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

SEMESTER 2 EXAMINATIONS 2019/20

Machine Learning for Data Analytics

Dr. Bojan Božić
Dr. Deirdre Lillis
Professor Eleni Mangina

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

1. (a) Why is machine learning an **ill-posed problem**?

(5 marks)

(b) What is the **inductive bias** of a machine learning algorithm?

(5 marks)

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

(5 marks)

(d) 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 **classification accuracy**

(4 marks)

$$classification\ accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

(iii) The **precision**, **recall**, and **F$_1$ measure**

(15 marks)

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

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	true	true	12	false	false
3	false	false	13	false	false
4	true	true	14	false	false
5	false	false	15	true	true
6	false	false	16	false	false
7	true	false	17	true	true
8	true	true	18	true	true
9	true	true	19	false	false
10	true	true	20	false	false

2. (a) Table 2, on the next page, lists a dataset containing examples described by two descriptive features, **Feature 1** and **Feature 2**, and labelled with a target class **Target**. Table 3, also on the next page, lists the details of a query for which we want to predict the target label. We have decided to use a **3-Nearest Neighbor** model for this prediction and we will use Euclidean distance as our distance metric:

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n} \left((x_1.f_i - x_2.f_i)^2 \right)}$$

(i) With which target class (**TypeA** or **TypeB**) will our **3-Nearest Neighbor** model label the query? Provide an explanation for your answer.

(8 marks)

(ii) There is a large variation in range between **Feature 1** and **Feature 2**. To account for this we decide to normalize the data. Compute the normalized versions of Feature 1 and Feature 2 to four decimal places of precision using range normalization

$$x_i.f' = \frac{x_i.f - min(f)}{max(f) - min(f)}$$

(4 marks)

(iii) Assuming we use the normalized dataset as input, with which target class (**TypeA** or **TypeB**) will our **3-Nearest Neighbor** model label the query? Provide an explanation for your answer.

(8 marks)

(b) A dataset showing the decisions made by a professional basketball team on whether to draft college players based on 4 features (1 continuous and 3 categorical) as listed in Table 4 on the next page. (Note that Table 5, also on the next page, lists some equations that you may find useful for this question.)

(i) Given that the DRAFT column lists the values of the target variable, compute the entropy for this dataset.

(5 marks)

(ii) What does the entropy tell us about a dataset? What happens with the distribution of a dataset if the entropy is increased?

(5 marks)

Table 2: Dataset for the 3-Nearest Neighbor question

ID	Feature 1	Feature 2	Target
101	180000	4	TypeA
102	120000	3	TypeB
103	360000	7	TypeB
104	420000	5	TypeA
105	480000	8	TypeB

Table 3: Query instance for the 3-Nearest Neighbor question.

ID	Feature 1	Feature 2	Target
250	240000	4	?

Table 4: A dataset showing the decisions made by a professional basketball team on whether to draft college players.

ID	AGE	SPEED	AGILITY	ABILITY	DRAFT
1	20	1	1	3	F
2	21	2	2	1	F
3	20	2	1	2	F
4	22	2	1	1	F
5	22	4	4	4	T
6	21	5	4	5	T
7	23	5	5	4	T
8	19	4	5	5	T
9	22	5	5	5	T
10	21	1	1	2	F
11	20	5	5	4	T
12	21	3	1	1	F

Table 5: Equations for entropy.

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

3. Table 6 lists a dataset of books and whether or not they were purchased by an individual (i.e., the feature PURCHASED 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 (PURCHASED=Yes, and PURCHASED=No) for the following book (marks will be deducted if workings are not shown, round your results to four places of decimal)

   2ND HAND=False, GENRE=Literature, COST=Expensive

   (10 marks)

   (c) What prediction would a **naive Bayes** classifier return for the new book from (b)?

   (2 marks)

Table 6: A dataset describing the a set of books and whether or not they were purchased by an individual.

ID	2ND HAND	GENRE	COST	PURCHASED
1	False	Romance	Expensive	Yes
3	True	Romance	Cheap	Yes
4	False	Science	Cheap	Yes
10	True	Literature	Reasonable	Yes
2	False	Science	Cheap	No
5	False	Science	Expensive	No
6	True	Romance	Reasonable	No
7	True	Literature	Cheap	No
8	False	Romance	Reasonable	No
9	True	Science	Cheap	No

4. (a) A multivariate linear regression model has been built to predict the HEATING LOAD in a residential building based on a set of descriptive features describing the characteristics of the building. Heating load is the amount of heat energy required to keep a building at a specified temperature, usually $65°$ Fahrenheit, during the winter regardless of outside temperature. The descriptive features used are the overall surface area of the building, the height of the building, the area of the building's roof, and the percentage of wall area in the building that is glazed. This kind of model would be useful to architects or engineers when designing a new building. The trained model is

$$\begin{aligned}\text{HEATING LOAD} = &-26.030 + 0.0497 \times \text{SURFACE AREA} \\ &+ 4.942 \times \text{HEIGHT} - 0.090 \times \text{ROOF AREA} \\ &+ 20.523 \times \text{GLAZING AREA}\end{aligned}$$

Use this model to make predictions for each of the query instances shown in the Table 7 on the next page.

(12 marks)

(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 8 on the next page.

(18 marks)

Table 7: The queries for the multivariate linear regression HEATING LOAD question

ID	SURFACE AREA	HEIGHT	ROOF AREA	GLAZING AREA
1	784.0	3.5	220.5	0.25
2	710.5	3.0	210.5	0.10

Table 8: 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