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.**

IT IS ON THE NOTEBOOK.

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

IT IS ON THE NOTEBOOK.

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, the training R2 and test R2 is really low because garage area is not the only factor which affects the sales price and they are not that much related. In part b, we look all the features which affects the sales price so training R2 score is higher. However, the model overfits the training data therefore test R2 is low and we need to apply ridge regression to resolve this issue.

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

IT IS ON THE NOTEBOOK.

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

Ridge Regression is a method that we use to ensure that we don't overfit our training data by adding a degree of bias to the regression estimates. It makes coefficients of the features less sensitive to the data by making the coefficients smaller. We use Ridge Regression when we use complex models and avoid overfitting of these complex models.

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?**

Coefficients are printed in notebook.

Ridge regression made the absolute value of most of the feature weights smaller. (Especially the features with extremely huge weights)

Therefore, the effects of the attirbutes on the model have decreased.

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

	Orange	Lemon	Apple	Precision
Orange	**1**	**0**	**0**	**1/1 = 1**
Lemon	**2**	**1**	**0**	**1/3 = 0.33**
Apple	**2**	**1**	**2**	**2/5 = 0.4**
Recall	**1/5 = 0.2**	**1/2 = 0.5**	**2/2 = 1**	

Macroavg precision = (1+0.33+0.4)/3 = 0.576 Macroavg recall = (0.2+0.5+1)/3 = 0.56
Macroavg F1-score = 0.568
Accuracy = 4/9 = 0.44

CONTINGENCY TABLE

Class Orange	TO	TN	Class Lemon	TL	TN	Class Apple	TA	TN
SO	1	0	SL	1	2	SA	2	3
SN	4	4	SN	1	5	SN	0	4

POOLED

	True YES	True NO
System YES	4	5
System NO	5	13

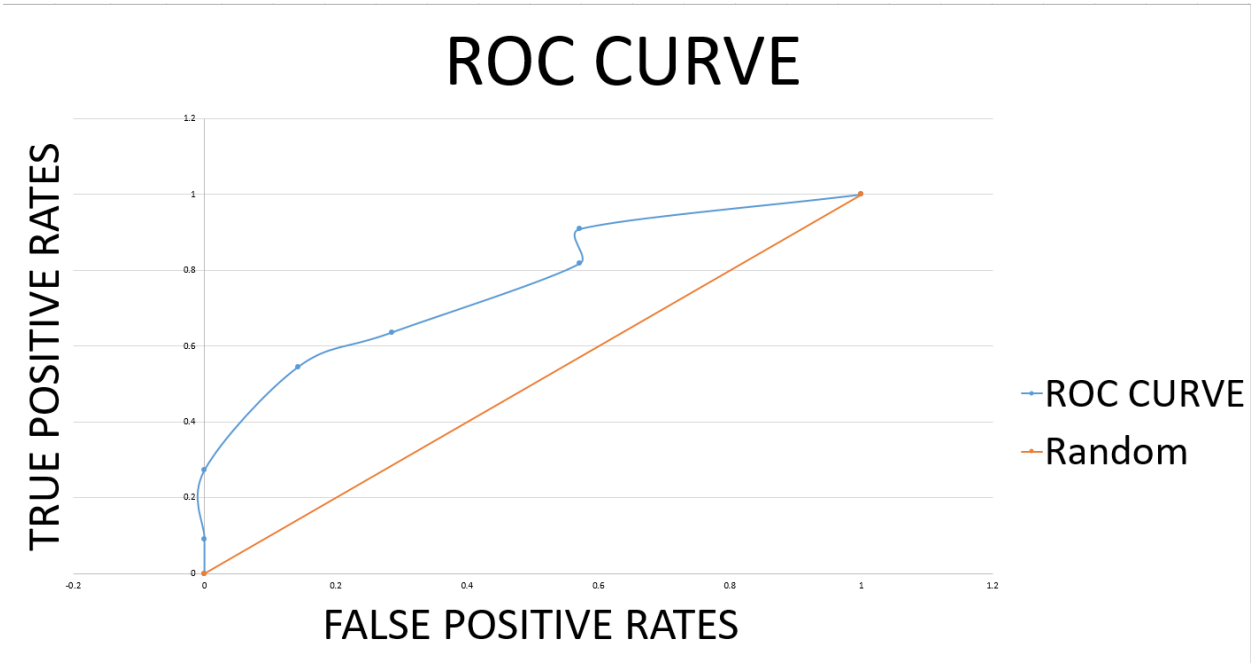
Microaverage Precision = 4/9 = 0.44 Microaverage Recall = 4/9 = 0.44
Microaverage F1-Score = 0.44

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

Graph will be in the next page.

In my case, Decoy will be positive, Active will be negative.



If threshold = 0, TPR=11/11 FPR=7/7
If threshold = 0.15 TPR=10/11 FPR=4/7
If threshold = 0.30 TPR=9/11 FPR=4/7
If threshold = 0.45 TPR=7/11 FPR=2/7
If threshold = 0.60 TPR=6/11 FPR=1/7
If threshold = 0.75 TPR=3/11 FPR=0
If threshold = 0.9 TPR=1/11 FPR=0
If threshold = 1 TPR=0 FPR=0