Solutions to Assignment

Rongfei Jin

February 27, 2025

1   Chapter 5 - Conceptual 3

(a) K-fold CV is implemented the following way:

     (a) Split the data into K folds.

     (b) For each fold i:

         i. Use the other $K - 1$ folds as training data.

         ii. Train the model on the training data.

         iii. Evaluate the model on the i-th fold.

     (c) Average the evaluation results over all K folds.

(b) Compare to the validation set approach, K-fold CV ensures that all data points are used for both training and validation, which can lead to a more reliable estimate of the model's performance. It also reduces the variance of the performance estimate by averaging over multiple folds.

     However, K-fold CV can be computationally more expensive, especially for large datasets or complex models, as it requires training the model K times.

(c) LOOCV is a special case of K-fold CV where K is equal to the number of data points. The K-fold CV where k is less than N is less computationally expensive, as it only requires training the model K times instead of N times. More over, since almost all points are used for training, the variance of error will be greater than K-fold CV, making it more prone to overfitting. However, if a dataset is small, LOOCV can provide a more accurate estimate of the model's performance, as it uses all but one data point for training.

2   Chapter 5 - Conceptual 4

(a) Repeated sample the original with replacement to create B dataset

(b) Train the model on each dataset

(c) Predict the Y with the particular X value

(d) Compute the standard deviation of the predictions