University of Massachusetts Dartmouth

Department of Computer and Information Science

CIS 490 Machine Learning – Exam I (Spring 2022)

Tuesday, March 15, 2022

Printed Full Name: \_\_\_\_\_Pranav Vinod\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student ID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_01984464\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

DO NOT TURN THE PAGE OVER UNTIL YOU ARE INSTRUCTED TO DO SO

Please read the following instructions:

1. You have 75 minutes to complete the examination.
2. This examination is OPEN Notes
3. Type your answer in space provided on the examination sheets, any work not on the examination sheets will not be graded.
4. Type your answers legibly.
5. Submit your answer according to the instruction for grading by the end of the examination.
6. DO NOT communicate any of your classmates during the examination.

Honor Policy: copying in whole or in part of the examination will be considered to be an act of scholastic dishonesty. Students who violate university rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the university. Since such dishonesty harms individuals, all students, and the integrity of the university, policies on scholastic dishonesty will be strictly enforced.

I have read the above instructions and I will act in accordance with all of them.

Pranav Vinod

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_03/15/2022\_\_\_\_\_\_\_\_\_\_\_\_

Student Signature Date

Type your name and date to agree the policy before you start!

This examination contains three sections. The whole Exam I carries 100 points.

**Section I. Single-Choice Questions (20 points, 2 points per question; only ONE choice is correct). Please write your answers in the table provided below.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Question** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Answer | b | a | a | b | d | b | a | a | a | a |

1. If we have a PDF expressed as , identify which probability distribution does this PDF describe?



1. Poisson
2. Normal
3. Uniform
4. Gamma
5. If we have a PMF expressed as for x = 0, 1,2,… where λ is the shape parameter which indicates the average number of events in the given time interval, which probability distribution has this PMF?



1. Poisson
2. Normal
3. Uniform
4. Gamma
5. Given a random variable, Gender, which has two values, (0=male; 1= female), this variable should be regarded as:
6. Discrete
7. Continuous
8. Given a random variable, GRE scores, this variable should be regarded as:
9. Discrete
10. Continuous
11. Which method only grows One big tree?
12. Bagging
13. Random forests
14. Boosting
15. Decision trees (CART)
16. Which model or graph looks more like overfitting?
17. -2.2 + 3.1 X – 0.30 X2
18. -1.1 + 4,700,910.7 X – 8,585,638.4 X2 …

1. Given MSE is fixed, MSE = Variance + Bias2

if we decrease the variance, what would happen to bias?

1. Increase
2. Decrease
3. No change
4. Given a model f(x), if we re-measure response variable Y, would stochastic or deterministic noise change?
5. stochastic noise
6. deterministic noise
7. What is the major difference between the two regularization types, L1 and L2?
8. L1 seeks sparse solution
9. L2 seeks sparse solution
10. L1 penalizes the complexity of coefficients
11. L2 penalizes the complexity of coefficients
12. If Classifier 1 with AUC =.95, Classifier 2 with AUC =0.70, and Classifier 3 with AUC =0.50, which classifier is better?
13. Classifier 1
14. Classifier 2
15. Classifier 3

**Section II. True or False questions (20 points, 2 points per question).**

|  |  |  |
| --- | --- | --- |
| Questions | True | False |
| 1. Supervising learning methods only work on attributes (X), not outcome (Y) |  | False |
| 1. If outcome Y is a binary variable, we consider linear regression |  | False |
| 1. If outcome Y has more than two categories, we should consider logistic regression |  | False |
| 1. If outcome Y are counts (e.g., 2, 3, 4, 5, etc), the underlying distribution of Y is likely to be Bernoulli |  | False |
| 1. The probability of distribution of continuous random variables is described by probability density function | True |  |
| 1. Standard normal distribution is a special case of normal distribution when μ=1 and σ2=1. |  | False |
| 1. If a classifier has its ROC falls on the 45o line, then this classifier is the best compared to the one with its ROC hugging the top left corner. |  | False |
| 1. The kurtosis of a Gaussian distribution is negative and the skewness of a Gaussian distribution is 0 |  | False |
| 1. To check the model fit of linear regression, we could examine residual standard error, R-squared, adjusted R-squared, F-statistic and MSE | True |  |
| 1. If outcome Y is discrete, we use regression tree |  | False |

**Section III. Short problems (60 points)**

1. **(10 points)** Let’s assume the number of spams follows the Poisson distribution and we randomly draw samples of size =50, with the mean of sample means of 4. Estimate the population mean and the population standard deviation based on CLT.

Note: Show you step-wise calculation. If you only give an answer, no scores.

Your Answer:

For a Poisson distribution, we know that the mean of the population is

And the variance () of the population is

Since CLT holds true, as the size of sample > 30,

4

Also, since the underlying population is a Poisson distribution, the SD of sample means = ,

Therefore, SD of sample means = 2

2 =

=

**Population mean = 4**

**Population standard deviation =**

1. **(10 points)** Given two random variables X (0 = male; 1 = female) and Y (0= low risk; 1= medium risk; 2 = high risk) with joint pmf given in the Table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y=0 | Y=1 | Y=2 |
| X=0 | 1/25 | 1/10 | 1/5 |
| X=1 | 2/5 | 4/25 | 1/10 |

Compute

1. p (X = female, Y= high risk) = ? (3 points)
2. p( X = female ) = ? (3 points)
3. p (Y= high risk|X=female) = ? (4 points)

Note: Show you step-wise calculation. If you only give an answer, no scores.

Your Answer:

1. p (X = female, Y = high risk) = p (X = 1, Y = 2) = 1/10
2. p (X = female) =

= p (X = 1, Y = 0) + p (X = 1, Y = 1) + p (X = 1, Y = 2)

= 2/5 + 4/25 + 1/10

= 33/50

1. p (Y= high risk |X=female) = p (Y=high risk, X = female) / p (X = female)

= (1/10) / (33/50)

= 5/33

1. **(10 points)** Given historical credit scores, gender, age and family income, a credit card company would like to predict the credit risk of a new customer and decide if they should issue a credit card to this new customer.
2. Which machine learning method would this company use to help their decision, unsupervised or supervised? (3points)

Your Answer: Based on historical credit scores a decision can be made on whether to issue a credit card to a new customer. Therefore, since outcome variable is given, we should use supervised machine learning method to take the decision.

1. Please justify your decision based on your understanding of unsupervised or supervised learning methods in this case study. (3 points)

Your Answer: Supervised learning methods are those in which the outcome variable is also given along with the predictor variables.

Unsupervised learning methods are those in which the outcome variable is absent, and the predictor variables are present.

In this problem, we have been given historical credit scores. We can create a threshold of credit scores and say that any score below the threshold is high risk and any score above the threshold is low risk. By creating these 2 categories we can change this into a classification problem.

In the given problem, our outcome variable, which is credit score, is present. Therefore, we can use supervised learning methods to model this problem.

1. What specific supervised learning methods/models you would like to propose to your supervisor (4 points)

Your Answer: Since the decision to be made is a binary decision, i.e., should credit card be given or not, we can use machine learning methods for classification to implement in this case. Particularly, I would propose logistic regression to solve this problem since we require more accuracy in our predictions.

1. **(8 points, 2points \* 4)** Suppose that 9 out of 10 males are admitted to an engineering school while 5 of 10 females are admitted.

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| Admitted | 9 | 5 |
| Not Admitted | 1 | 5 |

Note: Show you step-wise calculation. If you only give an answer, no scores.

1. Compute the probability for admitting a male and the probability for not admitting a male?

Your Answer:

P (admitting male) = number of admitted male / total number of applications

= 9/20

P (not admitting male) = number of non-admitted male / total number of applications

= 1/20

1. Compute the probabilities for admitting a female and for not-admitting a female?

Your Answer:

P (admitting female) = number of admitted female / total number of applications

= 5/20 = 1/4

P (not admitting female) = number of non-admitted female / total number of applications

= 5/20 = 1/4

1. Compute the admission odds for males and females, respectively?

Your Answer:

Odds (male) = P (admitting male) / P (not admitting male)

= (9/20) / (1/20)

= 9

Odds (female) = P (admitting female) / P (not admitting female)

= (1/4) / (1/4)

= 1

1. Compute the odds ratio for admission for males and females?

Your Answer:

Odds ratio = odds (male) / odds (female)

= 9/1

1. **(12 points)** Write pseudo-code or steps of 5-fold cross-validation for choosing the best tuning parameter in L1 regularized linear regression.

Your Answer:

The following are the steps for a 5-fold cross validation for choosing the best tuning parameter in L1 regularized linear regression:

1. Establish a range of values to run the CV algorithm over.
2. Divide the dataset into 5 partitions.
3. Use 4 of them as the training set and the remaining one as the testing set.
4. Run L1 regularized linear regression over the 4 training sets and fit the model on the remaining testing set.
5. Get the value of MSE from the testing set and name it .
6. Now repeat steps 4 and 5 using a different testing set and the other remaining sets as the training set 4 times to get .
7. Now get cross validation error as follows:

1. Repeat steps 2 – 7 for all values of within the range established.
2. Use the which corresponds to the least CV error out of all calculated.
3. **(10 points; 2 points \*5)** Given the confusion matrix for classification evaluation

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Target |  |
|  |  | Y=1 | Y=0 |
| Predicted | Y=1 | 100 | 10 |
|  | Y=0 | 20 | 40 |

TP = 100

FP = 10

FN = 20

TN = 40

1. Compute classification accuracy

Your Answer: TP + TN / TP + FP + TN + FN

= 140/170

1. Compute Sensitivity

Your Answer: TP/ TP + FN

= 100/120 = 5/6

1. Compute Specificity

Your Answer: TN / FP + TN

= 40/ 50 = 4/5

1. Compute Positive predictive value

Your Answer: TP / TP + FP

= 100/110

1. Compute Negative predictive value

Your Answer: TN / TN + FN

= 40/60 = 2/3