KBTU KAZAKH-BRITISH TECHNICAL UNIVERSITY

Introduction to Machine Learning
Week 13

Olivier JAYLET

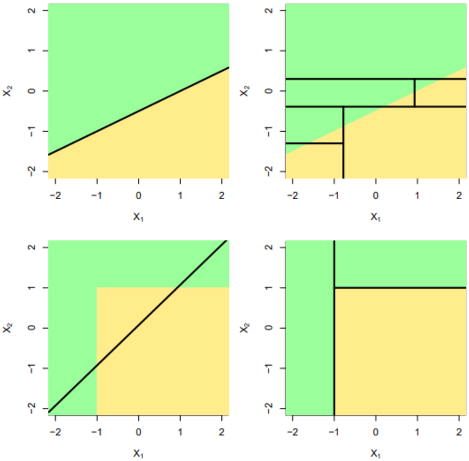
School of Information Technology and Engineering

Tree Vs. Linear model

Which model is better ?

- Linear regression : $f(X) = \beta_0 + \sum_{j=1}^{K} X_j \beta_j$
- Regression Tree : $f(X) = \sum_{j=1}^{J} c_j \mathbb{I}(x \in R_j)$
- If the relationship between y and $x_1, \ldots, x_K$ is linear: a linear model should perform better.
- If the relationship between y and $x_1, \ldots, x_K$ is highly non-linear and complex: a tree model should perform better.

Trees Vs. Linear Models



Pro

- Graphical decision making tool Some people believe that decision trees more closely mirror human decision-making than do the regression and classification methods seen in previous chapters.
- Trees can easily handle qualitative predictors without the need to create dummy variables.

Cons

- Trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches.
- Trees are unstable (non-robust).

Bootstrapping

”How much would our results vary if we had drawn a different sample from the same population ?”

We would like to estimate the variability of our results.

Problem, traditional statistical inference methods can be impractical:

- for complex statistics
- non-standard scenarios
- limited data

Bootstrapping

Bootstrapping is a powerful method that allows us to :

- Estimate the sampling distribution of a statistic without requiring knowledge of the true population distribution.
- Quantify uncertainty (e.g., confidence intervals) for estimates derived from small or complex datasets.
- Avoid relying on restrictive assumptions (e.g., normality or independence) that may not hold in real-world data.

Idea behind bootstrapping :

- The bootstrap experiment is based on the plug-in principle : if something is unknown, then substitute an estimate for it
- We can substitute unknown parameters and/or distributions

Example :

Suppose you want to figure out the average number of steps people in Almaty walk each day.

- Measuring every single resident in the city is impossible, as there are too many people in Almaty.
- So, you collect a small sample by interviewing 100 or more random people in Almaty and asking them how many steps they walk each day.

Is this small sample of 100 people enough ?

If you had chosen a different group of 100 people, would your average change significantly ?

Idea behind Bootstrapping

Instead of going back and interviewing another 100 or more people (which is time-consuming and expensive), we can use bootstrapping.

Idea behind Bootstrapping

- Collect the daily step counts of the 100 people you surveyed (for example):

$$8000, 7000, 6000, 7500, 9000, ..., 8500$$

- Use these 20 people, and resample with replacement to create new ”sample groups.” For example:
  - Group 1: [8000, 9000, 7000, 7000, 8500, ...]
  - Group 2: [7500, 7500, 8000, 9000, 8000, ...]
  - Group 3: [6000, 9000, 8500, 6000, 7000, ...]
- For each new group, calculate the average number of steps
- Repeat this process to create hundreds or thousands of resampled groups and calculate their averages.

What do we learn ?

By looking at all the averages from resampling, you can estimate:

- The variability in the average step count for Almaty residents.
- A confidence interval (e.g., ”We’re 95 sure the true average daily steps fall between x1 and x2”).

Even though you only surveyed 100 people, bootstrapping allows you to better understand the potential range of the average.

Connecting with Bagging

If we applied this to machine learning:

- The step counts are like your dataset.
- Instead of calculating averages, you train different models on separate bootstrap samples of the data.
- Each model makes predictions, and you combine them (e.g., by averaging) to improve the overall accuracy and stability of your predictions.

Ensemble Methods

- An **ensemble method** combines multiple simple models, called *building blocks* or *weak learner*.
- These simple models, may give mediocre predictions on their own.
- These ensemble methods aim to produce a single powerful model.
- Examples of ensemble methods:
  - Bagging
  - Random Forests
  - Boosting
  - Bayesian Additive Regression Trees
- The *weak learners* in these methods are often regression or classification trees.

Bagging and Random Forest

How Bagging and random forest work intuitively :

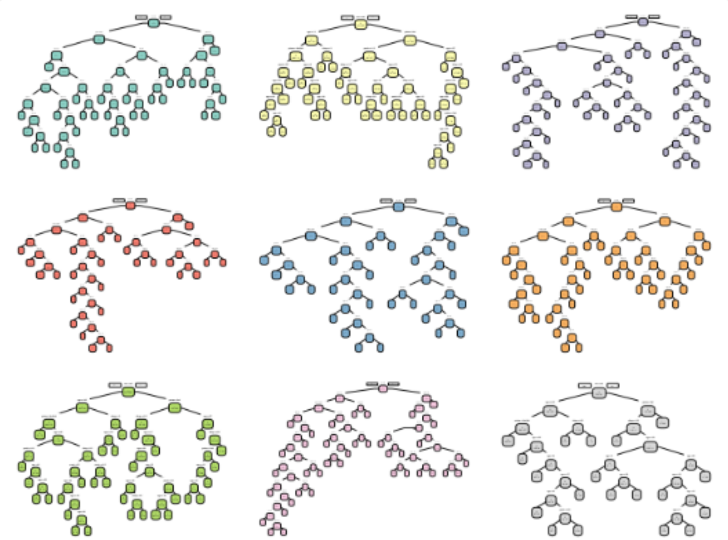
- Based on your symptoms, suppose a doctor diagnoses an illness that requires surgery
- Instead asking one doctor, you may choose to ask several
- If one diagnosis occurs more than any others, you may choose this one as the final diagnosis

Bagging

Algorithm Bagging

```
Select the number of trees $B$ and tree depth $D$;
for $b \leftarrow 1$ to $B$ do
    Generate a bootstrap sample from the original data
    Estimate a tree model of depth $D$ on this sample
```

Generate several trees by bootstrapping



Why bootstrapping CART model ?

Bagging = Boostrap aggregating

Prediction

- Regression : Average the resulting predictions
- Classification : Take a majority vote

Impact of bootstrapping :

- Averaging a set of observations reduces variance
- Hence increase the prediction accuracy
- Loss of interpretability

Random Forest

Algorithm Random Forest

```
Select the number of trees $B$, subsampling parameter $m$, and tree depth $D$;
for $b \leftarrow 1$ to $B$ do
    Generate a bootstrap sample from the original data
    Estimate a tree model on this sample
    for each split do
        Randomly select $m$ of the original covariates ($m < P$)
        Split the data with the best covariate (among the $m$ )
```

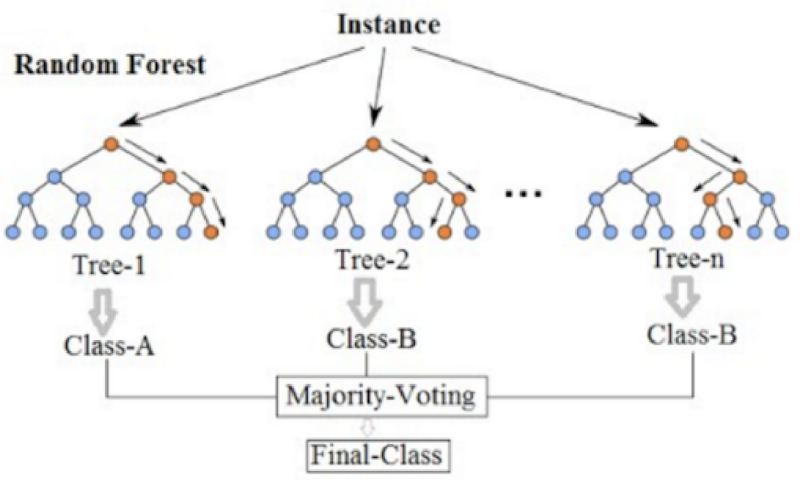
- Random Forest = Bagging + subsample covariates at each nodes.
- Bagging is a special case of RF with m = P.

Random Forest : Why subsampling covariates ?

- Suppose there is one very strong covariate in the sample
  - Most or all trees will use this covariate in the top split
  - All of the trees will look quite similar to each other
  - Hence the predictions will be highly correlated
- Averaging many highly correlated quantities does not lead to a large reduction in variance
- Random forests overcome this problem by forcing each split to consider only a subset of the covariates

Random forests decorrelate the trees.

Random Forest



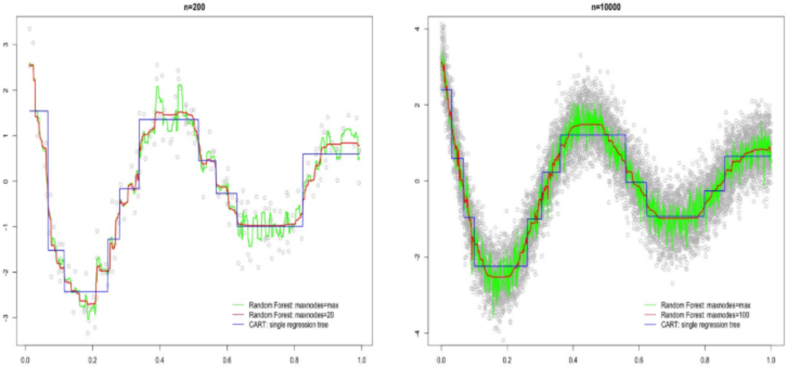
Random Forest : Overfitting

- As B increases: average effect over trees $\rightarrow$ no overfitting
- As D increases: overfitting is argued to be minor

The goal is to grow trees with as little bias as possible. The high variance that would result from deep trees is tolerated because of the averaging over a large number of trees.

... However, a simple example shows that it can be problematic.

Random Forest : Overfitting



- improvement of random forest over a single regression tree
- overfitting can be very large without controlling tree depth

Random Forest : Out-of-Bag (OOB)

No need to perform cross-validation:

- By bootstrapping, each tree uses around 2/3 of the obs. The remaining 1/3 obs are referred to as the out-of-bag (OOB) obs
- Use OOB observations for out-sample predictions
- An OOB-MSE can be computed over all OOB predictions
- OOB are automatically used to control and validate each tree, during the training process.

The OOB approach for estimating the test error is particularly convenient with large sample, for which CV would be onerous.

Bagging and Random Forest

Advantages :

- They tend to work well for problems where there are important nonlinearities and interactions.
- They are robust to the original sample and more efficient than single trees.

Disadvantages :

- The results are not intuitive and difficult to interpret