MATH2319 – MACHINE LEARNING
FINAL EXAM

IMPORTANT NOTE: You must show all your work for full credit.

Problem 1: (20 points) Consider the prediction problem below where the customer's *Preferred Channel* is a binary target feature with *Gender*, *Age*, and *Policy Type* being categorical descriptive features.

GENDER	AGE	POLICY TYPE	PREF CHANNEL
female	young	planC	phone
male	young	planA	email
male	young	planA	email
female	middle-aged	planC	email
female	middle-aged	planB	phone
male	middle-aged	planC	phone
male	mature	planB	email
male	mature	planC	phone
female	mature	planA	phone

A) (10 points) What target level will a Naïve Bayes model predict for the following customer?
Gender = male, *Age* = middle-aged, *Policy Type* = planA

The factor probabilities corresponding to this dataset are calculated as below.

$P(phone)$	=	0.56	$P(email)$	=	0.44
$P(\text{GENDER} = female \mid phone)$	=	0.6	$P(\text{GENDER} = female \mid email)$	=	0.25
$P(\text{GENDER} = male \mid phone)$	=	0.4	$P(\text{GENDER} = male \mid email)$	=	0.75
$P(\text{AGE} = young \mid phone)$	=	0.2	$P(\text{AGE} = young \mid email)$	=	0.5
$P(\text{AGE} = middle\text{-}aged \mid phone)$	=	0.4	$P(\text{AGE} = middle\text{-}aged \mid email)$	=	0.25
$P(\text{AGE} = mature \mid phone)$	=	0.4	$P(\text{AGE} = mature \mid email)$	=	0.25
$P(\text{POLICY} = planA \mid phone)$	=	0.2	$P(\text{POLICY} = planA \mid email)$	=	0.5
$P(\text{POLICY} = planB \mid phone)$	=	0.2	$P(\text{POLICY} = planB \mid email)$	=	0.25
$P(\text{POLICY} = planC \mid phone)$	=	0.6	$P(\text{POLICY} = planC \mid email)$	=	0.25

Pr(Channel = phone | **q**) = 0.56 * 0.4 * 0.4 * 0.2 = 0.018
Pr(Channel = email | **q**) = 0.44 * 0.75 * 0.25 * 0.5 = 0.042
Prediction: Channel = email.

B) (10 points) In Part (A), suppose the *Policy Type* of the customer was incorrectly recorded as planA and the true *Policy Type* is actually planC. In light of this new information, what target level will a Naïve Bayes model predict for this customer?

Gender = male, *Age* = middle-aged, *Policy Type* = planC
Pr(Channel = phone | **q**) = 0.56 * 0.4 * 0.4 * 0.6 = 0.054
Pr(Channel = email | **q**) = 0.44 * 0.75 * 0.25 * 0.25 = 0.021
Prediction: Channel = phone.

Problem 2: (15 points) Suppose that we have a set of observations, each with measurements on p = 1 feature, X. We assume that X is uniformly (evenly) distributed on [0,1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observation's that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with X = 0.5, we will use observations in the range [0.45, 0.55]. In each part below, please also explain your reasoning.

A) (5 points) On average, what fraction of the available observations will we use to make the prediction?

Answer: 10%

B) (10 points) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centred around the test observation that contains, on average, 10% of the training observations. For p = 10, what is the length of each side of the hypercube? Note that a hypercube is a generalization of a cube to an arbitrary number of dimensions. When p = 1, a hypercube is simply a line segment, when p = 2, it is a square, etc.

Answer: (0.1)^(1/10) = 0.794

Problem 3: (15 points) Email spam filtering models often use a bag-of-words representation for emails. In a bag-of-words representation, the descriptive features that describe a document (in this case, an email) each represent how many times a particular word occurs in the document. One descriptive feature is included for each word in a predefined dictionary. The dictionary is typically defined as the complete set of words that occur in the training dataset. The table below lists the bag-of-words representation for the following five emails and a target feature, SPAM, whether they are spam emails or genuine emails:

- "*money, money, money*"
- "*free money for free gambling fun*"
- "*gambling for fun*"
- "*machine learning for fun, fun, fun*"
- "*free machine learning*"

	Bag-of-Words							
ID	MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	SPAM
1	3	0	0	0	0	0	0	true
2	1	2	1	1	1	0	0	true
3	0	0	1	1	1	0	0	true
4	0	0	1	0	3	1	1	false
5	0	1	0	0	0	1	1	false

What target level would a nearest neighbour model with k=3 using Manhattan distance return for the following email: "*machine learning for free*"?

The bag-of-words representation for this query is as follows:

ID	Bag-of-Words MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	SPAM
Query	0	1	1	0	0	1	1	?

The table below shows the calculation of the Manhattan distance between the query bag-of-words vector and each instance in the dataset:

ID	$abs(\mathbf{q}[i[- \mathbf{d}_j[i[)$ MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	Manhattan Distance
1	3	1	1	0	0	1	1	7
2	1	1	0	1	1	1	1	6
3	0	1	0	1	1	1	1	5
4	0	1	0	0	3	0	0	4
5	0	0	1	0	0	0	0	1

Based on these Manhattan distance calculations, the three nearest neighbors to the query are instances $\mathbf{d}_5$, $\mathbf{d}_4$, and $\mathbf{d}_3$. The majority of these three neighbors have a target value of SPAM = *false*. Consequently, the 3-NN model using Manhattan distance will return a prediction of SPAM = *false*.

Problem 4: (15 points) The table below 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 where true corresponds to the positive level.

ID	Target	Prediction
1	true	false
2	false	false
3	true	false
4	false	true
5	true	true
6	false	false
7	false	false
8	true	true
9	false	false
10	true	true

A) (5 points) The confusion matrix with row and column labels

		Prediction	
		True	False
Target	True	3	2
	False	1	4

B) (5 points) The average class accuracy (harmonic mean)

$Recall_{true} = 3/5 = 0.6$
$Recall_{false} = 4/5 = 0.8$
$ACA_{HM} = 1/ ((1/2) * (1/0.6 + 1/0.8)) = 0.685$

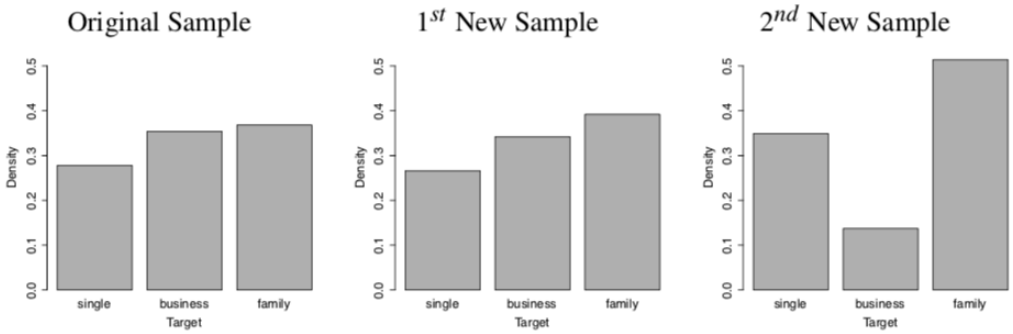
C) (5 points) The precision, recall, and F_1 measure

Precision =3/4 = 0.75
Recall =3/5 = 0.6
$F_1 = 2 * (0.75 * 0.6) / (0.75 + 0.6) = 0.67$

Problem 5: (20 points) A retail supermarket chain has built a prediction model that recognizes the household that a customer comes from as being one of *single*, *business*, or *family*. After deployment, the analytics team at the supermarket chain uses the stability index to monitor the performance of this model. The table below shows the frequencies of predictions of the three different levels made by the model for the original validation dataset at the time the model was built, for the month after deployment, and for a month-long period that is six months after deployment.

Target	Original Sample	1st New Sample	2nd New Sample
single	123	252	561
business	157	324	221
family	163	372	827

Bar plots of these three sets of prediction frequencies are shown in the following images.



Calculate the stability index for the two new periods and determine whether the model should be retrained at either of these points.

The stability index is calculated as

$$\textit{stability index} = \sum_{l \in \textit{levels}} \left(\left(\frac{|\mathcal{A}_{t=l}|}{|\mathcal{A}|} - \frac{|\mathcal{B}_{t=l}|}{|\mathcal{B}|} \right) \times \log_e \left(\frac{|\mathcal{A}_{t=l}|}{|\mathcal{A}|} / \frac{|\mathcal{B}_{t=l}|}{|\mathcal{B}|} \right) \right)$$

where l is a target level, $|\mathcal{A}|$ refers to the size of the test set on which performance measures were originally calculated, $|\mathcal{A}_{t=l}|$ refers to the number of instances in the original test set for which the model made a prediction of level l for target t, $|\mathcal{B}|$ and $|\mathcal{B}_{t=l}|$ refer to the same measurements on the newly collected dataset. The following table shows the components of calculating this for the two new periods.

	Original		1^{st} New Sample			2^{nd} New Sample		
Target	Count	%	Count	%	SI_t	Count	%	SI_t
single	123	0.2777	252	0.2658	0.00052	561	0.3487	0.01617
business	157	0.3544	324	0.3418	0.00046	221	0.1374	0.20574
family	163	0.3679	372	0.3924	0.00157	827	0.5140	0.04881
Sum	443		948		0.003	1,609		0.271

For the first new sample, the stability index is 0.003, which indicates that there is practically no difference between the distribution of target levels predicted for the original validation dataset and for the data in the newt period.

For the second sample, the stability index is 0.271, which indicates a massive difference between the distribution of target levels at this point in time compared to the distribution predicted for the original validation set. This suggests that concept drift has occurred and that the model should be retrained.

Problem 6: (15 points) Consider a supervised machine learning problem where there are four binary descriptive features and a binary target feature. You are given a training dataset that has 10 observations with unique combinations of the descriptive features. You are told to assume that there is no noise in the training data.

A) (5 points) What is the total number of prediction models that can be used?

Total number of combinations of descriptive features = 2^4 = 16
Total number of prediction models = 2^16 = 65,536

B) (5 points) What is the total number of prediction models that are consistent with the training dataset?

Out of the 16 possible combinations, 10 of them are known, leaving 6 unknown combinations.
Total number of consistent prediction models = 2^6 = 64

C) (2 points) Select the correct choice: If the training data is assumed to be noisy, the total number of prediction models would be (i) less than (ii) more than (iii) equal to what you calculated in part (A) in general.
Answer: (iii) equal to

D) (2 points) Select the correct choice: If the training data is assumed to be noisy, the total number of prediction models that are consistent with the training data would be (i) less than (ii) more than (iii) equal to what you calculated in part (B) in general.
Answer: (iii) equal to

E) (1 point) How should the blank space be filled? Because a single consistent model cannot be determined based on the training data in general, machine learning is said to be __an ill-posed problem_.