TECHNOLOGICAL UNIVERSITY DUBLIN

KEVIN STREET CAMPUS

BSc. (Honours) Degree in Information Systems/Information Technology

Year 4

SEMESTER 2 OPEN BOOK EXAMINATIONS 2020/21

Machine Learning for Data Analytics

Bojan Božić

Duration 9hrs

Exam script available 9am on date of the exam.
All exams submissions should be uploaded before 6pm on the date of the exam

Answer ALL questions.

Question 1 carries 40 marks and questions 2 and 3 carry 30 marks each.

1. (a) Explain why machine learning is often described as an **ill-posed problem** and give an **example**.

(5 marks)

(b) What is the **inductive bias** of a machine learning algorithm? Give an **example**.

(5 marks)

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

(5 marks)

(d) Table 1 shows 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 and **explain exactly** what they would tell us with regard to a **spam detection** dataset (assume that false is 'ham' and true is 'spam').

(i) A **confusion matrix**

(6 marks)

(ii) The **classification accuracy**

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

(4 marks)

(iii) The **precision, recall** and **F1 measure**

$$precision = \frac{TP}{(TP + FP)}$$

$$recall = \frac{TP}{(TP + FN)}$$

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

(15 marks)

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 lists a dataset containing examples described by two descriptive features, **Feature 1** and **Feature 2**, and labelled with a target class **Target**. Table 3 lists the details of a query for which we want to predict the target label. We have decided to use a **3-Nearest Neighbour** 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 Neighbour** 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 **normalise** the data. Compute the normalised versions of **Feature 1** and **Feature 2** to four decimal places of precision using range normalisation. **Explain** why range normalisation is important and give an **example**.

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

(4 marks)

(iii) Assuming we use the normalised dataset as input, with which target class (**TypeA** or **TypeB**) will our **3-Nearest Neighbour** model label the query? Provide an **explanation** and **example** 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. (Note that Table 5 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 and explain what you did and what your result tells you..

(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? Give 2 examples from other possible datasets.

(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.

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

3. (a) 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).

Calculate the probabilities (to four places of decimal) that a **naïve Bayes** classifier would use to represent this domain. **Explain** your results and **outline** why **naïve Bayes** is a good solution for this type of problem.

(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) and **explain** how you did it:

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

(10 marks)

(c) What prediction would a **naïve Bayes** classifier return for the new book from (b)? **Explain** what the result means and how to interpret probabilistic predictions.

(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