Assignment 1

Statistical Learning – Fall 2018

Assignment Date: 1397/08/01

Due Date: Section 1 and 2) 1397/08/12, Section 3) 1397/08/19

Note1: Please submit your thoughts on the following exercises. Don't forget to include your codes.

Note2: For the lab sections, you can use any software which you are familiar with (Python, Matlab or R).

Section 1

1) What are the advantages and disadvantages of a very flexible (versus a nonflexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

2) What is the meaning of "confidence interval" for a given parameter? Suppose that a given confidence interval includes "zero" value. What we can conclude about the *null hypothesis?*

3) Suppose that we have n training samples in the form of $D = (x^1, y^1), \dots, (x^n, y^n)$ which, each of the inputs x has n dimensions. We want to fit a linear regression model with MSE cost function which has been defined below:

$$J(w) = \Sigma_{i=1}^{n} \left(y^i - w^T x^i\right)^2$$

As you know, the optimum value of the w which optimizes the above objective function is:

$$\widehat{w} = (XX^T)^{-1}XY^T$$

Which $X \in \mathbb{R}^{d\times n}$, $Y \in \mathbb{R}^{1\times n}$

Now, if we use a probabilistic framework, the optimum value for the w will be:

$$\widehat{w} = argmin_w E_{x,y}[(y - w^T x)^2]$$

a) Find the optimum value of $\widehat{w}$ in terms of the autocorrelation matrix $R = E_x(xx^T)$ and the cross-correlation matrix $C = E_{x,y}(x, y)$.

b) Afterwards, show that we can write the expected error value as the summation of two terms of "structural error" and "estimation error":

$$E_{x,y}[(y - \widehat{w}x)^2] = E_{x,y}[(y - w^{*T}x)^2] + E_x[(w^{*T}x - \widehat{w}^T x)^2]$$

Explain the interpretation of each term.

c) Explain the relation between these two terms and the "non-reducible error" and "reducible error"? any connection between them.

4) Consider a $M - order$ polynomial linear regression model on a given data with N training samples. We have derived the values of the coefficients by optimizing the given objective function:

$$E(w) = \frac{1}{2}\Sigma_{n=1}^{N}(y(x_n, w) - t_n)^2$$

x is the input value which has been normalized between 0, 1 and t is the corresponding output value. Values of the coefficients w^* which obtained from polynomials of various order, have been listed in Table 4.1.

	$M=0$	$M=1$	$M=4$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

I. What is your inference about the obtained values of w^*?

II. What is your expectation about the values of training error and testing error, when we arise the values of M from $M = 0$ to $M = 9$?

III. What methods do you suggest to improve the regression algorithm?

IV. What is the influence of changing N value on testing and training errors?

5) Consider the fitted values that result from performing linear regression without an intercept. In this setting, the i^{th} fitted value takes the form:

$$\hat{y_i} = x_i\hat{\beta}$$

Where

$$\hat{\beta} = \frac{\Sigma_{i=1}^{n} x_i y_i}{\Sigma_{i'=1}^{n} x_i}$$

Show that we can write

$$\hat{y_i} = \Sigma_{i'=1}^{n} a_{i'} y_{i'}$$

What is $a_{i'}$?

6) It is claimed that in the case of linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y. Prove that this is the case. For simplicity you may assume that $\bar{x} = \bar{y} = 0$.

Section 2

Implement the following instructions (in R, Matlab or Python) and analyze the results if it's necessary.

A) Load the "iris" dataset. (If you're using R, you can load the dataset from ISLR package and if you're not, you can download it from the link: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/).
B) Explain the structure of the data and show the list of the features.
C) Show the scatterplot of the data. Then, explain how these features are related to each other(for example Petal.Width-Sepal.Length, …)
D) Fit a linear model to Petal.Length as a function of Petal.Width. Report the t-value and the p-value of this model.
E) Create the feature V by Multiplying the Petal.Width and Sepal.Width. Now, fit a linear model to Petal.Length as a function of V. Compare this model with the model of part (D) based on t-values and p-values. What is your conclusion about these models?

Section 3

Use of Stochastic gradient descent to estimate logistic regression parameters. Comes up in a separate file.