Cairo University
Faculty of Computers and Artificial Intelligence

Midterm Exam

Department: CS
Course Name: Machine Learning **Date: 1/12/2021**
Course Code: CS467 **Duration: 1 hour**
Instructor(s): Dr. Hanaa Bayomi **Total Marks: 20**
Name:……………………………………… **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.

Question 1 ***[5 marks]***

- <u>Answer the following Questions:</u>

1. Assume the following data

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Construct a parametric classifier using Naïve byes to predict whether this person with a new instance

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

Will be mammals or non-mammals.

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

Question 2 Mark each statement with T or F in the right side: ***[5 marks]***

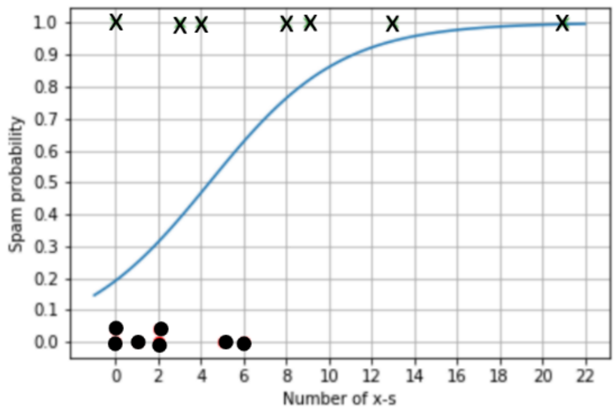
1) *We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.*	**(F)**
2) *When a decision tree is grown to full depth, it is more likely to fit the noise in the data.*	*(T)*
3) *When the feature space is larger, over fitting is more likely.*	*(T)*
4) *5-NN is more robust to outliers than 1-NN.*	*(T)*
5) *Since classification is a special case of regression, logistic regression is a special case of linear regression.*	**(F)**
6) *The Gradient descent will always find the global optimum*	**(F)**
7) *Overfitting Indicates limited generalization*	**(T)**
8) *In Support Vector Machines (SVM) ,Inputs are mapped to lower dimensional space where data becomes likely to be linearly separable*	**(F)**
9) *When the trained system matches the training set perfectly, overfitting may occur*	**(T)**
10) *Algorithms for supervised learning are not directly applicable for unsupervised learning*	**(T)**

Question 3 ***[5 marks]***

Kim is building a spam filter. She has the hypothesis that counting the occurrences of the letter 'x' in the e-mails will be a good indicator of spam or no-spam. She collects 7 spam messages and 7 no spam messages and counts the number of x-s in each. Here is what she finds.

- Number of 'x'-s in each spam: [0, 3, 4, 8, 9, 13, 21]
- Number of 'x'-s in each no-spam: [0, 0, 1, 2, 2, 5, 6]

She trains a logistic regression classifier on the data and plots the classifier against the data.



a) How many x-s must an e-mail contain to guarantee it is a spam mail?

You can never be 100% sure with a logistic regression model.

b) How is a logistic regression model normally turned into a binary classifier? If you turn the model into a classifier in this way, what is the accuracy of the classifier on the training data?

This is normally done by choosing the class 1 if P(1 | x) > 0.5.
We see from the graph that this classify 4 spams correctly and 3 spams incorrectly and 5 no-spams correctly and 2 incorrectly. Altogether 9 out of 14 are classified correctly, yielding and accuracy of 9/14.

c) Can use the SVM to solve this problem? explain. if you use it what is the training error rate after using SVM?

Yes, because this is a classification and linearly separable problem

Zero

Question 4 ***[5 marks]***

a) While minimizing a convex objective function using gradient descent, the algorithm does not converge even after 10,000 iterations. Mention any two reasons and the possible solutions?

1) Very small learning rate: increase learning rate
2) Data is not normalized: perform normalization

b) The training error of 1-NN classifier is 0. (true/false) Explain

True: Each point is its own neighbor, so 1-NN classifier achieves perfect classification on training data.

c) We consider the following models of logistic regression for a binary classification with a sigmoid function

$$g(z) = \frac{1}{1+e^{-z}}$$

- Model 1: $P(Y=1 \mid X, w_1, w_2) = g(w_1 X_1 + w_2 X_2)$
- Model 2: $P(Y=1 \mid X, w_1, w_2) = g(w_0 + w_1 X_1 + w_2 X_2)$

We have three training examples:

$$\begin{array}{lll} x^{(1)} = [1,1]^T & x^{(2)} = [1,0]^T & x^{(3)} = [0,0]^T \\ y^{(1)} = 1 & y^{(2)} = -1 & y^{(3)} = 1 \end{array}$$

Does it matter how the third example is labeled in Model 1? i.e., would the learned value of w = (w1, w2) be different if we change the label of the third example to -1? Does it matter in Model 2? Briefly explain your answer. (Hint: think of the decision boundary on 2D plane.)

It does not matter in Model 1 because x $^{(3)}$ = (0, 0) makes $w_1x_1 + w_2x_2$ always zero and hence the likelihood of the model does not depend on the value of w. But it does matter in Model 2.

Good Luck

Dr. Hanaa Bayomi