Machine Learning 2023 spring

HW1: Linear Regression

Deadline: 112.3.24

Part I. Linear Regression

In this exercise, you will learn how to use cross-validation to select model parameters for curve fitting. Please write a **Python** program to implement linear regression. You are given a *HW1.csv* file, which contains two arrays:

x: $\{x_1, x_2, \ldots, x_{70} \mid 0 \le x_i \le 3\}$ represent the input values and

t: $\{t_1, t_2, \ldots, t_{70}\}$ represent the target values.

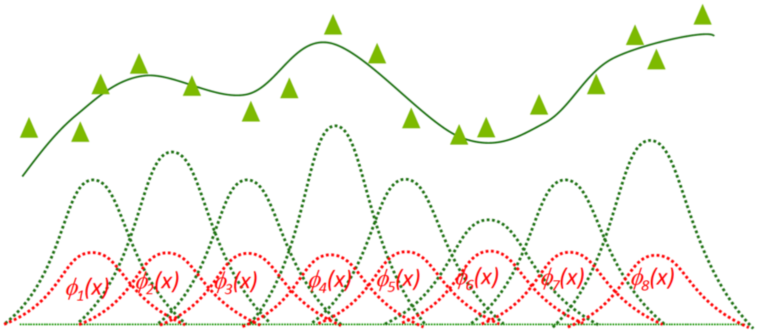
The number of data points is 70. You need to split them into the Training Set (the first 50 points) and the Testing Set (the last 20 points).

Training stage:

Please fit the data by minimizing the error function:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y(x_i, \mathbf{w}) - t_i\}^2 \text{ , where } y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M} w_j \phi_j(x)$$

$$y = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_{M-1}\phi_{M-1}(\mathbf{x})$$



Basis Function:

Please use the following set of basis functions $\phi = (\phi_0, \ldots, \phi_{M-1})$

$$\phi_j(x) = \begin{cases} 1 & j = 0 \\ \sigma\left(\dfrac{x - \mu_j}{s}\right) & j = 1, \ldots, M-1 \end{cases}$$

where $\sigma(a)$ is the logistic sigmoid function defined as $\sigma(a) = \dfrac{1}{1 + \exp(-a)}$.

Please take the following parameter settings for the basis functions:

$$s = 0.1 \text{, and } \mu_j = \frac{3j}{M} \text{ with } j = 1, \ldots, (M\text{-}1).$$

1. Please plot the data points(only the Training Set) and the fitting curve for *M*=1,3,5,10,20 and 30, respectively.

2. Please plot the Mean Square Error evaluated on the Training Set and the Testing Set separately for *M* from 1 to 30.
3. Please apply the 5-fold cross-validation in your training stage to select the best order *M* and then evaluate the mean square error on the Testing Set. Plot the fitting curve and data points (only the Training Set). You should briefly express how you select the best order M step-by-step.
4. Considering regularization, please use the modified error function

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y(x_i,\mathbf{w}) - t_i\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

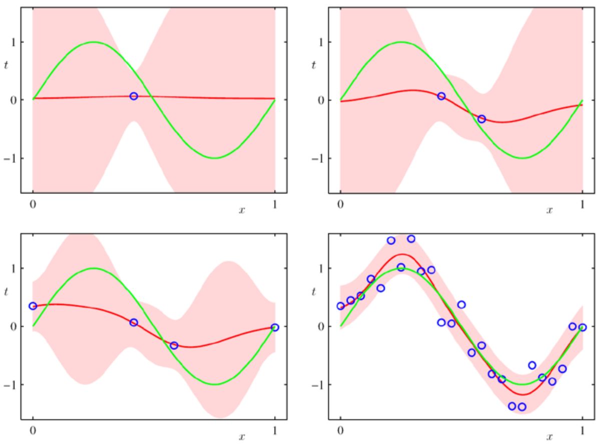
where $\|\mathbf{w}\|^2 \equiv w_1^2 + w_2^2 + \ldots + w_M^2$. Repeat *Part I -1. and Part I-2.* with $\lambda = \frac{1}{10}$.

(You can also try to change the value of λ and discuss what happens under different λ values.)

Part II. Bayesian Linear Regression

In this part, use the Training Set in Part I, apply the sigmoidal basis functions in Part I with *M*=10, and implement Bayesian linear regression. In order to discuss how the amount of training data affects the regression process, please implement a sequential estimation:

Please ***sequentially*** compute the mean m_N and the covariance matrix S_N for the posterior distribution $p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ with the given prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$ with $m_0 = 0, S_0^{-1} = 10^{-6}\mathbf{I}$. The predictive distribution $p(t \mid X, w, \beta)$ is chosen to be $\beta = 1$. Similar to the following figures, please plot the curve of the posterior mean m_N versus x and the region spanning one standard deviation on either side of the mean curve for N = 1, 2, 3, 4, 5, 10, 20, 30, 40, 50.



Homework Rules and Grading Policy

Homework will be graded by:

1. The correctness of your fitting lines.
2. Your discussion of what you observe in this regression problems.

Upload:

[web]	E3
[File Name]	hw4_StudentID.zip (ex: hw4_1234567.zip)

Remind:

1. Your report in the format of .pdf.
2. Deadline:

   If you have a late submission by 1 to 7 days, you will only get 70% of the score. We DO NOT accept any late submission after 7 days after the deadline.