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**~~ **Notebook Solution**
https://colab.research.google.com/drive/1UQvLhd-tdXAvoCzoYdhfaXSB1gpYIDQE?usp=sharing
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) **15pt - 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.**
   Y = 240.14x + 67120.46
   There is a positive correlation between them.
   Others in the notebook.

b) **15pt - Apply multiple linear regression by taking all input features. Print the train and test R2.**
   In the notebook.

c) **10 pt - 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?**
   R2 is low in part a, because we used the limited data just one feature. So underfitting occurs here. We need to have more information to estimate better.
   In part b, the model is over-fitted. Since we have lots of features, many of them are not correlated with the price and they cause the model to be over-fitted. When we test the model, it produces values that are different from true values and since the values are in thousands, we encounter a quite big nominator for some cases.

d) **15pt - Apply ridge regression with cross-validation by taking all input features. Print optimal alpha. Print also the train and test R2.**

   In the notebook.

e) **10pt - Discuss on regularization. What is ridge regression? When do we use it? And what is the effect on features?**
Ridge regression belongs to a class of regression tools that use L2 regularization. The other type of regularization, L1 regularization, limits the size of the coefficients by adding an L1 penalty equal to the absolute value of the magnitude of coefficients. This sometimes results in the elimination of some coefficients altogether, which can yield sparse models. L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. All coefficients are shrunk by the same factor (so none are eliminated). Unlike L1 regularization, L2 will not result in sparse models.

A tuning parameter (λ) controls the strength of the penalty term. When λ = 0, ridge regression equals least squares regression. If λ = ∞, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

f) **10pt - 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?**
Ridge regression forces coefficients to be closer to 0 so that we decrease the complexity of the model and prevent overfitting.

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

	Actual labels			
Predicted	orange	lemon	apple	precision
orange	1	0	0	1
lemon	2	1	0	0.33333333
apple	2	1	2	0.4
recall	0.2	0.5	1	

Accuracy	0.444
Error	0.556

	Precision	Recall	F1-Score
orange	1	0.2	0.571
lemon	0.333	0.5	0.4
apple	0.4	1	0.333
Macroaverage	0.578	0.567	0.435

Microaverage: 0.44 for all.

-3pts: if macro or micro are missing or incorrect.

-2pts: if one of the metrics is missing or incorrect.

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

Decay: positive class
Active: negative class
If you did the opposite, It is OK.
0 pts: if written with code. I need to be sure you don't draw it with code. For that reason, I need to see the numbers at the discrete points.