**CS412 Machine Learning - Homework 4 Linear Regression and Evaluation Metrics**

**Deadline:** 30 April 2020, 23:55

**Late submission:** till 2 May 2020, 23:55

(-10pts penalty for **each** late submission day)

**Submission**

For your notebook results, make sure to run all of the cells and the output results are there.

Please submit your homework as follows:

* Download the .ipynb and the .py file and upload both of them to sucourse.
* Submit also a single pdf document by solving questions on the sheet.
* Link to your Colab notebook (obtained via the share link in Colab) in the sheet:

**Objective**

The topic of this homework assignment is supervised learning. The first half is concerned with linear regression, and the second half, performance measure on classification tasks.

**Startup Code**

<https://colab.research.google.com/drive/1W80EpGJYudkQ7Sz2pbAHffvt9bo_ITHH>

To start working for your homework, take a copy of this folder to your own google drive.

**Software:** You may find the necessary function references here:

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html>

**Question 1: 75 pts - Predict the price of houses.**

**Dataset Description**

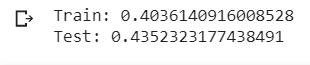
<https://raw.githubusercontent.com/OpenClassrooms-Student-Center/Evaluate-Improve-Models/master/house_prices.csv>

In this dataset, there are 2930 observations with 305 explanatory variables describing (almost) every aspect of residential homes.

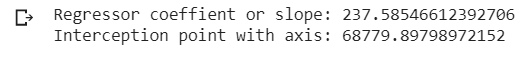
1. **Find the correlation between garage area and sale price by applying linear regression. Print the bias and slope. Print the train and test R2. Plot the test set with a scatter plot and add the linear regression model line.**

**From the linear regression we can see that as the garage area gets bigger, sale prices increase with it.**

**Train and Test r2:**

****

**Bias and Slope:**

****

**harita içeren bir resim

Açıklama otomatik olarak oluşturulduScatter Plot:**

1. **Apply multiple linear regression by taking all input features.** **Print the train and test R2.**

**Train and test r2 accordingly:**

****

1. **Comment on part a and b results. Why R2 is low in part a? Why test R2 is low although train R2 is quite high in part b?**

**In part a, we trained our linear regression model only with the “Garage Area” feature, therefore we missed the chance of identifying more related features that is affecting the Sales Price. Hence, we made an underfit for our model which results that our train and test r2 score became low. In the second part, since we used all the features in our dataset, our model became very complex and it memorized the train dataset. Hence, it became unsuccessful for predicting the outcome in the test data which results in overfitting problem. Therefore, our train score was quite high but the test score was really low. This is because, some features were not positively correlated in the dataset.**

1. **Apply ridge regression with cross-validation by taking all input features. Print optimal alpha. Print also the train and test R2.**

**Train and test r2: Optimal alpha:**

****

****

1. **Discuss on regularization. What is ridge regression? When do we use it? And what is the effect on features?**

**Ridge regression is a regularization technique in order to prevent the overfitting problem. We use ridge regression, when we want our weight coefficients in our regression model stable and reasonable so that particular features will not dominate our model. In ridge regression, there is a total cost function in order to penalize the extreme weights so that we overcome the overfitting problem. By appyling ridge regression we obtain weights that are closer to zero.**

1. **Print regression coefficients for multiple linear regression and ridge regression. Comment on the change of feature weights. What is the effect of ridge regression on feature weights?**

**Below is a partial image of regression coefficients of multiple regression:**

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Below is a partial image of ridge regression coefficients:**

**metin, gazete içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**As seen from the two tables, there is a significant decrease in the overall absolute value of the coefficients in the ridge regression part compared with the multiple linear regression part. In the ridge regression part, there are less number of extreme values for the weights, and they are more regularized. Hence, we conclude that that we are less likely to overfit in the regression part.**

**Question 2: 25 pts - Evaluation metrics.**

1. **15 pts - Provide the Confusion Matrix, Accuracy, Error, Precision, Recall, and F1-Score for the fruit classification problem. The output of test data classification results is given in the following table.**

Use both macro and micro averaging methods.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **mass** | **width** | **height** | **color\_score** | **class** | **prediction** |
| 154 | 7.1 | 7.5 | 0.78 | orange | lemon |
| 180 | 7.6 | 8.2 | 0.79 | orange | lemon |
| 154 | 7.2 | 7.2 | 0.82 | orange | apple |
| 160 | 7.4 | 8.1 | 0.80 | orange | orange |
| 164 | 7.5 | 8.1 | 0.81 | orange | apple |
| 152 | 6.5 | 8.5 | 0.72 | lemon | lemon |
| 118 | 6.1 | 8.1 | 0.70 | lemon | apple |
| 166 | 6.9 | 7.3 | 0.93 | apple | apple |
| 172 | 7.1 | 7.6 | 0.92 | apple | apple |

**Confusion matrix:**

**Gold Labels:**

**Orange Apple Lemon**

**Orange**

**System**

**Output Apple**

|  |  |  |
| --- | --- | --- |
| **1** | **0** | **0** |
| **2** | **2** | **1** |
| **2** | **0** | **1** |

**Lemon**

**General Accuracy: 4/9**

**Confusion matrix for 3 different classes:**

**Gold Labels**

**TO TN**

|  |  |
| --- | --- |
| **1** | **0** |
| **4** | **4** |

**System SO**

**Output SN**

**TA TN**

|  |  |
| --- | --- |
| **2** | **3** |
| **0** | **4** |

**SA**

**SN**

**TL TN**

|  |  |
| --- | --- |
| **1** | **2** |
| **1** | **5** |

**SL**

**SN**

TO: True Orange TA: True Apple

TN: True Negative SA: System Apple

SO: System Orange TL: True Lemon

SN: System Not SL: System Lemon

**Contingency table for micro-average:**

**True Positive True Negative**

|  |  |
| --- | --- |
| **4** | **5** |
| **5** | **13** |

**System Positive**

**System Negative**

**Micro-average: Macro-average:**

**Accuracy: 17/27 Accuracy: 17/27**

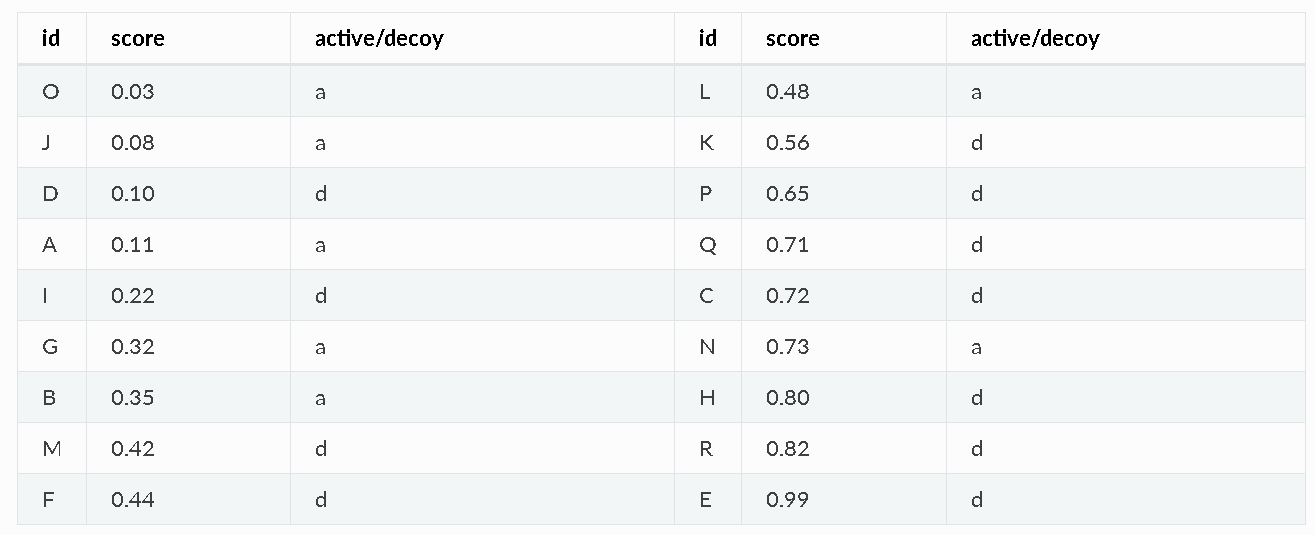
**Precision: 4/9 Precision: 26/45**

**Recall: 4/9 Recall: 17/30**

**Error: 10/27 Error: 10/27**

**F1 score: 4/9 F1 score: 2652/4635 = 0.57**

1. **10 pts - The table shows 18 data and the score assigned to each by a classifier. It is a binary classification problem. The active/decoy column shows the ground truth labels. Plot the corresponding ROC curve.**



**ROC CURVE**

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Colab Link:**

[**https://colab.research.google.com/drive/1JgWEWFyPMS2IbNL2d9r0MHUouUU1iySx?usp=sharing**](https://colab.research.google.com/drive/1JgWEWFyPMS2IbNL2d9r0MHUouUU1iySx?usp=sharing)