Assignment #3

Elements of Machine Learning

Saarland University – Winter Semester 2024/25

Rabin Adhikari
7072310
raad00002@stud.uni-saarland.de

Dhimitrios Duka
7059153
dhdu00001@stud.uni-saarland.de

1 Problem 1 (Generalization)

1.1 Validation Error for All Classification Methods

Question: Assume you are only given training points for a binary classification problem and a small validation set. Does it make sense to compute the validation error for all classification methods (Logistic Regression, LDA, QDA) and report minimal validation error over all methods to estimate the test error? Justify your answer.

Answer: No, fitting all the models for the same training and validation set would overfit the validation set. Since we only have one validation set, the model parameters would be tailored for that one and may not generalize well to the test error. This sampling bias induced by the specific selection of the validation set doesn't capture the variance of the metric for other combinations of those parameters, so it underestimates the actual test error.

1.2 Overfitting in Cross-Validation

Question: Is it possible that model selection using cross-validation overfits? If yes, describe with an example; if no, explain the reason why overfitting is impossible.

Answer: Yes, the model selection using cross-validation can overfit. The example scenario would be us fitting a large number of models to a given dataset. Selecting the best model from a large number of models would result in them picking up the noise in the data, leading to better performance due to random chance. Also, there may be a case where we have a small number of data, and in that case, the validation metrics may not be able to generalize to the test metrics. Finally, the last case would be when the testing dataset has a different distribution to the dataset used to construct train and validation sets in cross-validation, in that case performing best on the validation set wouldn't exactly be transferrable to the test set.

1.3 Bias in K-Fold Cross-Validation

Question: Why does K-fold CV result in a higher bias than LOOCV?

Answer: One could argue that LOOCV is a specific case of a K-fold CV with $K = N$. The K-fold CV with $K < N$ would result in a higher bias than LOOCV because it has seen fewer data points in the training set. For K-fold CV, the number of data points in the training and validation set is $\frac{K-1}{K}N$ and $\frac{N}{K}$, respectively. We can see that as K increases the number of data points in the training set increases and it would reach the maximum $(N - 1)$ when $K = N$, i.e., for LOOCV. So, this higher bias in K-fold with $K < N$ comes from the models seeing fewer data points than LOOCV.

Additionally, from a modeling perspective, decreasing the training set would result in the best models trained on that set being simpler ones. Since the simpler models generally have a higher bias, this becomes applicable to K-fold as well.