CS 4320: Machine Learning

Classification: Ensembles

Use the Fashion MNIST data set at Kaggle.

You will use Ensemble Methods for this assignment.

In class we discussed the following ensemble methods.

- Bagging
- Random Forest
- Ada Boosting
- Gradient Boosting

Choose 2 of these ensemble methods (one choice must be either bagging or random forest, the other must be ada or gradient boosting.). In one of the ensemble methods, you may use DecisionTreeClassifier for the ensemble members, but not for both. Use hyper-parameter searching with cross validation to train the best possible model using the “micro-averaged” f1 metric for multiclass scoring.

Required Steps

- Use the data from the previous assignment.
- Use the same split of training data with an 80/20 split into your training and your testing data.
- Build a pipeline to process and fit the data.
- Use either grid or random search to search for the best hyper-parameters for your models.
- Maximize the F1 score (micro-averaged) as the training metric.
- Document the hyper-parameters you have found with the hyper-parameter search.
- Document rough estimates of the amount of time required to run the parameter searches, and how many hyper-parameter combinations were used.
- For each model type, train a model with your best parameters, predict how well it will do on your testing data using cross-validation. Measure its actual performance on your testing data. Record both of these results.
- Combine your results and a discussion of your process in a report.
- Include the cross-validation score and performance on your testing data from the previous two assignments.
- Submit the report (as PDF), and all source code.