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 Predictive Analytics

Dr. Bojan Božić
Dr. Deirdre Lillis
Professor Eleni Mangina

SOLUTIONS

*** *SOLUTIONS* ***

*** *SOLUTIONS* ***

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

(5 marks)

Supervised machine learning techniques automatically learn the relationship between a set of **descriptive features** and a **target feature** from a set of historical **instances**. Supervised machine learning is a subfield of machine learning. Machine learning is defined as an automated process that extracts patterns from data. In predictive data analytics applications, we use **supervised machine learning** to build models that can make predictions based on patterns extracted from historical data.

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

(5 marks)

- If the inductive bias of the learning algorithm constrains the search to only consider simple hypotheses we may have excluded the real function from the hypothesis space. In other words, the true function is **unrealizable** in the chosen hypothesis space, (i.e., we are **underfitting**).
- If the inductive bias of the learning algorithm allows the search to consider complex hypotheses, the model may hone in on irrelevant factors in the training set. In other words the model with **overfit** the training data.

(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)

The confusion matrix can be written as

		Prediction	
		'true'	*'false'*
Target	*'true'*	8	1
	'false'	0	11

(ii) The **misclassification rate**

(4 marks)

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

Misclassification rate can be calculated as

$$\begin{aligned} misclassification\ rate &= \frac{(FP+FN)}{(TP+TN+FP+FN)} \\ &= \frac{(0+1)}{(8+11+0+1)} \\ &= 0.05 \end{aligned}$$

(iii) The **precision**, **recall**, and **F_1 measure**

(12 marks)

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

We can calculate precision and recall as follows (assuming that the *'true'* target level is the positive level):

$$\begin{aligned} precision &= \frac{TP}{(TP+FP)} \\ &= \frac{8}{(8+0)} \\ &= 1.000 \\ recall &= \frac{TP}{(TP+FN)} \\ &= \frac{8}{(8+1)} \\ &= 0.889 \end{aligned}$$

Using these figures, we can calculate the F_1 measure as

$$\begin{aligned} F_1\ measure &= 2 \times \frac{(precision \times recall)}{(precision + recall)} \\ &= 2 \times \frac{(1.000 \times 0.889)}{(1.000 + 0.889)} \\ &= 0.941 \end{aligned}$$

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

(8 marks)

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

First, we calculate the recall for each target level:

$$recall_{\text{'true'}} = \frac{8}{9} = 0.889$$

$$recall_{\text{'false'}} = \frac{11}{11} = 1.000$$

Then we can calculate a harmonic mean as

$$\begin{aligned} average\ class\ accuracy_{HM} &= \frac{1}{\frac{1}{|levels(t)|}\sum_{l \in levels(t)} \frac{1}{recall_l}} \\ &= \frac{1}{\frac{1}{2}\left(\frac{1}{0.889} + \frac{1}{1}\right)} \\ &= 0.941 \end{aligned}$$

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)

In answering this question students should realise that with an imbalanced training set, the majority class will dominate the feature space. Therefore, if the analyst set k close to the number of training examples in this imbalanced training set, the model may start using the majority class as the prediction for all queries.

(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)

Students should realise that yes, if distance weighted voting is used (particularly if a $\frac{1}{d^2}$ type distance weight is used) then examples that are far away from the query will have very little impact on the result. Again to score well students should mention that when distance weighted voting is used the value of k in k-NN classifiers is much less important.

(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)

Students should realise that no, the distance metrics only indicate the distance between data points in feature space. If the training set is imbalanced, the majority class will start to dominate the feature space and the model will use this majority class as the prediction most of the time, independent of the distance metric in use. By using the a weighted averaging function we still be a good idea to mitigate the problem as by any distance metric a weighted model would put more importance into closer samples and less importance into distant samples.

(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)

$$
\begin{aligned}
&H\left(\textsc{Annual Income}, \mathcal{D}\right) \\
&= - \sum_{l \in \left\{ \substack{<25K, \\ 25K-50K, \\ >50K} \right\}} P(\textsc{An. Inc.} = l) \times log_2\left(P(\textsc{An. Inc.} = l)\right) \\
&= -\left(\left(\frac{2}{8} \times log_2\left(\frac{2}{8}\right)\right) + \left(\frac{5}{8} \times log_2\left(\frac{5}{8}\right)\right) + \left(\frac{1}{8} \times log_2\left(\frac{1}{8}\right)\right)\right) \\
&= 1.2988\ bits
\end{aligned}
$$

(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)

The students should realise the information gain is a metric that will give us intuitions about the informativeness of features and therefore be used to create the partitions. Students should also realise that information gain is also used as a measure of the reduction in the overall entropy of a set of instances that is achieved by testing a descriptive feature.

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

(5 marks)

The students should realise that a threshold must be created for the continuous descriptive feature. They must realize that a set of thresholds must be created and each of these thresholds must be tested by creating the partitions and measuring the information gain in that partition. The threshold with the highest information gain is then selected to create the partition.

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)

   A naive Bayes classifier would require the prior probability for each level of the target feature and the conditional probability for each level of each descriptive feature given each level of the target feature:

   $P(Waited = Yes) = 0.4$ $P(Waited = No) = 0.6$
   $P(Bar = True|Waited = Yes) = 0.5$ $P(Bar = True|Waited = No) = 0.5$
   $P(Bar = False|Waited = Yes) = 0.5$ $P(Bar = False|Waited = No) = 0.5$
   $P(Patrons = None \mid Waited = Yes) = 0.25$ $P(Patrons = None \mid Waited = No) = 0.1667$
   $P(Patrons = Some \mid Waited = Yes) = 0.5$ $P(Patrons = Some \mid Waited = No) = 0.3333$
   $P(Patrons = Full \mid Waited = Yes) = 0.25$ $P(Patrons = Full \mid Waited = No) = 0.5$
   $P(Price = Cheap \mid Waited = Yes) = 0.5$ $P(Price = Cheap \mid Waited = No) = 0.5$
   $P(Price = Reasonable \mid Waited = Yes) = 0.25$ $P(Price = Reasonable \mid Waited = No) = 0.3333$
   $P(Price = Expensive \mid Waited = Yes) = 0.25$ $P(Price = Expensive \mid Waited = No) = 0.1667$

   (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)

   The initial score for each outcome is calculated as follows:
   $(Waited = Yes) = 0.5 \times 0.25 \times 0.25 \times 0.4 = 0.0125$
   $(Waited = No) = 0.5 \times 0.1667 \times 0.1667 \times 0.6 = 0.0083$
   However, these scores are not probabilities. To get real probabilities we must normalise these scores. The normalisation constant is calculated as follows:
   $\alpha = 0.0125 + 0.0083 = 0.0208$
   The actual probabilities of each outcome is then calculated as:
   $P(Waited = Yes) = \frac{0.0125}{0.0208} = (0.600961...) = 0.6010$
   $P(Waited = No) = \frac{0.0083}{0.0208} = (0.399038...) = 0.3990$

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

   (2 marks)

   A naive Bayes classifier returns outcome with the maximum a posteriori probability as its prediction. In this instance the outcome WAITED=Yes is the MAP prediction and will be the outcome returned by a naive Bayes model.

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)

> Students should explain that this is a simple linear regression model which can be effectively used to make predictions. x is a vector of feature values for a query instance and w is a vector of feature weights. An diagram of a simple one dimensional linear function would help.

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

(8 marks)

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

> Students should explain that the simple linear regression model is attractive because it is linear with respect to w but has severe limitations because it is also linear with respect to x. These greatly limits the kinds of predictions that this model will be able to make. However, the introduction of *basis functions*, shown as ϕ above, goes some way towards solving this problem. The introduction of a non-linear basis function means that models can be made non-linear functions of input x but remain linear in w which makes them computationally easier to solve.
>
> Students might give the example of polynomial regression in which $\phi_j(x) = x^j$ or some other suitable example.

(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)

Calculating the predictions made by the model simply involves inserting the descriptive features from each query instance into the prediction model. With this information, the predictions can be made as follows:

1: $Logistic(-3.82398 + -0.0299 \times 56 + 0.74572 \times 1.6 + 0.02999 \times 109.32)$
$= Logistic(-1.02672) = \frac{1}{1-e^{1.02672}}$
$= 0.26372 \Rightarrow$ *no*

2: $Logistic(-3.82398 + -0.0299 \times 21 + 0.74572 \times 4.92 + 0.02999 \times 11.28)$
$= Logistic(-0.44465) = \frac{1}{1-e^{0.44465}}$
$= 0.390633 \Rightarrow$ *no*

3: $Logistic(-3.82398 + -0.0299 \times 48 + 0.74572 \times 1.21 + 0.02999 \times 161.19)$
$= Logistic(0.477229) = \frac{1}{1-e^{-0.477229}}$
$= 0.6205 \Rightarrow$ *yes*

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