University of Massachusetts Dartmouth

Department of Computer and Information Science

CIS 530-02 Advanced Data Mining – Exam I (Spring 2024)

Tuesday, March 19, 2024

Printed Full Name: Banoth Jeevan Kumar

Student ID: 02105145

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 materials, including notes, slides and books.
3. Type your answer in the 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.

\_\_\_\_B Jeevan Kumar\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_March 19th, 2024\_\_\_\_\_\_\_\_\_\_\_

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 | a | 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 l 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 the best among the three?
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 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 | True |  |

**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:**

Spam Count Distribution and Sample Analysis: Spams follow a Poisson distribution.

We have random samples of size 50 with a known mean of sample means (4).

Here's a breakdown of the analysis:

1. Population Mean Estimation: The population mean is assumed to be the same as the mean of the sample means, which is 4. Therefore, we estimate the population is meant to be 4.

2. Population Standard Deviation Estimation:

Since the sample size is large, the distribution of sample means approximates a normal distribution. The standard deviation of this distribution is calculated as: **Population Standard Deviation / √Sample Size**

This formula helps us estimate the standard deviation of the population from the sample mean and sample size.

Given that the mean of the sample means is 4 and the sample size is 50,

we can estimate the population standard deviation using the formula:

population standard deviation = sample mean / √sample size. - Substituting the values: **population standard deviation = 4 / √50 ≈ 0.5657.**

Therefore, based on the Central Limit Theorem, the estimated population mean is 4 and the estimated population standard deviation is approximately **0.5657** when considering the Poisson distribution and random samples of size 50 with a mean of sample means of 4.

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:

## **Computing Probabilities from the Joint PMF Table**

**Calculations:**

**P (X = female, Y = high risk):**

This represents the probability of a female (X = 1) and high risk (Y = 2) occurring together.

From the table, this value is at the intersection of the row Y = 2 and the column X = 1.

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

**P (X = female):**

This represents the probability of being female (X = 1) regardless of the risk level (Y).

We need to sum the probabilities for X = 1 across all Y values.

So, p (X = female) = (2/5) + (4/25) + (1/10) = 11/25.

**P (Y = high risk | X = female):**

This represents the conditional probability of high risk (Y = 2) given that the person is female (X = 1). We can use the formula for conditional probability:

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

We already calculated both these values in steps 1 and 2. Substituting:

P (Y = high risk | X = female) = (1/10) / (11/25) = 1/11 (approximately 0.091).

**Answer:**

* P (X = female, Y = high risk) = 1/10
* P (X = female) = 19/25
* P (Y = high risk |X = female) = 5/38 = 0.091 Aprox

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 data mining method would this company use to help their decision, unsupervised or supervised? (3points)

Your Answer:

The credit card company would use a supervised data mining method to help make decisions regarding issuing credit cards to new customers.

Supervised Learning: In supervised learning, a model is trained on data where each feature (e.g., credit score, gender) is paired with a target variable (e.g., credit risk).

This training allows the model to grasp the relationship between the features and the intended outcome.

Credit Card Scenario: To predict credit risk based on features, the company needs a supervised learning approach. The model studies historical data to learn the connections between the features and credit risk.

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

Your Answer:

Supervised Learning: - Supervised learning involves training a model on labeled data where the algorithm learns the relationship between input variables (features like credit scores, gender, age, and family income) and the target variable (credit risk).

- In this scenario, historical data with labeled credit risk information can be used to train a supervised learning model to predict credit risk for new customers.

- By using supervised methods, the model can learn from past data to make predictions about the credit risk of new customers, enabling the company to make informed decisions about issuing credit cards.

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

Your Answer:

When proposing specific supervised methods/models to predict credit risk for new customers based on historical credit scores, gender, age, and family income, the credit card company could consider the following options:

Logistic Regression: - Logistic regression is a commonly used model for binary classification tasks like predicting credit risk (low or high risk).

- It can estimate the probability of a new customer being in a particular credit risk category based on their credit scores, gender, age, and family income.

- The model provides interpretable results and is relatively simple to implement, making it a practical choice for this scenario.

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.

Since the table shows counts for admitted and not admitted for both males and females, we can find the total number of applicants (both males and females) by summing all the values in the table:

Total Applicants = Admitted (Male) + Not Admitted (Male) + Admitted (Female) + Not Admitted (Female)

Total Applicants = 9 + 1 + 5 + 5 = 20

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

Your Answer:

To compute the probability of admitting a male:

* + **Total Males**: The total number of males is *9+1*=10.
  + **Males Admitted**: 9 males are admitted.
  + *P*(Admit Male)=Males Admitted/ Total males = 9/10

To compute the probability of not admitting a male:

Males not admitted = 1 male not admitted

P(not admitted males) = males not admitted / total males = 1/10

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

Your Answer:

**Probability of Admitting a Female:**

P(Admitted | Female) = Number of Admitted Females / Total Females

P(Admitted | Female) = 5 / (5 + 5) = 1/2

**Probability of Not Admitting a Female:**

P(Not Admitted | Female) = Number of Not Admitted Females / Total Females

P(Not Admitted | Female) = 5 / (5 + 5) = 1/2

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

Your Answer:

**Admission Odds:**

The odds ratio represents the ratio of the probability of an event happening to the probability of it not happening. We can calculate the admission odds for males and females separately.

**Admission Odds for Males:**

Odds (Admitted | Male) = P(Admitted | Male) / P(Not Admitted | Male)

Odds (Admitted | Male) = (9/10) / (1/10) = 9

**Admission Odds for Females:**

Odds (Admitted | Female) = P(Admitted | Female) / P(Not Admitted | Female)

Odds (Admitted | Female) = (1/2) / (1/2) = 1

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

Your Answer:

**Odds Ratio:**

The odds ratio compares the odds of admission for males and females.

Odds Ratio = Odds (Admitted | Male) / Odds (Admitted | Female)

Odds Ratio = 9 / 1 = 9

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

Your Answer:

Steps for 5-Fold Cross-Validation for Tuning L1 Regularized Linear Regression:

1. Data Split: Divide the dataset into five approximately equal portions called "folds."

2. Parameter Iteration: - Select a range of candidate tuning parameter values.

- For each value: - Create a list to hold cross-validation scores.

3. Fold Loop: - For each fold (1 to 5): - Reserve the current fold for validation.

- Merge the remaining folds to form the training set.

4. Model Training: Train an L1 regularized linear regression model on the training set using the current tuning parameter value.

5. Evaluate Model: - Calculate the performance metric (e.g., mean squared error, R-squared) on the validation set for this fold.

6. Store Cross-Validation Score: - Add the performance metric to the list of cross-validation scores for the current tuning parameter.

7. Average Scores: - Calculate the average of the cross-validation scores obtained for the current tuning parameter.

8. Select Best Parameter: - Repeat steps 3 to 7 for each candidate tuning parameter to compare their average cross-validation scores.

9. Choose Optimal Parameter: - Select the tuning parameter that yields the best average performance across all folds.

10.Train Final Model: - Train the L1 regularized linear regression model using the selected optimal tuning parameter on the entire dataset.

By following these steps, we can effectively use 5-fold cross-validation to choose the best tuning parameter for L1 regularized linear regression.

1. **(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 |

1. Compute classification accuracy

Your Answer:

Accuracy represents the overall proportion of correctly classified instances.

**Calculation:**

Accuracy = (True Positives + True Negatives) / Total Samples

From the matrix:

True Positives (TP) = 100

True Negatives (TN) = 40

Total Samples = TP + TN + False Positives (FP) + False Negatives (FN) = 100 + 10 + 20 + 40 = 170

Accuracy=100+10+20+40/100+40

=170/140

≈0.824

1. Compute Sensitivity

Your Answer:

Sensitivity, also known as Recall, measures the proportion of correctly classified positive cases (Y=1 actually predicted as Y=1).

**Calculation:**

Sensitivity = TP / (TP + FN)

**Sensitivity = 100 / (100 + 20) = 83.33% (approximately)**

1. Compute Specificity

Your Answer:

Specificity measures the proportion of correctly classified negative cases (Y=0 actually predicted as Y=0).

**Calculation:**

Specificity = TN / (TN + FP)

**Specificity = 40 / (40 + 10) = 80% (approximately)**

1. Compute Positive predictive value

Your Answer:

PPV measures the proportion of true positives among all predicted positives (cases classified as Y=1).

**Calculation:**

PPV = TP / (TP + FP)

**PPV = 100 / (100 + 10) = 90.91% (approximately)**

1. Compute Negative predictive value

Your Answer:

NPV measures the proportion of true negatives among all predicted negatives (cases classified as Y=0).

**Calculation:**

NPV = TN / (TN + FN)

**NPV = 40 / (40 + 20) = 66.67% (approximately)**