Cairo University
Faculty of Computers and Artificial Intelligence

Midterm Exam

Department: CS
Course Name: Machine Learning
Course Code: CS467
Instructor(s): Dr. Hanaa Bayomi
Name:…………………………………….

Date: 19/11/2019
Duration: 1 hour
Total Marks: 20
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 ***[10 marks]***

- **Answer the following Questions:**

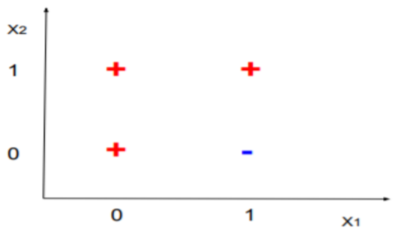
a. [2 marks] Suppose we have trained a linear regression model y = ax+b where a = 0.5 and b = 1.0, on a set of training data points D = {(1.0,1.6),(1.5,1.5),(3.0,2.4)}. Calculate the mean squared errors of this model on D.

Solution: $MSE = \frac{0.1^2+0.25^2+0.1^2}{6} =$ **0.01375**

b. [2 marks] Briefly describe the difference between a maximum likelihood hypothesis and a maximum a posteriori hypothesis.

Solutions:
ML: maximize the data likelihood given the model, i.e., $\arg\max_W P(Data|W)$
MAP: $\arg\max_W P(W|Data)$

c. In the following Figure, we show three positive samples ("+" for Y = 1) and one negative samples ("-" for Y = 0). Complete the following questions.



1. [1.5 marks] Which model is better: Logistic Regression or Linear Regression? Explain why.

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.

2. [1.5 marks] Is there any logistic regression classifier using X1 and X2 that can perfectly classify the examples in Figure? How about if we change label of point (0, 1) from “+” to”-”?

Solution: Logistic regression forms linear decision surface. Because data points in Figure is linear separable, they can be perfectly classified. But if we make the change, then data points are not linearly separable so no logistic regression classifier can perfectly classify the examples.

d. [3 marks] Consider Y (output) Binary , X_i (features) continuous, X=< X_1 , X_2 ,…….., X_n > , the number of estimated parameters in :-

1. Naïve bayes

   4n+2

2. in logistic regression

   n+1

3. in linear regression

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

Question 2 Mark each statement with T or F in the right side: ***[4 marks]***

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) As model complexity increases, bias will decrease while variance will increase.	**T**
3) Consider a cancer diagnosis classication problem where almost all of the people being diagnosed don't have cancer. The probability of correct classication is the most important metric to optimize.	**F**

4) *For logistic regression, with parameters optimized using a gradient method, setting parameters to 0 is an acceptable initialization.*	**T**
5) *In logistic regression, The optimal weight vector can be found using MLE.*	**T**
6) *For small training sets, Naive Bayes generally is more accurate than logistic regression*	**T**
7) *When a decision tree is grown to full depth, it is more likely to fit the noise in the data.*	**T**
8) *When the feature space is larger, over fitting is more likely.*	**T**

Question 3 ***[6 marks]***

We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

1.[1 points] What is the entropy H(Passed)?

★ **ANSWER:**

$$H(Passed) = -(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6})$$

$$H(Passed) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$$

$$H(Passed) = \boxed{\log_2 3 - \frac{2}{3} \approx 0.92}$$

2.[1 point] What is the entropy H(Passed j GPA)?

★ **ANSWER:**

$$H(Passed|GPA) = -\frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(1\log_2 1)$$

$$H(Passed|GPA) = \frac{1}{3}(1) + \frac{1}{3}(1) + \frac{1}{3}(0)$$

$$H(Passed|GPA) = \boxed{\frac{2}{3} \approx 0.66}$$

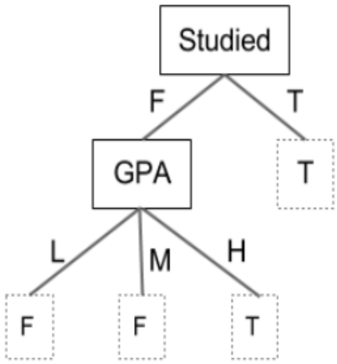
3.[1 point] What is the entropy H(Passed j Studied)?

★ **ANSWER:**

$$H(Passed|Studied) = -\frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) - \frac{1}{2}(1\log_2 1)$$
$$H(Passed|Studied) = \frac{1}{2}(\log_2 3 - \frac{2}{3}$$
$$H(Passed|Studied) = \boxed{\frac{1}{2}\log_2 3 - \frac{1}{3} \approx 0.46}$$

4. [3 points] Draw the full decision tree that would be learned for this dataset.

★ **ANSWER:** We want to split first on the variable which maximizes the information gain $H(Passed) - H(Passed|A)$. This is equivalent to minimizing $H(Passed|A)$, so we should split on "Studied?" first.



$$j(\theta) = \sum_{1}^{m}\left[y_i \log\left(\frac{1}{1+e^{-\theta^t x}}\right) + (1-y_i)\log\left(1-\frac{1}{1+e^{-\theta^t x}}\right)\right]$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$H(S) = -p_{(+)}\log_2 p_{(+)} - p_{(-)}\log_2 p_{(-)}$$

$$Gain(S,A) = H(S) - \sum_{V \in Values(A)} \frac{|S_V|}{|S|} H(S_V)$$

$$\hat{P}(a_{jk} \mid c_i) = \frac{n_c + mp}{n + m}$$

$$SplitEntropy(S,A) = -\sum_{V \in Values(A)} \frac{|S_V|}{|S|}\log\frac{|S_V|}{|S|}$$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitEntropy(S,A)}$$

$$\hat{P}(x_j \mid c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}}\exp\left(-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

$$d(x,y) = \sum_{i=1}^{k}(x_i - y_i)^2$$

$$P(x_1, x_2, \ldots, x_n \mid C) = P(x_1 \mid C) \times P(x_2 \mid C) \times \ldots \times P(x_n \mid C)$$

$$P(x_1, x_2, \ldots, x_n \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$$

Good Luck

Dr. Hanaa Bayomi