Ensemble of Models and Random Forests

Lesson 5 – Section 2

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Overview

Random Forest

Demonstration in Python

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Why Ensemble?

- Think about a patient with some complicated disease
  - A group (panel) of doctors are involved in diagnosis
  - Each doctor may diagnose based on a specific set of data, and/or on his own specific domain expertise (model)
  - The final diagnosis is made by majority voting, weighted average (some doctors might be more experienced, their diagnosis take higher weights than others)
- Benefits of ensemble models:
  - Usually perform better than each individual model
  - Reduce the variance in the predictions, generalize better than individual models
  - Make the process of building the machine learning solutions more scalable

Different Ways of Ensembling

- Bagging:
  - Each model is trained on a subset of observations and/or features independently
- Boosting:
  - Model i+1 is trained on a sampled subset of observations, where observations that are not classified corrected by model i have higher probability of being sampled

Different Ways of Ensembling

- Different ways of making the final decision from the decisions of multiple models to be ensembled:
  - –Simple average
  - –Weighted average
    - Based on performance of each model (Random Forest, Boosted Decision Tree)
    - Weights are determined by another machine learning model

Random Forest (Decision Forests)

Ensemble of multiple independently trained decision trees

- –Each tree is trained using a sample of observations and a sample of independent variables
  - Think about three doctors diagnosing heart disease. One doctor is trained by just looking at ECG, one doctor is a Chinese medicine doctor who is trained only by only touching the pulse, and one doctor is trained by looking at the ultrasound image
  - Each doctor is trained on data of different patients (there might be overlapping among the sets of patients)

Random Forest (Decision Forests)

Advantages of Random Forest:

- Significantly better performance than individual trees
- Automatic Feature Selection
- Less risk of overfitting
- Can be parallelized easily (training of multiple doctors can happen at the same time independently)

Disadvantages:

- Less interpretability than decision trees
- In some algorithms, data is copied in order to train each tree. Has higher requirement in memory space than individual trees.

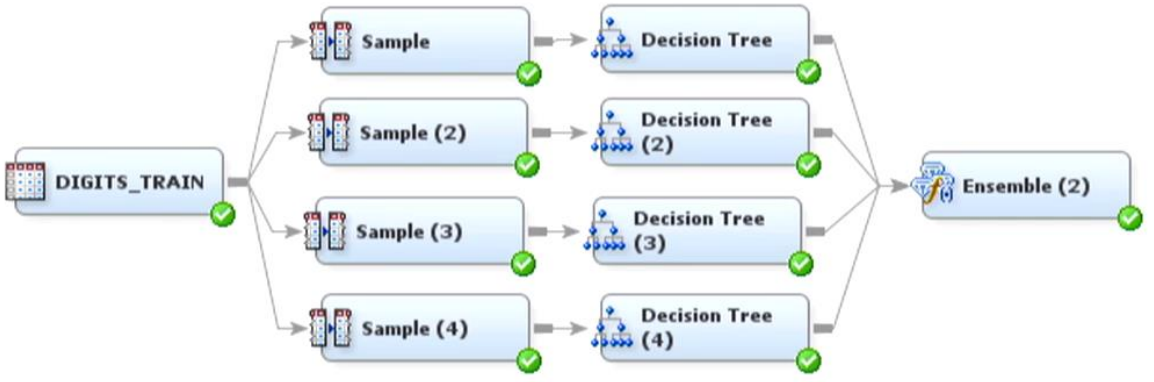
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Random Forests

- Combination of decision trees and bagging concepts
- A large number of decision trees is trained, each on a different bagging sample
- At each split, only a random number of the original variables is available (i.e. small selection of columns)
- Data points are classified by majority voting of the individual trees

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Random Forests



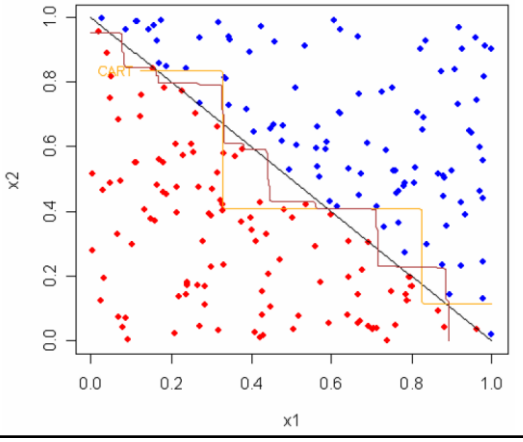
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Bagging: reduces variance – Example 1

- Two categories of samples: blue, red
- Two predictors: x1 and x2

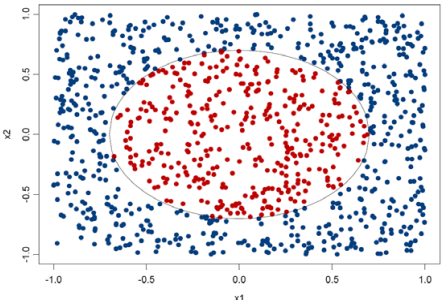
Diagonal separation…hardest case for tree-based classifier

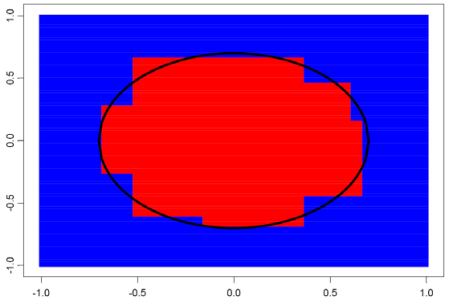
- Single tree decision boundary in orange.
- Bagged predictor decision boundary in red.



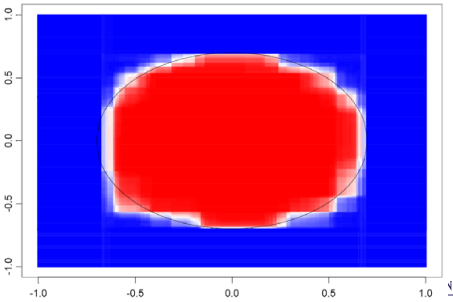
Bagging: reduces variance – Example 2

Ellipsoid separation →
Two categories,
Two predictors





Single tree decision boundary



100 bagged trees..

Random forests

```
D = training set                          F = set of tests
k = nb of trees in forest                 n = nb of tests

for i = 1 to k do:
   build data set Di by sampling with replacement from D
   learn tree Ti (Tilde) from Di:
      at each node:
         choose best split from random subset of F of size n
      allow aggregates and refinement of aggregates in tests

make predictions according to majority vote of the set of k trees.
```

Random Forest: How Many Trees to Train?

- Rule of thumb:
  - –Classification problem: $\sqrt{p}$
  - –Regression problem: p/3
- Optimal number is still case by case
  - –Start with rule of thumb
  - –Tune it to optimize performance

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

>Introduced Random Forest
- –A Random forest of decision Trees
- –Bagging
- –Boosting

>Practiced Random Forest in Python