Problem Set 7

Data Science 602: Data Analysis and Machine Learning

Spring 2022

Housekeeping note: This problem set is due after return from spring break, on March 31. Please remember final project proposals are also due 3/31.

This problem set is an extension of Problem Set 6. You will need the following artifacts from last week:

- The MNIST 784 dataset from OpenML, with dimensionality reduced to about 75%.
- The Support Vector Machine classifier.
- Recoded target variables, such that the target variable is 1 if the digit is less than 5, and 0 otherwise.

As with last week, please use the first 60,000 observations as training data, and the remaining 10,000 images as test data.

1. **Classifiers**. Construct 3 classifiers using different algorithms, not including the SVM model built last week, that classify the MNIST dataset with an F_1 score of at least 0.9. At least one classifier must use gradient boosting (AdaBoost, Gradient Boost, or xgboost). Show the F_1 score and classification report for each model.
2. **Voting ensemble model** Build a voting ensemble model that combines the three classifiers from the previous problem, in addition to the SVM model developed last week. What is the F_1 score of the ensemble model?
3. **Stacking ensemble model** Stacking uses a final classifier (often a logistic regression) that outputs an aggregate of the predictors. Repeat the previous problem using a StackingClassifier rather than voting to compute the final prediction. What is the F_1 score of the stacking classifier?
4. **Evaluation** At this point in the assignment, you have six classifiers:

   - the support vector classifier from last week,
   - the three classifiers from problem 1,
   - the voting classifier from problem 2, and
   - the stacking classifier from problem 3

   Identify the model with the highest F_1 score, and train this model with the full training dataset. Finally, score the test data against this model. Does the model demonstrate predictive validity (i.e., are the F_1 scores for the test data comparable to the training data)?