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.

a) 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.

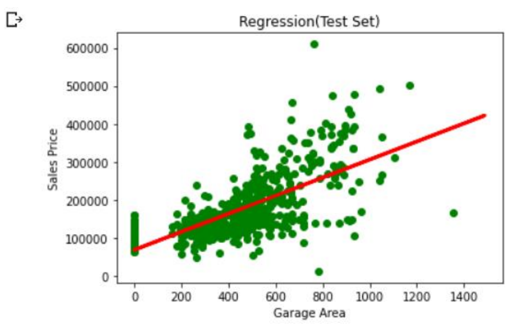
Train and Test r2:

```
Train: 0.4036140916008528
Test: 0.4352323177438491
```

Bias and Slope:

```
Regressor coeffient or slope: 237.58546612392706
Interception point with axis: 68779.89798972152
```

Scatter Plot:



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

Train and test r2 accordingly:

```
Train: 0.943037793827781
Test: -8.624046079362404e+18
```

c) **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.

d) **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:

```
Train: 0.9211536946303333
Test: 0.841271168982664
```

Optimal alpha:

```
Alpha: 5.0
```

e) **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.

f) 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:

```
 4.80568808e+04  4.11215228e+04  3.59411591e+03  3.95993125e+04
-2.16708190e+15 -5.85926112e+14 -8.96935385e+14  2.34600822e+15
-1.56581113e+15 -6.79143035e+14 -3.49931326e+14  1.74571004e+15
 2.21047166e+03 -5.04361678e+03  1.70746438e+04  4.99687074e+03
-3.58923590e+04 -5.17516200e+04  1.49878076e+04  3.40462934e+04
 1.08907499e+04  1.67972144e+04  2.64127044e+04  1.08575386e+04
-7.66197968e+03 -5.07445071e+03 -1.14140527e+04  2.38318799e+04
 9.12120085e+03 -2.07558408e+04 -1.52824219e+03 -2.89031250e+03
-1.11960262e+15 -1.11960262e+15 -1.11960262e+15 -1.11960262e+15
-1.11960262e+15 -1.11960262e+15 -1.11960262e+15  7.14978461e+14
 7.14978461e+14 -6.12562500e+02 -2.80484375e+03  1.75493945e+15
 1.75493945e+15  1.75493945e+15  1.75493945e+15  1.93281398e+14
 1.93281398e+14  1.93281398e+14  1.93281398e+14 -4.70618917e+14
-4.70618917e+14 -4.70618917e+14 -2.09894509e+15 -2.09894509e+15
-2.09894509e+15 -2.09894509e+15 -2.09894509e+15 -1.92519606e+15
-1.92519606e+15 -1.92519606e+15 -1.74506977e+12 -1.74506976e+12
-1.74506977e+12 -1.74506978e+12 -1.74506979e+12 -1.74506978e+12
-1.74506977e+12 -1.74506979e+12 -1.74506978e+12 -1.74506977e+12
-1.74506967e+12 -1.74506978e+12 -1.74506978e+12 -1.74506977e+12
-1.74506979e+12 -1.74506979e+12 -1.74506975e+12 -1.74506979e+12
-1.74506975e+12 -1.74506976e+12 -1.74506979e+12 -1.74506978e+12
```

Below is a partial image of ridge regression coefficients:

```
 3.05615859e+04  1.75659684e+04  4.93787974e+03  4.16958358e+04
 4.44751522e+04  1.11670034e+04  3.85290916e+03  4.53451722e+04
 6.94857589e+04  4.92754654e+04  5.11296997e+03  8.25199198e+04
 1.88330356e+04  2.74594773e+03  2.70704980e+04  1.00103564e+04
-6.87684290e+03 -1.11986756e+04  2.71132357e+04  3.43808528e+04
 2.93306543e+03  3.02124218e+04  2.16554029e+04  1.38166376e+04
 2.17693238e+03  5.21428789e+03  2.05788887e+03  2.38469165e+04
 4.64894464e+03  3.37674364e+03 -2.70177469e+03 -4.22464990e+03
-2.74176134e+03 -6.17726657e+03  3.00246759e+03  2.10882728e+03
 3.78362835e+03  2.14302179e+03 -2.11891710e+03 -3.96240197e+03
 3.96240197e+03 -4.74760984e+02 -9.33829604e+02  2.01049257e+03
 8.69345341e+03 -1.16151924e+04  9.11246444e+02 -6.35354146e+03
 6.09403042e+03  5.62294689e+02 -3.02783655e+02  5.73874039e+03
-3.73370108e+03 -2.00503931e+03  9.30400381e+01  6.46626747e+03
-3.18817549e+03 -4.41182244e+03  1.04069042e+03 -3.12590312e+02
 2.77317659e+03 -2.46058628e+03 -5.37933017e+03  2.59292022e+03
 3.17546524e+03 -3.53662164e+03 -4.33426061e+03 -9.11769969e+03
 9.13271524e+03 -1.33849262e+04 -1.17402374e+04 -2.73621149e+02
 2.37741580e+04 -8.03621921e+03  0.00000000e+00 -6.47099547e+03
[illegible]
```

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.

a) **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
System Output	Orange	1	0	0
	Apple	2	2	1
	Lemon	2	0	1

General Accuracy: 4/9

Confusion matrix for 3 different classes:

Gold Labels

		TO	TN
System Output	SO	1	0
	SN	4	4

	TA	TN
SA	2	3
SN	0	4

	TL	TN
SL	1	2
SN	1	5

TO: True Orange
TN: True Negative
SO: System Orange
SN: System Not

TA: True Apple
SA: System Apple
TL: True Lemon
SL: System Lemon

Contingency table for micro-average:

	True Positive	True Negative
System Positive	4	5
System Negative	5	13

Micro-average:

Accuracy: 17/27
Precision: 4/9
Recall: 4/9
Error: 10/27
F1 score: 4/9

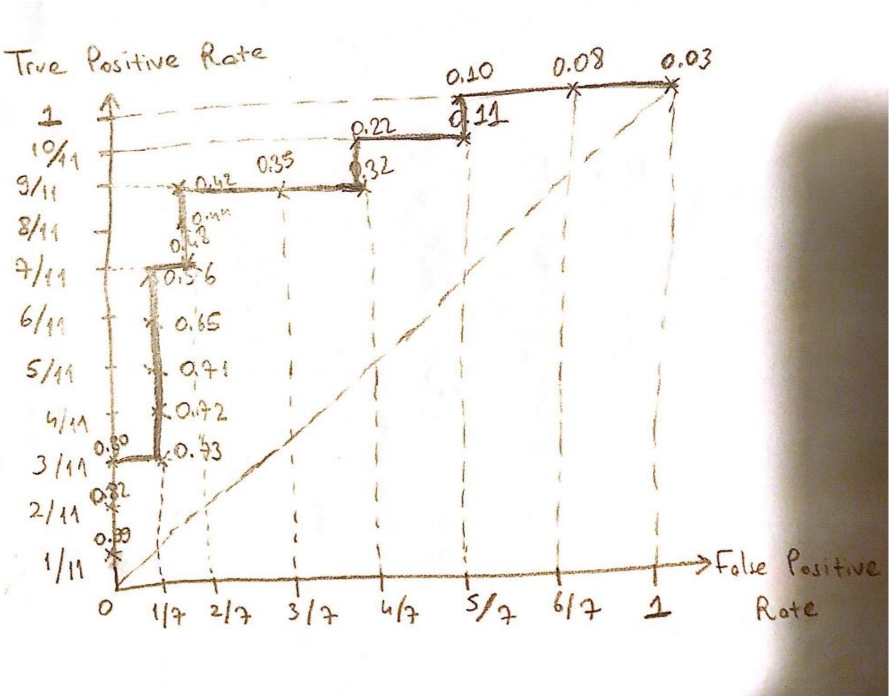
Macro-average:

Accuracy: 17/27
Precision: 26/45
Recall: 17/30
Error: 10/27
F1 score: 2652/4635 = 0.57

b) 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.

id	score	active/decoy	id	score	active/decoy
O	0.03	a	L	0.48	a
J	0.08	a	K	0.56	d
D	0.10	d	P	0.65	d
A	0.11	a	Q	0.71	d
I	0.22	d	C	0.72	d
G	0.32	a	N	0.73	a
B	0.35	a	H	0.80	d
M	0.42	d	R	0.82	d
F	0.44	d	E	0.99	d

ROC CURVE



Colab Link:

https://colab.research.google.com/drive/1JgWEWFyPMS2IbNL2d9r0MHUouUU1iySx?usp=sharing