Machine Learning Fall, 2021
Sharif University of Technology CE 717
S. Abbas Hosseini

Date Due: 1400/09/05

Homework 3

Theoretical (50)

1. Part1: In building a linear regression model for a particular data set, you observe the coefficient of one of the features having a relatively high negative value. Which one is true and why. (2)

   (a) This feature has a strong effect on the model (should be retained)

   (b) This feature does not have a strong effect on the model (should be ignored)

   (c) It is not possible to comment on the importance of this feature without additional information

   Part2: Discuss whether the following statements are true or false and explain your reasons:

   (a) If the bias is high, increasing the training data will not help reduce the bias. (1)

   (b) Reducing training error leads to reducing test error. (1)

   (c) Increasing model complexity in regression always reduces the training error and increases the test error. (1)

   (d) In a regression problem, When 6th degree polynomial regression results in a significant training error, linear regression should be used instead. (1)

2. For part "a", "b", and "c" assume multiple linear regression model which is $Y = X\beta + \epsilon$ where X is the matrix of data, Y is a vector of response values, and ϵ is a vectorize Gaussian noise($\epsilon \sim N(0, \Sigma)$). Consider least square as cost function.

   (a) Show that $\omega^*_{opt} = (X'X)^{-1}X'Y$. (5)

   (b) what would be the closed form solution for ω if we add L2 regularization term to SSE function? (5)

   (c) In the same problem when $E[\epsilon] = 0$ and $Var(\epsilon) = \sum$ one of the students propose another estimator $\omega^*_{new} = (X' \sum^{-1} X)^{-1} X' \sum^{-1} Y$. Prove that this estimator is equal to ordinary least-square estimator (ω^*_{opt}), if and only if there exist a nonsingular matrix F, such that:

   $$\sum X = XF$$

   (4)

   (d) consider the following loss function:

   $$L(\omega, \lambda_1, \lambda_2) = |y - X\omega|^2 + \lambda_1 ||\omega||_2^2 + \lambda_2 ||\omega||_1$$

   Show this loss function is equivalent to L1 regularized loss function by adding some data points. (5)

3. In mathematical statistics, the Kullback–Leibler divergence, (also called relative entropy), is a measure of how one probability distribution is different from a second. Consider two probability distribution P and Q then the $KL(P||Q)$ calculates the amount of difference between P and Q with following formula:

$$KL(P||Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

Suppose that p(x) is some fixed distribution and that we wish to approximate it using a Gaussian distribution $q(x) = N(x|\mu, \Sigma)$. By writing down the form of the KL divergence $KL(p||q)$ for a Gaussian q(x) and then differentiating, show that minimization of $KL(p||q)$ with respect to μ and Σ leads to the result that μ is given by the expectation of x under p(x) ($\mu = E_p[x] = \int xP(x)dx$) and that Σ is given by the covariance ($\Sigma = Var_p(x)$). (10)

4. Given n training data with m features, let the target value vector be $y = [y^{(0)}, \ldots, y^{(n)}] \in \mathbb{R}^n$ and data samples be $X = [x^{(0)}; \ldots; x^{(n)}] \in \mathbb{R}^{n \times m}$. In this context, x_j denotes the j^{th} column of this matrix.

   (a) Show that if we train the regressor on just one of the features (from m features), we then have $w_j = \frac{x_j^T y}{x_j^T x_j}$. (5)

   (b) Suppose that the columns of matrix X are orthogonal. Prove that the optimal parameters from training the regressor on all features is the same as the optimal parameters resulting from training on each feature independently. (5)

   (c) Now, suppose we want to train a regressor on the bias term and one feature of the data samples ($w = [w_j, w_0]$). Show that we will have:

$$w_j = \frac{cov[x_j, y]}{var[x_j]}$$
$$w_0 = \mathbb{E}[y] - w_j \mathbb{E}[x_j]$$

   (5)

Practical (50)

In this part, we are going to train a linear regressor with the help of various basis functions. In this homework, you are meant to use the closed form of the optimal parameters (with Least Square Error loss function) that you learnt in the class. You are not going to find the parameters iteratively.

Dataset: You are asked to work on the Boston house prices dataset. This dataset consists of 506 data samples and 13 real atributes. The target value is the Median value of owner-occupied homes in \$1000's (*'MEDV'* feature).

Allowed packages: Pandas, matplotlib, and numpy. Sklearn is allowed only for getting the dataset.

Assignment: Hand in your report in pdf and your codes in Python. (You may also use Jupyter Notebooks instead.)

1. First of all, we recommend to check whether if the dataset includes missing parts. Then split the dataset into train set (80% of the data) and test set (20% of the data).
   Note: Do NOT use the test set unless for loss computation.

2. Using the tools that you learnt before, try to play with the dataset. You may want to plot the target value based on 13 different features and recognize the correlation between features and the target values. Put your plots in the report and talk about them and their meanings.

3. Now, using the closed form of the linear regression parameters, find the optimal weight parameters. Plot the target value and the predicted value based on *'LSTAT'*, *'DIS'*, and any other features so that you can see how the distributions vary. Put the plots in your report.

4. In this part, you add the 2^{nd}-order of each feature to the original feature vectors. Again, find the optimal parameters in this manner and plot the target and predicted values, same as before.

5. Now, we want to use Gaussian basis functions along with the original features.

$$\phi_j(x) = exp\{\frac{||x - \mu_j||_2^2}{2s^2}\}$$

   Here, we use 10 basis functions with the spatial scale $s = 1$. You may randomly select different μ_js from the train set. Again, find the optimal parameters with these new features and plot the target and predicted values, as before.

6. Report the train and test MSE loss and plots for each of the three above-mentioned parts (Overall, you have to report 6 loss values and 6 figures for these three approaches!). Afterward, discuss on the results in a paragraph. Which feature approach works better in this dataset?

Good Luck ;)