LOTI.05.046 PATTERN RECOGNITION

HOMEWORK 10 - MAJORITY VOTING

May 22, 2019

Quazi Saimoon Islam
Msc. Robotics and Computer Engineering

1 INTRODUCTION

Briefly, in majority voting Every model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes. If none of the predictions get more than half of the votes, we may say that the ensemble method could not make a stable prediction for this instance. Although this is a widely used technique, one may try the most voted prediction (even if that is less than half of the votes) as the final prediction.

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. Define 5 classifiers
3. Perform Cross Validation with 10 folds on each classifier Classification.
4. Implement majority vote classifier based on the 5 chosen classifiers
5. Perform Cross Validation with 10 folds on the majority vote classifier.

The following is a short description of the chosen classifiers for this assignment:

2.1 Decision Tree

From Homework 8 9, the decision tree classifier is very well known by now. It was included to obtain a comparitive understanding of how it performs compared to the other four. It would be interesting to compare its performance to others in thsi scope.

2.2 Random Forest

A random forest is an estimator that fits multiple of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but can be replaced if the right flags are used.

2.3 Gaussian Process

Gaussian Processes (GP) are a generic supervised learning method designed to solve regression and probabilistic classification problems. The Gaussian Process Classifier implements Gaussian processes (GP) for classification purposes, more specifically for probabilistic classification, where test predictions take the form of class probabilities.

2.4 K-Nearest Neighbour

K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. In K means algorithm, for each test data point, we would be looking at the K nearest training data points and take the most frequently occurring classes and assign that class to the test data. Therefore, K represents the number of training data points lying in proximity to the test data point which we are going to use to find the class.

2.5 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

3 RESULTS

3.1 Classification using decision tree

Using appropriate tools from the Sklearn library, it as easy to set up the classifiers and the following are the results obtained. The results of the performance of each ase are as below:

3.2 Performance of individual classification

1. The Accuracy of the Decision Tree Classifier with 10-fold Cross Validation is : 92.666667% (+/- 0.13)

2. The Accuracy of the Random Forest Classifier with 10-fold Cross Validation is : 92.666667% (+/- 0.13)

3. The Accuracy of the Gaussian Process Classifier with 10-fold Cross Validation is : 94.000000% (+/- 0.09)

4. The Accuracy of the K-Nearest Neighbour Classifier with 10-fold Cross Validation is : 94.000000% (+/- 0.09)

5. The Accuracy of the Support Vector Machine Classifier with 10-fold Cross Validation is : 96.000000% (+/- 0.11)

The support vector machines(SVM) seems to be the best classifier out of the 5 chosen. The decision tree and random forest have similar performance which make sense since the sample size is quite small.

3.3 Majority Vote

1. The Accuracy of the Majority Vote Classifier with 10-fold Cross Validation is : 94.000000% (+/- 0.09)

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].