MCE5901 Homework Set 1

September 13, 2023

NOTICE: The homework is due on Sept. 24 (Sunday) 11:59pm. **Please draft your solution with steps and provide the codes that you use for the applied problems.** You will need to submit two files to the blackboard: a PDF file including the solutions and results for all problems, and a file containing the codes for the applied problems. You are allowed, and even encouraged, to discuss the homeworks with your classmates. However, you must **write up the solutions on your own**. Plagiarism and other anti-scholarly behavior will be dealt with severely.

Problem 4 is optional. You will receive 20 points bonus if you solve it and the maximum score of the assignment is 100 points.

Problem 1. Bias-variance tradeoff

Suppose that we have a training data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ and

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

where ϵ_i are independent and have the same normal distribution $\mathcal{N}(0, \sigma^2)$.

By means of some learning method, we have found a function $\hat{f}(\boldsymbol{x})$ to model the true but unknown function $f(\boldsymbol{x})$. We decompose the expected prediction error of $\hat{f}$ on new point $\boldsymbol{x}_0$ as follows:

$$\mathrm{E}\Big[\big(f(\boldsymbol{x}_0) + \epsilon - \hat{f}(\boldsymbol{x}_0)\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(\boldsymbol{x}_0)\big]^2 + \mathrm{Var}\big[\hat{f}(\boldsymbol{x}_0)\big] + \sigma^2, \tag{1}$$

where

$$\mathrm{Bias}\big[\hat{f}(\boldsymbol{x}_0)\big] = \mathrm{E}\big[\hat{f}(\boldsymbol{x}_0)\big] - f(\boldsymbol{x}_0) \tag{2}$$

and

$$\mathrm{Var}\big[\hat{f}(\boldsymbol{x}_0)\big] = \mathrm{E}\Big[\big(\hat{f}(\boldsymbol{x}_0) - \mathrm{E}\big[\hat{f}(\boldsymbol{x}_0)\big]\big)^2\Big]. \tag{3}$$

(a) Show how to derive the equation (1).

(b) Explain the Bias -Variance Trade-Off with aid of this decomposition. Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches.

Problem 2. K-nearest neighbour (KNN) regression

Continuing with Problem 1, now we choose our learning method as KNN regression. Suppose the data points in the training data set $\mathcal{D}$ are ordered such that

$$||\boldsymbol{x}_1 - \boldsymbol{x}_0||_2 \leq ||\boldsymbol{x}_2 - \boldsymbol{x}_0||_2 \leq \cdots \leq ||\boldsymbol{x}_n - \boldsymbol{x}_0||_2$$

where $||\boldsymbol{x}_i - \boldsymbol{x}_0||_2$ is the Euclidian distance between the two points. That is, for the prediction point $\boldsymbol{x}_0$, the K training observations that are closest to it are $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$.

(a) What would be the predicted output $\hat{f}_{knn}(\boldsymbol{x}_0)$ by KNN?

(b) Show by means of equations (1), (2) and (3) that

$$\mathrm{E}\Big[\big(f(\boldsymbol{x}_0) + \epsilon - \hat{f}_{knn}(\boldsymbol{x}_0)\big)^2\Big] = \sigma^2 + \left(f(\boldsymbol{x}) - \frac{1}{K}\sum_{i=1}^{K} f(\boldsymbol{x}_i)\right)^2 + \frac{\sigma^2}{K}.$$

(c) How does the choice of K influence the Bias -Variance Trade-Off of KNN regression in view of this decomposition?

Problem 3. Applied problem: linear regression with one variable

In this exercise, you will implement linear regression and get to see how it work on data. The file "ex1data1.txt" contains the dataset for our linear regression problem. The first column is the population of a city (in 10,000s) and the second column is the profit of a food truck in that city (in $10,000s). A negative value for profit indicates a loss. We are interested in predicting the profit based on the population. Using programming to answer the following questions.

(a) How many samples are there in this dataset? Are there any missing values?

(b) Use a scatter plot to visualize the data.

(c) Implement the batch gradient descent. Initialize the parameters to 0 and the learning rate to 0.01. To monitor the convergence of your gradient descent implementation, plot the cost function $J(\beta^{(t)})$ versus the iteration step $t = 0, 1, 2, \ldots$.

(d) What are the final estimated values for the parameters?

(e) Plot the regression line and the data in the same figure.

(f) Use your model to predict the profit when the population is 35000.

(g) To understand the cost function better, plot the cost over a two dimension grid of β_0 and β_1 values.

If you are not familiar with programming, please refer to the instructions in the following link:

https://github.com/TSunny98/CS229/blob/658d268d251721abc92368374e17664ab339d144/Exercise1/exercise1.ipynb

Problem 4. Ridge regression (optional)

Consider the ridge regression problem

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{4}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$. Let β^* denote the optimal solution to (4).

(a) Show that (4) is equivalent to the problem

$$\min_{\tilde{\beta}} \sum_{i=1}^{n} \left[y_i - \tilde{\beta}_0 - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\tilde{\beta}_j \right]^2 + \lambda \sum_{j=1}^{p} \tilde{\beta}_j^2 \tag{5}$$

where $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ and $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. Let $\tilde{\beta}^*$ denote the optimal solution to (5).

(b) Give the correspondence between $\tilde{\beta}^*$ and the original β^*.

(c) Show that $\tilde{\beta}_0^* = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Then the coefficients $\tilde{\beta}_1, \ldots, \tilde{\beta}_p$ can be estimated by a ridge regression without intercept, using centered inputs: $x_{ij} - \bar{x}_j$.