9/1/2019

09.30 - 11.30am

Courtyard, DIT Aungier Street

DUBLIN INSTITUTE OF TECHNOLOGY

DT211C/4 BSc. (Honours) Degree in Computer Science (Infrastructure)

DT228/4 BSc. (Honours) Degree in Computer Science

DT282/4 BSc. (Honours) Degree in Computer Science (International)

DT8900/1 International Pre Masters for MSc in Computing

WINTER EXAMINATIONS 2017/2018

MACHINE LEARNING FOR DATA ANALYTICS [CMPU4011]

DR. SVETLANA HENSMAN
DR. DEIRDRE LILLIS
DR. DAVID MALONE – DT211C
DR. MARTIN CRANE – DT228/DT282

WEDNESDAY 9TH JANUARY 9.30 A.M. – 11.30 A.M.

TWO HOURS

QUESTION 1 IS **COMPULSORY**

ANSWER QUESTION 1 (40 MARKS) **AND**

ANY 2 OTHER QUESTIONS (30 MARKS EACH).

1. (a) Briefly explain the term **predictive data analytics**.

(5 marks)

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

(5 marks)

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

(10 marks)

(d) Discuss what is meant by the **curse of dimensionality**, and suggest how it may be dealt with.

(10 marks)

(e) In the context of machine learning discuss what is meant by **inductive bias**, and explain what the inductive bias is for the **k-NN** algorithm.

(10 marks)

2. Table 1 below contains students exam performance data for six students, listing details such as if they achieved an 'A' last year (yes/no), their gender (male/female), did they have good attendance (yes/no), did they spend too much time partying (yes/no) and did they achieve A this year (yes, no).

No.	Student	A last year	Gender	Attendence	Partying	A this year
1	Mick	yes	male	no	yes	yes
2	Alan	yes	male	yes	no	yes
3	Ellie	no	female	yes	no	yes
4	Jack	no	male	no	yes	no
5	George	yes	female	yes	yes	yes
6	Simon	no	male	yes	yes	no

Table 1: Dataset showing exam performance of a set of six students.

Table 3 at the end of this exam paper contains equations you may find useful when answering this question.

(a) Using the dataset in Table 1, what is the entropy of this set of training examples with respect to the target feature classification?

(5 marks)

(b) Construct the decision tree that would be generated by the **ID3** algorithm using entropy-based information gain. Show the steps when building the tree.

(20 marks)

(c) Using the decision tree build in the section 2(b) above, what will be the classification for the new instance below

No.	Student	A last year	Gender	Attendence	Partying	A this year
7	John	no	male	no	yes	?

(5 marks)

3. For the dataset of training examples in Table 2 each feature can take on one of three possible values: a, b, or c and the target classification is either category '+' or '-'.

ID	F1	F2	F3	Category
1	a	c	a	+
2	c	a	c	+
3	a	a	c	-
4	b	c	a	-
5	c	c	b	-

Table 2: Dataset for Question 3.

(a) Calculate the probabilities that would be required by a Naïve Bayes classifier trained on the dataset in Table 2.

(15 marks)

(b) How would the resulting Naïve Bayes algorithm from part 3(a) classify the following new instance? F1 = a, F2 = c , F3 = b

(5 marks)

(c) Describe how a **3-nearest-neighbor** algorithm would classify the new instance
F1 = a, F2 = c , F3 = b
based on the training data in Table 2.
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. (a) Discuss what is meant by **linear regression** and **logistic regression**.

(10 marks)

(b) Discuss in what cases would we prefer **average class accuracy** over **classification accuracy** as a performance measure.

(10 marks)

(c) Discuss what is meant by **eager learner** and **lazy learner** in the context of machine learning. Give one example of each to illustrate your answer.

(10 marks)

$$
\begin{aligned}
H(\mathbf{f}, \mathcal{D}) &= -\sum_{l \in levels(f)} P(f = l) \times log_2(P(f = l)) \\
rem(\mathbf{f}, \mathcal{D}) &= \sum_{l \in levels(f)} \frac{|\mathcal{D}_{f=l}|}{|\mathcal{D}|} \times H(t, \mathcal{D}) \\
IG(\mathbf{d}, \mathcal{D}) &= H(\mathbf{t}, \mathcal{D}) - rem(\mathbf{d}, \mathcal{D})
\end{aligned}
$$

Table 3: Equations from information theory.

COLLEGE EXAMINATIONS

AMENDMENTS TO EXAMINATION QUESTION PAPER

COURSE REF

VENUE: *K103 | K104
*(2 students)

SUBJECT: mach Learn for Data Anayltics

DATE: 9/1/18

TIME: 9:30 – 11:30

SIGNED: Tracy Armstrong

INSTRUCTIONS:

Q2A The target feature is the final column in Table 1:

A this year

COLLEGE EXAMINATIONS

AMENDMENTS TO EXAMINATION QUESTION PAPER

COURSE REF VENUE:

SUBJECT: Computation mach learn for Data Analytics

DATE: 9/11/18

TIME: 9:30 – 11:30

SIGNED: Tracy Armstrong

INSTRUCTIONS:

predictions for new instances. Typically faster to make predictions. Examples: decision trees, Naive Bayes. Could have issues with concept drift.

5 marks)

$$H(\mathbf{f}, \mathcal{D}) = -\sum_{l \in levels(f)} P(f = l) \times log_2(P(f = l))$$

$$rem(\mathbf{f}, \mathcal{D}) = \sum_{l \in levels(f)} \frac{|\mathcal{D}_{f=l}|}{|\mathcal{D}|} \times H(t, \mathcal{D})$$

$$IG(\mathbf{d}, \mathcal{D}) = H(\mathbf{t}, \mathcal{D}) - rem(\mathbf{d}, \mathcal{D})$$

Table 3: Equations from information theory.

$H(t, D_{f=l})$