**Exam 02 – Part2. DSCI 35600 (70 points)**  Name\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Problem 1 (10 pts).** Assume that you are provided with three 2D arrays, named X\_train, X\_val, and X\_test. The contents of these arrays are provided below.

X\_train X\_val X\_test

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | 23 |  | 8 | 15 |  | 4 | 12 |
| 5 | 25 |  | 7 | 16 |  | 9 | 24 |
| 8 | 22 |  |  |  |  |  |  |
| 4 | 16 |  |  |  |  |  |  |

Complete the tables below to show the contents of the arrays Xs\_train, Xs\_val, and Xs\_test after executing the following code:

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

scaler.fit(X\_train)

Xs\_train = scaler.transform(X\_train)

Xs\_val = scaler.transform(X\_val)

Xs\_test = scaler.transform(X\_test)

X\_train X\_val X\_test

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | .77 |  | 1 | -.11 |  | -.33 | -44 |
| 0.5 | 1 |  | .83 | 0 |  | 1.16 | .88 |
| 1 | .66 |  |  |  |  |  |  |
| .33 | 0 |  |  |  |  |  |  |

**Problem 2 (10 pts).** Assume you are provided with a dataset containing 5 observations and three features to use in creating a regression model. The table below contains information for ONE such proposed model. The first four columns provide the proposed coefficient estimates for each model. Columns 5 – 9 contain the **predicted** values resulting from each of the models. The **true** values are provided in the bottom row of the table.

For the proposed model, calculate its LOSS using the **linear regression loss function**, the **ridge regression loss** function with , and the **LASSO regression loss** function with . In each case, use SSE rather than MSE as the “baseline” loss.

SHOW ALL THE FORMULAS AND THE COMPUTATIONS!

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  | **Linear Regression** | **Ridge Regression** | **LASSO Regression** |
| 6 | 5 | -3 | -2 | 8.1 | 12.9 | 13.8 | 17.6 | 18.3 | **9.22** | **47.22** | **19.22** |
| True Values 🡪 | | | | 9.8 | 12.1 | 14.6 | 15.7 | 19.5 |  |  |  |

**LinREG -> SSE=sum( yhat-ytrue)^2=(8.1-9.8)^2 + (12.9-12.1)^2 + (13.8-14.6)^2 + (17.6-15.7)^2 + (18.3-19.5)^2 = 9.22**

**RIDGE -> SSE+alpha\*(beta1^2+beta2^2+beta3^2) = 9.22 + (5^2 + 3^2 + 2^2) = 47.22**

**LASSO -> SSE +alpha\*( abs(beta1) + abs(beta2) + abs(beta3) ) =9.22 + (5+3+2) = 19.22**

**Problem 3. (10 pts)** A training set for a classification task is provided below. The dataset has two features and one categorical label, , which has two possible classes: “red” and “blue”. You are asked to score the logistic regression model. **Round all answers to three decimal places on this problem, and show your work.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 5 | 5 | 7 |
|  | 4 | 1 | 7 | 8 |
|  | blue | blue | red | blue |

1. The logistic regression model is provided below. In the models, is an estimate for the probability that an observation falls into the red class. We let for red observations and let for blue observations. For each model, find and for each observation.

**Model 1:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | .917 | .870 | .786 | .690 |
|  | .083 | .130 | .786 | .310 |

1. Calculate the negative log-likelihood score for each model. Show how you used the appropriate formula.

**Negative Log-Likelihood = - sum( ln(pi^hat) ) = - (ln.083 + ln.130 + ln.786 + ln.310) = 5.941**

**Problem 4 (10 pts).** The confusion matrix for a test set in a classification problem with two classes is provided below. Find the precision and recall for each class, as well as the overall accuracy of the model.  **Show the formulas and the computations. Round to three decimal places.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **Class 0** | **Class 1** | | **Class 0** | 20 | 4 | | **Class 1** | 4 | 4 | | |  |  |  | | --- | --- | --- | |  | **Precision** | **Recall** | | **Class 0** | 20/24=.833 | 20/24=.833 | | **Class 1** | ­­­­4/8 = 0.5 | 4/8 = 0.5 | |

**Accuracy =** (TP+TN)/(TP+TN+FP+FN) = 24/32 = .75

**Problem 5 (10 pts).** A scatter plot consisting of FOUR points is shown below. Using the KNN classification algorithm trained on this dataset with *K*=3 figure out how to divide the given square into one BLUE (shade it) and one RED (leave it blank) region.

A picture containing text

Description automatically generated

**Problem 6.**  **(20 pts)** Write **True** or **False** next to each of the following statements.

1. It is important to scale features when constructing a ridge regression model. \_\_\_\_\_T\_\_\_\_\_\_\_
2. It is important to scale features when constructing a basic linear regression model. \_\_\_\_\_\_F\_\_\_\_\_\_
3. It is important to scale features when constructing a KNN model. \_\_\_\_\_\_T\_\_\_\_\_\_
4. The coefficients in a logistic regression model are trained by minimizing the sum of

squared errors objective function. \_\_\_\_F\_\_\_\_\_\_\_\_

1. We use negative log-likelihood rather than likelihood when score a logistic regression

model because negative log-likelihood is less affected by rounding issues. \_\_\_\_\_F\_\_\_\_\_\_\_

1. Training a regression model using the mean-squared error loss function will tend to

produce smaller coefficient estimates than training with sum of squared error loss. \_\_\_\_\_F\_\_\_\_\_\_\_

1. Increasing the hyper-parameter in a ridge regression model will typically make it

less likely that the model will overfit. \_\_\_\_\_\_T\_\_\_\_\_\_

1. Increasing the hyper-parameter *K* in a KNN model will typically make it less likely

that the model will overfit. \_\_\_\_\_T\_\_\_\_\_\_\_

1. The output of a logistic regression model is an estimate of the probability that the

provided observation belongs to a particular class. \_\_\_\_\_\_T\_\_\_\_\_\_

1. The advantage of the lasso algorithm over standard linear regression is that the lasso

algorithm will generally produce a model with a lower training SSE than linear regression. \_\_\_\_\_F\_\_\_\_\_\_\_

1. The lasso algorithm is a regularized version of least squares regression that tries to avoid

overfitting by applying a penalty to models based on the size of their coefficients. \_\_\_\_\_\_T\_\_\_\_\_\_

1. The “out-of-bag” training instances are used for validation of their predictor. \_\_\_\_\_\_T\_\_\_\_\_\_\_
2. At each node the RandomForestClassifier searches for the best feature among

a random subset of features instead of searching thru all features. \_\_\_\_\_\_T\_\_\_\_\_\_\_

1. Bagging is the method which uses the same training algorithm for every predictor

and trains them on different random subsets of the training set. \_\_\_\_\_\_\_T\_\_\_\_\_\_

1. Pasting is the method which uses the same training algorithm for every predictor

and trains them on different random subsets of the training set. \_\_\_\_\_\_T\_\_\_\_\_\_\_

1. Bagging uses sampling of the training set with replacement. \_\_\_\_\_\_T\_\_\_\_\_\_\_
2. Pasting uses sampling of the training set with replacement.  \_\_\_\_\_\_\_F\_\_\_\_\_\_