 marks)

(c) One of your collegues has suggested that since the ID attribute has an unique value for each instance, it will be a good idea to be included in the model as it will be improve the performance.

Do you agree or disagree with the statement above? Explain your answer.

(5 marks)

ID	AGE	EDUCATION	MARITAL STATUS	OCCUPATION	ANNUAL INCOME
1	39	bachelors	never married	transport	$25K$–$50K$
2	50	bachelors	married	professional	$25K$–$50K$
3	18	high school	never married	agriculture	$<25K$
4	28	bachelors	married	professional	$25K$–$50K$
5	37	high school	married	agriculture	$25K$–$50K$
6	24	high school	never married	armed forces	$<25K$
7	52	high school	divorced	transport	$25K$–$50K$
8	40	doctorate	married	professional	$>50K$

Table 2: Census data for the ID3 Algorithm Question

3. QUESTION 3 (TOTAL MARKS 30)

The dataset in Table 3 shows a small set of five historical emails.

The features included are IMAGES (represents how many images the email contains, possible values are: none, one, few), KNOWN SENDER (represents if the sender email is in the address book, possible values: yes, no) and BWORDS (represents if the email contains blacklisted words, possible values: none, one, few).

The target feature SPAM that shows if that instance was a spam or not (possible values: yes, no).

ID	IMAGES	KNOWN SENDER	BWORDS	SPAM
1	none	no	none	yes
2	few	yes	few	yes
3	none	yes	few	no
4	one	no	none	no
5	few	no	one	no

Table 3: Dataset for Question 3.

(a) Calculate the probabilities that would be required by a Naïve Bayes classifier trained on the dataset in Table 3.

(15 marks)

(b) How would the resulting Naïve Bayes algorithm from part 3(a) classify the new instance below:

IMAGES = none, KNOWN SENDER = no , BWORDS = one

(5 marks)

(c) Describe how a **3-nearest-neighbor** algorithm would classify the new instance

IMAGES = none, KNOWN SENDER = no , BWORDS = one

based on the training data in Table 3.

To calculate the distance you can use hamming distance which is calculated as the number of positions at which the corresponding features have the same values.

(10 marks)

4. QUESTION 4 (TOTAL MARKS 30)

(a) Discuss the difference between **linear regression** and **logistic regression** models.

(10 marks)

(b) Discuss the limitations of the **simple linear regression** models, and explain how using **basis functions** can help.

(10 marks)

(QUESTION 4 CONTINUED ON NEXT PAGE)

(c) A multivariate logistic regression model has been built to predict the propensity of customers of a broadband company to renew their contract. The descriptive features used by the model are the age of the customer (AGE), the current speed of the broadband (SPEED), and the average number of calls the customer makes to the customer service help line per year (CUST SERV).

This model is being used by the marketing department to determine which customers should be approached with more competitive offers.

The trained model is

$$\begin{aligned}\text{RENEW} = &- 3.81 - (0.03 \times \text{AGE}) \\ &+ (0.02 \times \text{SPEED}) \quad +(0.81 \times \text{CUST SERV})\end{aligned}$$

The logistic function is defined as:

$$logistic(x) = \frac{1}{1+e^{-x}}$$

Assuming that the *yes* level is the positive level and the classification threshold is 0.5, use this model to make predictions for the query instance below.

ID	AGE	CUST SERV	SPEED
id1	30	1.25	120

(10 marks)

$H(\mathbf{t}, \mathcal{D})$	$= - \sum_{l \in levels(t)} P(t=l) \times log_2(P(t=l))$
$rem(\mathbf{d}, \mathcal{D})$	$= \sum_{l \in levels(d)} \frac{\lvert\mathcal{D}_{d=l}\rvert}{\lvert\mathcal{D}\rvert} \times H(t, \mathcal{D}_{d=l})$
$IG(\mathbf{d}, \mathcal{D})$	$= H(\mathbf{t}, \mathcal{D}) - rem(\mathbf{d}, \mathcal{D})$

Table 4: Equations from information theory.