Yunhao CHEN
yunhao.chen@polytechnique.edu

Lab session # 1
Data streaming 2022

12/02/22

Report

It is a classical dataset which concerns about predicting forest cover type from cartographic variables. There are continue variables (such as elevation and aspect) as well as categorical variables (such as soil type and wilderness area).

A **feature engineering** has been done on raw data. For example, combine the "Horizontal_Distance" and "Vertical_Distance" to compute an euclidian distance; convert the slope to a angle and compute its cosine value.

On this dataset, I've tested **PAClassifier, HoeffdingTreeClassifier, LogisticRegression** (converted to multi-class classifier by applying one-vs-rest strategy), **HoeffdingAdaptiveTreeClassifier and GaussianNaiveBayes**.

Fig. 1 shows the performance of these classifiers.

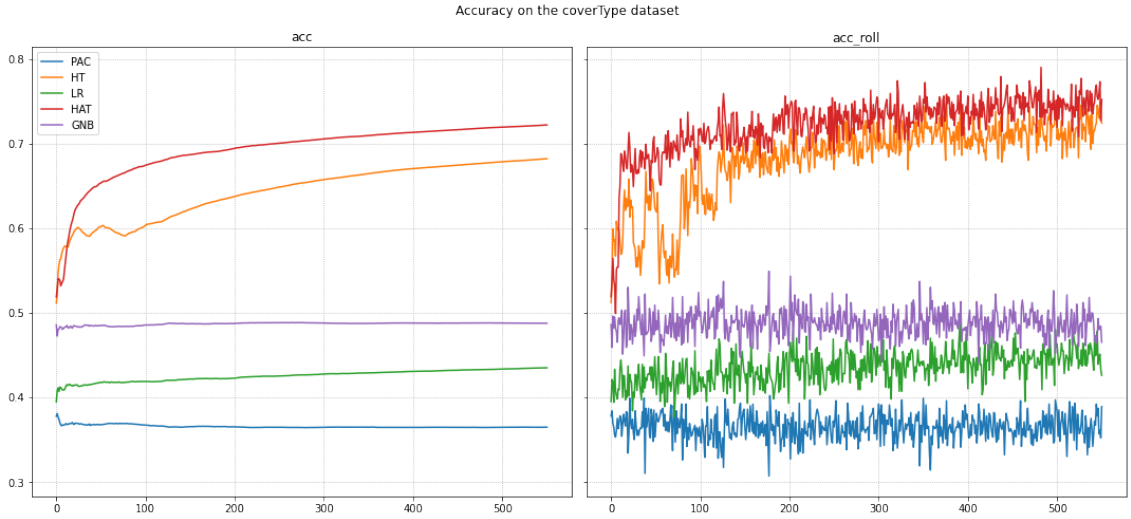

Figure 1: Accuracy

Respectively for PAClassifier, HoeffdingTreeClassifier, LogisticRegression, HoeffdingAdaptiveTreeClassifier and GaussianNaiveBayes:

- execution time: 50s, 716s, 236s, 1918s, 631s.
- test accuracy: 0.36, 0.72, 0.49, 0.75, 0.48

The accuracy on training set and test set shows that the tree-based models (HoeffdingTree & HoeffdingAdaptiveTree) are much more performant than others.
It is not surprising because generally the tree-based model handles very well this kind of classification problem with many features.
One thing to notice for GaussianNaiveBayes: due to the feature engineering, the independance between features is not valide anymore. This would degrade the performance of NaiveBayes for which the independance is a basic hypothsis.

The execution time of HoeffdingAdaptiveTree is more than HoeffdingTree (**1918 seconds vs. 716 seconds** for training); but its accuracy is higher than HoeffdingTree (**0.72 vs. 0.68** on training set; **0.75 vs. 0.72** on test set). There is a tradeoff between speed and performance. I would recommend **HoeffdingAdaptiveTree** as a classifier for this cover type predicting task.