LOTI.05.046 PATTERN RECOGNITION

HOMEWORK 9 - CROSS VALIDATION

May 17, 2019

Quazi Saimoon Islam
Msc. Robotics and Computer Engineering

1 INTRODUCTION

Over-fitting is a real problem in mach learning. Learning the parameters of a prediction function and testing it on the same data is a methodological mistake. A model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This is where cross validation is a strong tool to avoid over-fitting the predictive model.

In this assignment, two types of cross validation is explored, Leave One Out validation and Cross Validation with K-Folds. Leave one out validation is quite self explanatory where 1 sample is taken out for validation using the rest for training, and this process is iterated over the size of the training sample. K-fold validation splits the training sample into the number of k-folds specified and uses the separated batches for validation. The following image is a good illustration of this process:

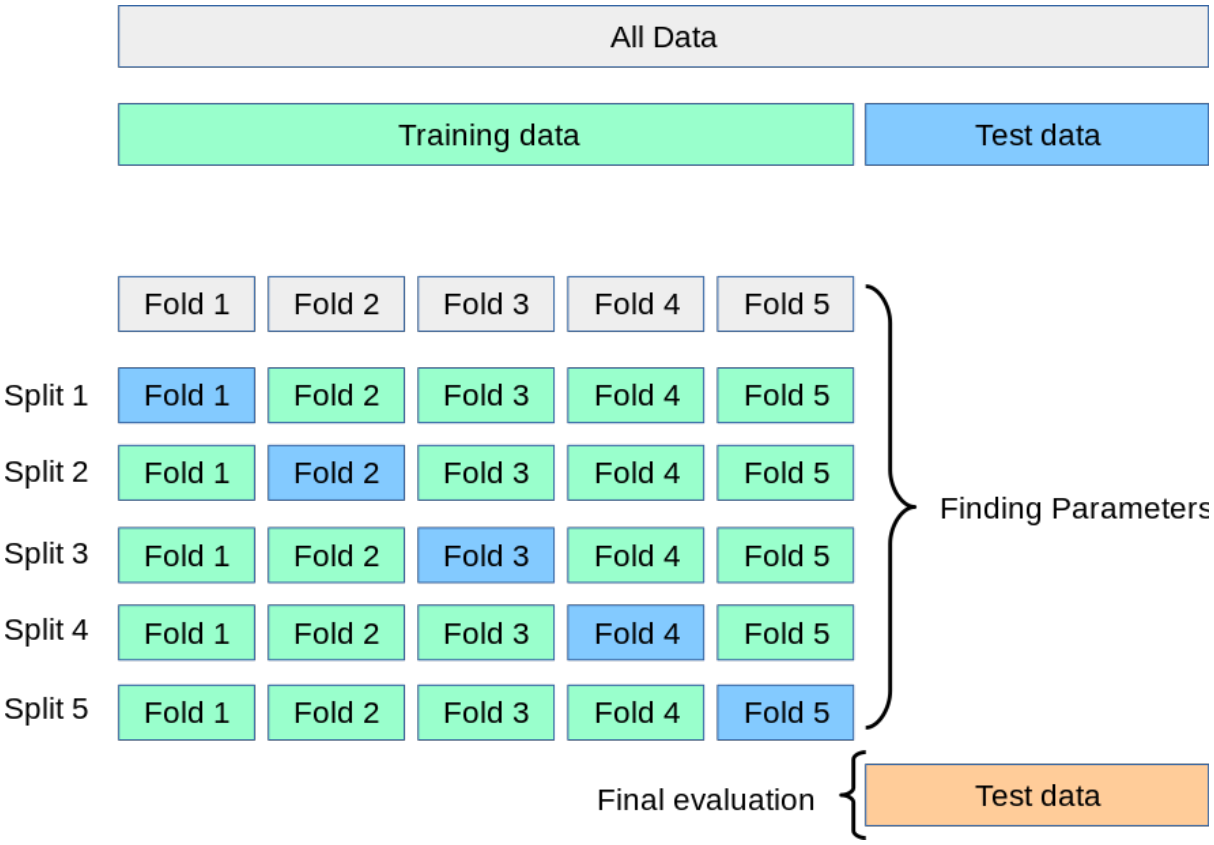

Figure 1: Illustration of how Cross Validation is structured

2 METHODOLOGY

The given methodology was followed and are as follows:

1. Read the data and separate features and labels. It should be noted that the data was not split into train and test sets and only cross validation performance is to be validated.
2. Perform Leave One Out validation on Decision Tree Classification
3. Perform Cross Validation with 10 folds on the Decision Tree Classification.
4. Perform Leave One Out Classification on Decision Tree Classification with PCA for feature reduction.
5. Perform Cross Validation with 10 folds on the Decision Tree Classification with PCA for feature Reduction.
6. Perform Leave One Out Classification on Decision Tree Classification with LDA for feature reduction.
7. Perform Cross Validation with 10 folds on the Decision Tree Classification with LDA for feature Reduction
8. Evaluate the accuracies and compare performance

3 RESULTS

3.1 Classification using decision tree

Using appropriate tools from the Sklearn library, it as easy to set up Leave One Out Validation and 10 fold cross validation for all three cases. The results of the performance of each ase are as below:

3.2 Leave One Out Validation

1. The Accuracy of the Decision Tree Classifier with Leave One Out Validation is : 92.000000%
2. The Accuracy of the Decision Tree Classifier + PCA with Leave One Out Validation is : 33.333333%

3. The Accuracy of the Decision Tree Classifier + LDA with Leave One Out Validation is : 93.333333%

As can be seen from the above, the most notable issue is that performace of the Decision Tree with PCA and this was down to the fact that PCA can't be properly performed with a sample size of 1, which is what is obtained when the test set is transformed. Therefore, it can be clearly said that Leave One Out is not a good validation technique for Classifiers that use PCA for feature reduction. LDA performed the best out of the three with this type of feature reduction

3.3 10-fold Cross Validation

1. The Accuracy of the Decision Tree Classifier with 10-fold Cross Validation is : 94.000000% (+/- 0.11)
2. The Accuracy of the Decision Tree Classifier + PCA with 10-fold Cross Validation is : 94.000000% (+/- 0.14)
3. The Accuracy of the Decision Tree Classifier with 10-fold Cross Validation is : 96.666667% (+/- 0.07)

Once again, Decision Tree Classifier with LDA performs the best during this validation.

4 DISCUSSION

Comparing the results of this assignment and Homework 8, we can see that cross validation provides a clearer view towards the performance of the classifier. In the previous assignment, the decision tree with LDA performed the best and the same is the case here, but the performance does not reach 100% as it did in the previous assignment.

REFERENCES

[1] Sci-Kit Learn Documentation *3.1. Cross-validation: evaluating estimator performance* [Online]. Available at: <https://scikit-learn.org/stable/modules/cross_validation.html>
[Accessed 10 May 2019].

[2] Sci-Kit Learn Documentation *1.10. Decision Trees* [Online]. Available at: <https://scikit-learn.org/stable/modules/tree.html#decision-trees>
[Accessed 3 May 2019].

[3] Stack Abuse *Implementing PCA in Python with Scikit-Learn* [online]. Available at: <https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/>
[Accessed 3 May 2019].