Machine Learning 2024 spring

HW1: Linear Regression

Deadline: 2024.3.27

1 Linear Regression

In this exercise, you will learn how to use cross-validation to select model parameters for curve fitting. Please write a **Python** or **C++** program to implement linear regression. You are given a HW1.csv file, which contains 12 arrays:

- $\mathbf{x}_k : \{x_{k,1}, x_{k,2}, \ldots, x_{k,10000}\}$, $\forall k = 1, \ldots, K$: the input values of the kth input item where the number of data points is 10000, and $K = 11$.
- $\mathbf{x}_1$: song_duration_ms, $\mathbf{x}_2$: acousticness, $\mathbf{x}_3$: danceability, $\mathbf{x}_4$: energy, $\mathbf{x}_5$: instrumentalness, $\mathbf{x}_6$: key, $\mathbf{x}_7$: liveness, $\mathbf{x}_8$: loudness, $\mathbf{x}_9$: speechiness, $\mathbf{x}_{10}$: tempo, $\mathbf{x}_{11}$: audio_valence.
- $\mathbf{t} : \{t_1, t_2, \ldots, t_{10000}\}$: the target values (song_popularity)

The total number of data points is 15818. You need to split them into the training set (the first 10000 points) and the testing set (the rest 5818 data). You need to normalize each input feature of the training data.

$$\hat{x}_{k,n} = \frac{x_{k,n} - \nu_k}{\sigma_k}, \text{ for } n = 1, \ldots, 10000$$

- Mean: $\nu_k = \frac{1}{10000} \sum_{n=1}^{10000} x_{k,n}$
- Standard deviation: $\sigma_k = \sqrt{\frac{1}{10000-1} \sum_{n=1}^{10000} (x_{k,n} - \nu_k)^2}$

You also need to normalize each input feature of the testing data **by using the mean and standard deviation of the training data**.

- Training stage:
  Please fit the data by minimizing the error function:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \{y(x_{i,1}, \ldots, x_{i,K}, \mathbf{w}) - t_i\}^2, \text{ where } y(x_1, \ldots, x_K, \mathbf{w}) = \sum_{k=1}^{K} \sum_{j=0}^{M} w_{k,j} \phi_{k,j}(x_k)$$

$$y = w_{1,0} + w_{1,1}\phi_{1,1}(x_1) + w_{1,2}\phi_{1,2}(x_1) + \ldots + w_{1,M-1}\phi_{1,M-1}(x_1) + \ldots$$
$$w_{K,0} + w_{K,1}\phi_{K,1}(x_K) + w_{K,2}\phi_{K,2}(x_K) + \ldots + w_{K,M-1}\phi_{K,M-1}(x_K)$$

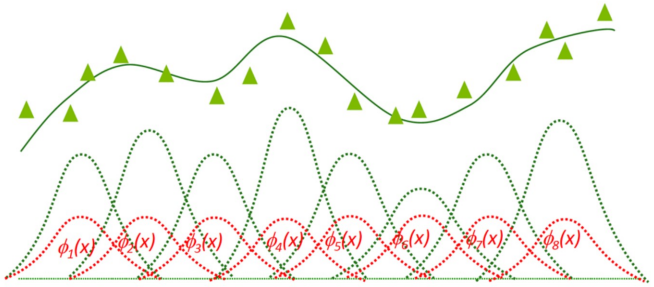
- Basis Function:
  Please use the following set of basis functions $\phi_k = (\phi_{k,0}, \ldots, \phi_{k,M-1})$

$$\phi_{k,j}(x_k) = \begin{cases} 1 & j = 0 \\ \sigma\left(\frac{x_k - \mu_j}{s}\right) & j = 1, \ldots, M-1 \end{cases}$$

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

  Please take the following parameter settings for the basis functions:

$$s = 0.1, \text{ and } \mu_j = \frac{3(-M + 1 + 2(j-1)\frac{M-1}{M-2})}{M} \text{ with } j = 1, \ldots, (M-1), \text{ for } M > 2.$$



1. Please plot the fitting curve of the third input feature ($\mathbf{x}_3$: daceability) for $M = 5, 10, 15, 20, 25, 30$, respectively. (Change M for all input features, but you only need to plot the fitting curve of the third input feature.)
2. Please plot the Mean Square Error evaluated on the training set and the testing set separately for $M = 5, 10, 15, 20, 25, 30$. Also, evaluate the accuracy of the training set and the testing set given as follows:

$$accuracy = 1 - \frac{1}{N_d}\sum_{i=1}^{N_d}\left|\frac{y - t_i}{t_i}\right|$$

   - N_d: number of data points in a data set
   - Only for evaluating the accuracy, if the target value $t_i = 0$, replace the denominator as 1.

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 of the third input feature ($\mathbf{x}_3$: daceability). You should briefly express how you select the best order M step-by-step.
4. Considering regularization, please use the modified error function

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

   where $\|\mathbf{w}\|^2 = w_{1,1}^2 + \ldots + w_{1,M}^2 + \ldots + w_{K,1}^2 + \ldots + w_{K,M}^2$. 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.)

Homework Rules and Grading Policy

Homework will be graded by:

1. Report 50%
   - The correctness of your fitting lines.
   - Your discussion of what you observe in the regression problems. For example,
     - If you change the number of basis functions, M, what effect will have on training and testing?
     - Which features do you consider the most important? Why do you consider your selected features to be the most important?
     - And any other topics you would like to discuss.
2. Demo 50%
   - Writing a program using C++ can earn you 10 points more in the demo segment compared to using Python. In other words, if you write the program with correctness in C++, you can get 100 in the demo. However, if you write the program with correctness in Python, you can get only 90 in the demo. If the program is incorrect in either language, you will receive a score of 0.
3. The final score is determined by an equal weight of 50% for both the demo and report. In order to achieve an average score of around B+, the report will be graded around 60.

Upload:

- [Web] E3
- [File Name] hw1_StudentID.zip (ex: hw1_1234567.zip)
  The file should include your code and report.

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 submissions after 7 days after the deadline.
3. We encourage open discussion to ensure program correctness, but plagiarism is strictly prohibited. Violators will receive a score of 0.

Demonstration:

1. In the demo part, you need to train your program based on the training data of HW1.csv (the first 10000 points) and test your program with the testing data (the last 5815 points) and the additional test data prepared by TAs, HW1_demo.csv.

2. We will use additional dataset HW1_demo.csv to test your program. You will be required to evaluate the accuracy of this dataset.

3. **You need to demonstrate the training accuracy (the first 10000 points), the testing accuracy (the last 5818 points), and the accuracy of additional testing data HW1_demo.csv.**

4. Your submitted program should be executable directly (with the dataset and program placed in the same folder).