NYU | COURANT

MATH-GA.2047-001 Data Science in Quantitative Finance

Homework 11

Profs. Ivailo Dimov and Petter Kolm
Due: December 8, 2020

Instruction

This homework is to be done individually. No collaboration and/or code sharing permitted.

Objective

In this assignment, you will:

- Compare the performance of various boosting and tree classification methods.

Methodology and Deliverables

1. Classification Trees and Boosting
   From the examples demoed in class and your knowledge so far, build a binary classifier for the MNIST dataset that determines whether a an input image is digit 7 or not. Before you proceed make sure you split your dataset into train 70-30 train-test. Using 5-fold cross validation on the training set, find the optimal model parameters of the following Tree models:
   (a) **Gradient Boosting** with 'deviance' loss. Use the `GradientBoostingClassifier` class of sklearn grid search over the parameter space. Carefully describe the procedure of which parameters you choose to search over and explain your logic. Save the optimal model `clf_gb_opt`
   (b) **AdaBoost**. Repeat the same exercise with `AdaBoostClassifier` and save the optimal model `clf_ab_opt`.
   (c) **Random Forest 1**. Repeat the same exercise with `RandomForestClassifier` with 'gini' criterion and save the optimal model `clf_rf1_opt`.
   (d) **Random Forest 2**. Repeat the same exercise with `RandomForestClassifier` with 'entropy' criterion and save the optimal model `clf_rf2_opt`.

(e) For each of the optimal classifiers in (a)-(d) compute the precision-recall (PR) and ROC curves. Does any of these classifier dominate the other ones? Comment on your findings.

(f) Combine the optimal classifiers (a)-(d) via (1) a `VotingClassifier` or (2) a `StackingClassifier` and calculate the PR and ROC curves. Does any of the newly formed classifiers dominate the classifiers in the previous part? Comment on your findings.

(g) Repeat the same exercise (a)-(f) for all the other 8 digits. For each digit, find the classifier that produces the best F1 score and plot 4 images that your classifier mislabels for the given digit. Comment on your findings.

2. Multivariate Polynomial Features

How many polynomial features $N_F(d, n)$ are there as a function of polynomial degree d and number of variables n? For example for a bivariate degree 3 polynomial there are 4 features (x^3, x^2y, xy^2, y^3).

3. Comparing Feature Map Regressions

Generate 500 points from the model $y(x) = (x-1)^2(x+1)^2 + \varepsilon$ where $\varepsilon \sim N(0, 0.3)$ and $x \sim N(0, 3)$. Using Implement and compare the performance of (1) polynomial regression (2) B-spline regression (3) GradientBoostingRegression (4) Gaussian Kernel Regression whith L_2. Choose your optimal model parameters using 5-fold cross-validation. Test the performance of each optimal model against a newly simulated 100 points. Comment on your results.