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

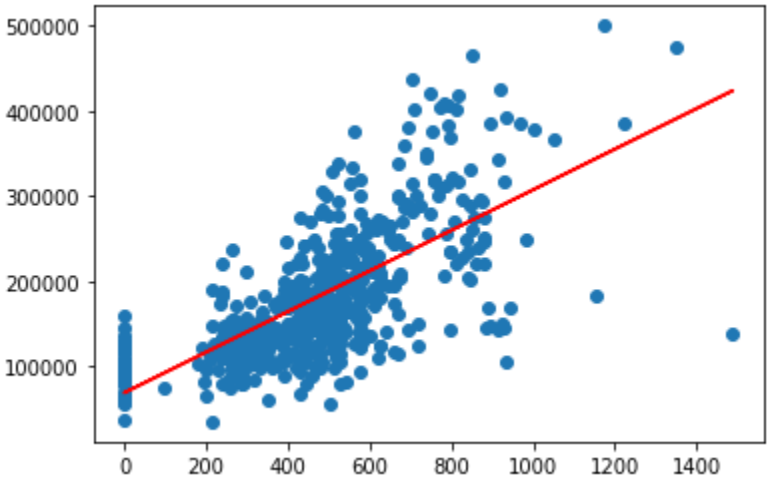
https://raw.githubusercontent.com/OpenClassrooms-Student-Center/Evaluate-Improve-Models/master/house_prices.csv

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.**

```
Regressor coeffient or slope: [[238.55255257]]
Interception point with axis: [68803.5794934]

Train: 0.4008804539088502
Test: 0.44940274425903814
```



Target and features are Linearly correlated.

b) **Apply multiple linear regression by taking all input features. Print the train and test R2. Plot the test set with a scatter plot and add the linear regression model line.**

```
Train: 0.9408590817831344
Test: -5.787705863267773e+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 model shows low R2 due to the fact that the data points fall further away from the regression line.There is noise in the data and high variability.R2 shows us the scattering around the regression line.

In part b model shows high R2 due to the fact that it is overfitting the data since it can be observed that it shows low R2 score over test data.The model uses scaling in order to deal with the scattering problem but it is still overfitting on data.We have checked the coefficients and see that they are very high.High coefficients occur when there is overfitting in the data.

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

```
Alpha: 5.0
Train: 0.9178609787326371
Test: 0.8612994065278898
```

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

Ridge Regression is a form of regression that discourages learning procedure to a more complex or flexible model causing the model to avoid the risk of overfitting.The fitting procedure uses a loss function(RSS:Residual Sum of Squares) where this function is tried to be minimized.Ridge Regression is when dealing with complex models and trying to avoid overfitting since the model can show low performance on unseen data even though it shows high performance on train dataset like in the part b case .This model contains a meta-parameter for regularization called "Alpha" or "L2 norm".Cross Validation is used in order to calculate the optimal "Alpha" value.Ridge Regression also needs scaling in order to perform.

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?**

The coefficients are reduced regarding multiple linear regression in ridge regression.Ridge Regression allowed the model to be less overfitting causing the model to be more normalised.

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 (true)	Prediction (system)

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

Orange	**True Orange**	**True Not**
System Orange	**1**	**0**
System Not	**4**	**4**

Apple	**True Apple**	**True Not**
System Apple	**2**	**3**
System Not	**0**	**4**

Lemon	**True Lemon**	**True Not**
System Lemon	**1**	**2**
System Not	**1**	**5**

Pooled	True Yes	True No
System Yes	4	5
System No	5	13

Macro Averaging:

Orange-Recall =1/5 , Apple-Recall =1 , Lemon-Recall = 1/2
Orange-Precision = 1, Apple-Precision =2/3 , Lemon-Precision =1/3
Orange-Accuracy =5/9 , Apple-Accuracy =6/9 , Lemon-Accuracy =6/9
Orange-Error =4/9 , Apple-Error =3/9 , Lemon-Error = 3/9
Orange-F1 = 0,3333, Apple-F1=0,5714, Lemon-F1 = 0,4000

Recall = 0,56666,
Precision = 0,66666,
Accuracy = 17/27,
Error = 10/27,
F1-Score = 0,4349

Micro Averaging:

Recall = 4/9,
Precision = 4/9,
Accuracy = 17/27,
Error = 10/27,
F1-Score = 0,4444

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

d 0.99
d 0.82
d .80
d .72
d .71
d .65
d .56
d .44
d .42
d .10
~~d 0.73~~
d 0.48
d 0.35
d 0.32
d 0.11
d .08
d .03

Hit = # of (T) called ~~ob~~ true

$\frac{3}{10}$

False Alarm = # called / # ~~la~~ tres