Due Date: 1399.08.13

Homework 3

Theoretical

1. Given n data samples $x_i \sim p(x)$, suppose we have a constant model. We know that *Mean Square Error* and *Mean Absolute Error* are defined as follows:

$$L_{MSE}(\theta) = \frac{1}{n}\sum_i (x_i - \theta)^2$$

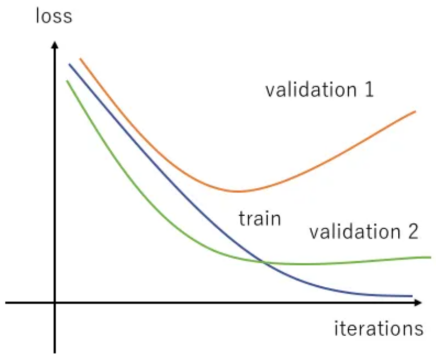
$$L_{MAE}(\theta) = \frac{1}{n}\sum_i |x_i - \theta|$$

   (a) By minimizing both of these loss functions, find the optimal parameter θ^* for each function.

   (b) Based on the previous part, briefly discuss the advantages and the drawbacks of *MSE* and *MAE* in a paragraph or two.

   (c) Read about *Huber* and *Log-Cosh* loss functions and state how using these functions may solve your mentioned drawbacks.

2. Suppose we have two validaion sets in the procedure of training and the loss curves look like this:



   Interpret these curves and the problem and propose your solutions in the following manners:

   - Two validation sets come from different distributions.
   - Two validation sets come from the same distributions.

3. Consider the expected loss for regression problems under the L_q loss function as follows:

$$\mathbb{E}[L_q] = \iint |f(x) - y|^q p(x, y) dx dy$$

   Write down the condition that $f(x)$ must satisfy in order to minimize the expected L_q loss. What does the solution represent when $q = 1$ and when $q \to 0$?

4. Rewrite the *tanh* function based on the *sigmoid* function. Then suppose we use a model f for predicting the output of input x as follows. g here can be *tanh*, *sigmoid*, or any other function.

$$f(x; \mathbf{w}^{(g)}) = w_0^{(g)} + \sum_{j=1}^{m} w_j^{(g)} g(x)$$

Find expressions to relate $[w_0^{(tanh)}, \ldots, w_m^{(tanh)}]$ to parameters $[w_0^{(sigmoid)}, \ldots, w_m^{(sigmoid)}]$.

5. Assume we have a linear model with *Sum Square Error* function. Show that if we independently add a noise ϵ_i to each dimension of data points x_i which $\epsilon \sim \mathcal{N}(0, \sigma^2)$, minimizing expected loss on the noisy distribution is equivalent to minimizing the expected *Sum Square Error* for noise-free data samples with the addition of a regularization term.

6. The weighted linear regression is an extension of linear regression in which we allocate a weight to each data sample. We can write the weighted loss for n training data samples as follows. Find the closed form of the optimal $\mathbf{w}$.

$$L(\mathbf{w}) = \sum_{i=1}^{n} F_i (y^{(i)} - \mathbf{w}^T x)^2$$

7. 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 $w_j = \frac{x_j^T y}{x_j^T x_j}$.

   (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.

   (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:

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

Practical

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.

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 ;)