Cairo University
Faculty of Computers and Artificial Intelligence

Midterm Exam

Department: CS
Course Name: Machine Learning
Course Code: CS462
Instructor(s): Pro. Abeer ElKorany
Name:………………………………………

Date: 21/11/2023
Duration: 1 hour
Total Marks: 12
ID:…………………………..

تعليمات هامة

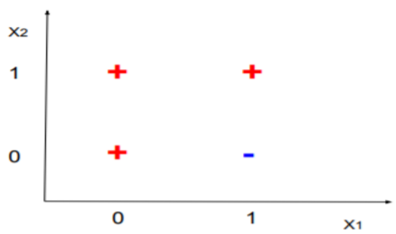
- حيازة التليفون المحمول مفتوحا داخل لجنة الإ\امتحان يعتبر حالة غش تستوجب العقاب وإذا كان ضرورى الدخول بالمحمول فيوضع مغلق فى الحقائب.
- لا يسمح بدخول سماعة الأذن أو البلوتوث.
- لايسمح بدخول أي كتب أو ملازم أو أوراق داخل اللجنة والمخالفة تعتبر حالة غش.

[4 marks, ½ for each]

1) The number of parameters in a parametric model is fixed, while the number of parameters in a nonparametric model grows with the amount of training data.	**T**
2) Naïve based is consider as an eager learner as it d*oes lot of work on training data*.	**T**
3) logistic regression is a special case of linear regression.	**F**
4) Max gain is the only method that could be used *to select best feature in decision tree*	**F**
5) KNN is considered as a generative Machine learning Model while logistic regression is considered as Discriminative Model	**F**
6) The Gradient descent will always find the global optimum	**F**
7) Overfitting Indicates limited generalization	**T**
8) KNN is more suitable in multi-classification problem than Logistic regression	**T**

Question 2 [4 marks]

In the following Figure, there are three positive samples ("+" for Y = 1) and one negative samples ("-" for Y = 0). Answer the following questions.



X1	*X2*	*y*
0	*0*	*1*
1	*0*	*0*
0	*1*	*1*
1	*1*	*1*

a) What is the suitable machine learning technique to be used with this dataset: Logistic Regression or Linear Regression? Explain why. [1 Mark]

Solution: Logistic Regression. Because Logistic Regression predicts values between 0 and 1, which is consistent with the target space Y, but Linear Regression predicts any values.

b) What is the number of estimated parameters that we need to calculate using the following ML techniques (explain, list those parameters :- . [3 Mark]

1. Naïve bayes **(1.5 mark)**

2 paramètres for Class Priors: Y1, Y0 = $P(Y=1)$, $P(Y=0)$ (*½ mark)* +

Feature Probabilities: (***1 mark)***

For discrete features *(as per the above table)*:

$P(X1=0|Y=1)$, $P(X1=1|Y=1)$, $P(X1=1|Y=0)$, $P(X2=0|Y=1)$,$P(X2=1|Y=1)$, $P(X2=0|Y=0)$

For continuous features :X1 and X2

mean$(X1|Y=1)$, *std*$(X1|Y=1)$, *mean*$(X1|Y=0)$, *std*$(X1|Y=0)$, *mean*$(X2|Y=1)$, *std*$(X2|Y=1)$, *mean*$(X2|Y=0)$, *std*$(X2|Y=0)$ *Note: std or variance can be accepted*

2. Logistic regression **(1 Mark)**

3 parameters (X1, X2, Base)

3. linear regression (***½ mark)***

this is a classification problem and can't use linear regression to solve it

Question 3 [4 marks]

For the following dataset

Given Birth	Can Fly	Live in Water	Have Legs	Mammal
Yes	No	Yes	No	Yes
Yes	Yes	No	Yes	Yes
No	No	Yes	No	Yes
No	Yes	No	Yes	No
Yes	No	No	Yes	No
Yes	Yes	Yes	No	Yes
No	No	Yes	Yes	No
No	Yes	No	No	Yes
Yes	No	No	Yes	No
No	Yes	Yes	Yes	Yes

Use the following ML techniques to predict whether this animal with the following features will be mammals or non-mammals.

Birth	Fly	Water	Legs	Mammal	Distance	Difference	
1	o	1	o	1	$sqrt((1-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2) = 0$	0	1st neighbor
1	1	O	*1*	1	$sqrt((1-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2) = \sqrt{3}$	$\sqrt{3}$	
O	o	*1*	o	1	***$sqrt((0-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2) = \sqrt{1}$***	**1**	2nd neighbor
O	*1*	o	1	o	***$sqrt((0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2) = \sqrt{4} = 2$***	**2**	
1	o	O	1	o	***$sqrt((1-1)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2) = \sqrt{2}$***	$\sqrt{2}$	4th neighbor
1	1	1	o	1	***$sqrt((1-1)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2) = \sqrt{1}$***	1	3rd neighbor
O	o	1	1	o	***$sqrt((0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2) = \sqrt{2}$***	$\sqrt{2}$	5th neighbor
O	1	O	o	1	***$sqrt((0-1)^2 + (1-0)^2 + (0-1)^2 + (0-0)^2) = \sqrt{3}$***	$\sqrt{3}$	
1	o	O	1	o	***$sqrt((1-1)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2) = \sqrt{2}$***	$\sqrt{2}$	
O	1	1	1	1	***$sqrt((0-1)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2) = \sqrt{3}$***	$\sqrt{3}$	

X= (Given Birth= "Yes", Can Fly= "no", Live in water = "Yes", Have legs="no")

- KNN with K = 5

- Convert yes/no to binary 1,0 (***½ mark)***
- Calculate the Euclidean distance between X and all training instances. (***1 mark)***
- **Find the 5 nearest neighbors and take the majority's class.** ***(½ mark)***

X= (Given Birth= "Yes", Can Fly= "no", Live in water = "Yes", Have legs="no")
X= 1 0 1 0

The 5 nearest neighbors are highlighted in green.
Class (majority's vote): Yes -> mammal

-For naïve bays

1- Calculate the prior probabilities of each class. ***½ mark***

P(Mammal)=Number of Mammal instances/Total number of instances = 0.6
P(Not Mammal)=Number of Not Mammal instances/Total number of instances =0.4

2- **Likelihoods:** Calculate the likelihoods of each feature given each class. ***1 Mark***

P(Given Birth=Yes|Mammal)= 3/6	P(Given Birth=Yes|Not-Mammal)= *2/4 = 1/2*
P(Can Fly=No|Mammal) = 2/6= 1/3	*P*(Can Fly=No| Not-Mammal) = ¾
P(Live in Water=Yes|Mammal)= 4/6=2/3	P(Live in Water=Yes| Not-Mammal)= ¼
p(Have Legs=No|Mammal)= 4/6 = 2/3	p(Have Legs=No| Not-Mammal)= *0/4 = 0*

½ mark

$P(\text{Mammal}|X)=$ ***(3/6) * (2/6) * (4/6) * (4/6) * 0.6 = 0.044***
$P(\text{Not Mammal}|X) =$ ***0***

Maximum posterior: 0.044, Class: Yes -> mammal

We can accept the answer if the student used m-estimate:
Let m = 1 (a different value can be accepted if calculations are correct)

1- **Likelihoods:** Calculate the likelihoods of each feature given each class.

P(Given Birth=Yes|Mammal)= 3.5/7	P(Given Birth=Yes|Not-Mammal)= 2.5/5
P(Can Fly=No|Mammal) = 2.5/7	*P*(Can Fly=No| Not-Mammal) = 3.5/5
P(Live in Water=Yes|Mammal)= 4.5/7	P(Live in Water=Yes| Not-Mammal)= 1.5/5
p(Have Legs=No|Mammal)= 4.5/7	p(Have Legs=No| Not-Mammal)= 0.5/5

$P(\text{Mammal}|X)=$ ***0.044***
$P(\text{Not Mammal}|X) =$ ***0.0042***

Maximum posterior: 0.044
Class: Yes -> mammal