Programme Code: DT249, DT255
Module Code: CMPU4011
CRN: 22564, 32371

TECHNOLOGICAL UNIVERSITY DUBLIN

KEVIN STREET CAMPUS

BSc (Hons) Information Systems/Information Technology (Part-Time)

BSc (Hons) Information Systems/Information Technology (Full-Time)

Year 4

SUPPLEMENTAL EXAMINATIONS 2018/19

Machine Learning for Data Analytics

Dr. Bojan Božić
Dr. Deirdre Lillis
Professor Eleni Mangina

Question 1 is compulsory
Answer Question 1 (40 marks) and
Any 2 other questions (30 marks each)

1. (a) What is **supervised machine learning**?

(5 marks)

(b) Explain what can go wrong when a machine learning classifier uses the wrong **inductive bias**.

(5 marks)

(c) Table 1, on the next page, shows the predictions made for a categorical target feature by a model for a test dataset. Based on this test set, calculate the evaluation measures listed below.

(i) A **confusion matrix**

(6 marks)

(ii) The **misclassification rate**

(4 marks)

$$misclassification\ rate = \frac{(FP + FN)}{(TP + TN + FP + FN)}$$

(iii) The **precision**, **recall**, and $\mathbf{F}_1$ **measure**

(12 marks)

$$precision = \frac{TP}{(TP + FP)}$$
$$recall = \frac{TP}{(TP + FN)}$$
$$F_1\ measure = 2 \times \frac{(precision \times recall)}{(precision + recall)}$$

(iv) The **average class accuracy (harmonic mean)**. (During this calculation you should round all long floats to 3 places of decimal.)

(8 marks)

$$average\ class\ accuracy_{HM} = \frac{1}{\frac{1}{|levels(t)|}\sum_{l \in levels(t)} \frac{1}{recall_l}}$$

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

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

2. (a) A data analyst building a k-nearest neighbour model for a continuous prediction problem is considering appropriate values to use for k on an *imbalanced training set*.

   (i) Initially the analyst uses a simple average of the target variables for the k nearest neighbours in order to make a new prediction. After experimenting with small values for k in the range $0 - 5$ it occurs to the analyst that they might get very good results if they keep increasing k to a value closer to the total number of instances in the training set. Do you think the analyst is likely to get good results using these values for k?

   (5 marks)

   (ii) If the analyst was using a distance weighted averaging function rather than a simple average for their predictions would this have made their idea any more useful?

   (5 marks)

   (iii) By using a different distance metric than the standard Euclidean Distance, would any of the previous answers change? Provide an explanation to your answer.

   (5 marks)

   (b) 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 3, also on the next page, lists some equations that you may find useful for this question.

   (i) Calculate the ENTROPY for this dataset.

   (5 marks)

   (ii) When building a decision tree, we must partition the data into homogeneous subsets. What is the metric used to decide on the partitions? How it relates to the entropy of the dataset?

   (5 marks)

   (iii) In the case that we have a continuous descriptive feature, what is the procedure to create the partitions using information gain?

   (5 marks)

Table 2: Census data for the ID3 Algorithm Question

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 3: Equations from information theory.

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

3. Table 4 lists a dataset of the previous decision made by a couple regarding whether or not they would wait for a table at a restaurant (i.e., the feature WAITED is the target feature in this domain).

   (a) Calculate the probabilities (to four places of decimal) that a **naive Bayes** classifier would use to represent this domain.

   (18 marks)

   (b) Assuming conditional independence between features given the target feature value, calculate the **probability** of each outcome (WAITED=Yes, and WAITED=No) for the following restaurant for this couple (marks will be deducted if workings are not shown, round your results to four places of decimal)

   BAR=False, PATRONS=None, PRICE=Expensive

   (10 marks)

   (c) What prediction would a **naive Bayes** classifier return for the above restaurant?

   (2 marks)

Table 4: A dataset describing the previous decisions made by an individual about whether to wait for a table at a restaurant.

ID	BAR	PATRONS	PRICE	WAITED
1	False	Some	Expensive	Yes
2	False	Full	Cheap	No
3	True	Some	Cheap	Yes
4	False	Full	Cheap	Yes
5	False	Full	Expensive	No
6	True	Some	Reasonable	No
7	True	None	Cheap	No
8	False	Some	Reasonable	No
9	True	Full	Cheap	No
10	True	None	Reasonable	Yes

4. (a) The following model is commonly used for continuous prediction tasks:

$$y(x) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

(i) Provide the name for this model and explain all of the terms that it contains.

(4 marks)

(ii) Explain how the following model can overcome some of the limitations of the model given above.

(8 marks)

$$y(x) = \sum_{j=0}^{M-1} w_j \phi_j(x)$$

(b) 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 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. This model is being used by the marketing department to determine who should be given the free gift. The trained model is

$$\begin{aligned}\text{REPEAT PURCHASE} = &- 3.82398 - 0.02990 \times \text{AGE} \\ &+ 0.74572 \times \text{SHOP FREQUENCY} \\ &+ 0.02999 \times \text{SHOP VALUE}\end{aligned}$$

And, 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 each of the query instances shown in Table 5 below.

(18 marks)

Table 5: The queries for the multivariate logistic regression question

ID	AGE	SHOP FREQUENCY	SHOP VALUE
1	56	1.60	109.32
2	21	4.92	11.28
3	48	1.21	161.19