Øving 4

2022-04-29

Problem 1

Explain how k-fold cross validation is implemented

b) Specify what is done, and in particular how the results from each fold are aggregated.

We split the dataset into K sets, with equal amounts of samples n_k in each. One is chosen as the validation set and the rest is used for training. The combination that makes the best fit is chosen. We calculate the MSE for each fold, and the cross-validation error is given by

$$CV_k = \frac{1}{n} \sum_{j=1}^{k} n_j \cdot MSE_j$$

Where n_j is the size of the n´th fold. And the MSE is given by

$$MSE_j = \frac{1}{n_j} \sum_{i \in V_j} (y_i - \hat{y}_i)^2$$

c) Relate to one example from regression.

Finding the optimal number of neighbours k in KNN regression. Finding the optimal degree in polynomial regression.

d) Relate to one example in classification.

K in KNN classification. Test/train split in logistic regression.

Problem 2

What are the advantages and disadvantages of k-fold cross-validation relative to

a) The validation set approach

This method is essentially just doing a split into test/train. The advantage is that k-fold CV optimizes the problem, we can gind the split that gives us the most optimal model, we get a model with less variance and less bias. However, for large datasets this is much more computational expensive than just a random split.

b) Leave one out cross-validation (LOOCV)

The LOOCV is the k-fold where k = n, which is an extreme version of normal k-fold CV and has the maximum computational cost. This results in a reliable and unbiased estimate of model performance. The disadvantage of this approach holds for large datasets and computationally expensive models.

k-fold has less variance than LOOCV (the variance in model selection), because in LOOCV we are averaging from n fitted models that are trained on nearly the same data, therefore we have positively correlated data.

k-fold has more bias, as in LOOCV we use a larger data set to fit the model, which gived a less biased version of the test error.

c) What are recommended values for k and why?

Experimental simulations has shown that K = 5 or K = 10 is optimal.

Very large k will lead to the estimator of test error having high variance and low bias, and is also very computational expensive.

For small k the estimator will have larger bias but lower variance.

Problem 3

Problem 4

Problem 5

Problem 6