Random Forests

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Lecture Objectives

- Understand the basics of ensemble learning and bagging.
- Fit and evaluate Random Forests models in Python.

Motivation

- Decision trees are simple to understand.
  - They provide human-readable explanations.
- At its core, there is an important trade-off:
  - Growing large trees leads to overfitting (excellent performance on training data, poor performance on test data).
  - Growing small trees leads to weak learners (similar but poor performance on both training and test data).
- **Random Forests** try to address this issue by combining many small trees.

Ensemble Learning

- Suppose we have multiple prediction models.
  - None of them clearly outperforms the other ones.
  - Their performance is okay, but not great.
- We can combine their predictions into a “super” prediction model.
  - If classification, take a majority vote.
  - If regression, take average.
- This is **ensemble learning**: weak classifiers can be combined into a better model.

Exercise

See Jupyter Notebook on Random Forests.

Bagging and Random Forests i

- The main issue with the approach above is that these trees tend to be highly correlated.
- *Better*: We want a diversity of opinions!
- **Bagging** (short for *bootstrap aggregating*) was first introduced as a general idea for combining weak classifiers.
    - Sample with replacement from the training data.
    - Fit model on bootstrapped dataset.
    - Repeat B times and combine the predictions/classifications.
- As a bonus, we can measure the **out-of-bag** performance of each model.

Bagging and Random Forests ii

    - Evaluate accuracy (or other metrics) on samples that were not selected.

- In Random Forests, we combine bagging with an other idea: randomly select a *subset* of the features.
    - At each iteration, we both sample with replacement and select a subset of the features.
- **Why?** By combining these two ideas, we reduce the correlation between each classifier, improving overall performance.
- There are two main hyper-parameters:
    - How many trees/iterations.
    - How many features are sampled.

Bagging and Random Forests iii

- The main drawback of random forests (over decision trees) is that they are harder to interpret.
    - You need to combine multiple decision trees.
    - It's also harder to visualize: you would need to look at all trees!

Exercise

See Jupyter Notebook on Random Forests.

Final Remarks

- Ensemble Learning can be used to combine prediction models built using different methods!
    - E.g. Logistic regression, Decision Trees, K-NN, and SVMs
    - See van der Laan, Polley, and Hubbard (2007) on *Super Learner*.
- Bagging is an important technique that increases the variability among the fitted decision trees.
- Another powerful approach is **boosting**, where each learner is weighted according to how well they perform on "hard" cases.
    - E.g. AdaBoost and XGBoost