QBUS6850 Week 6

Ensemble Methods 2 - Random Forests

Dr Stephen Tierney

The University of Sydney Business School

The Plan

- **Part 1 (Week 6): Understand decision trees and How to make an ensemble (forests)**
- Part 3 (Week 8): Advanced ensembles 1 (boosting)
- Part 4 (Week 9): Advanced ensembles 2 (gradient boosting)

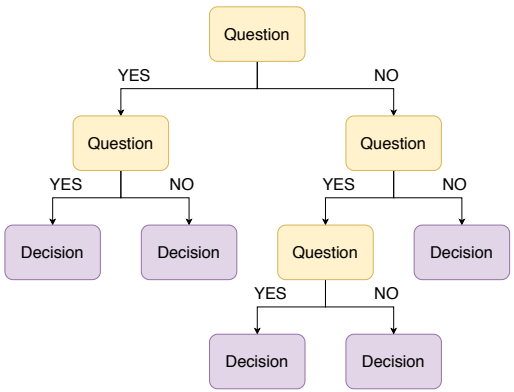
Random Forests

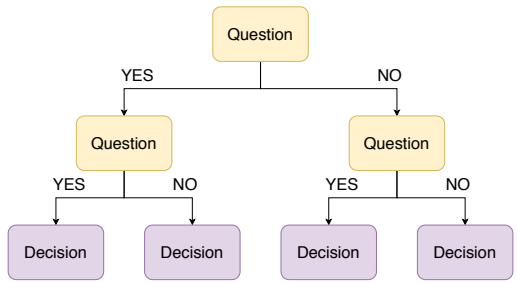
Random Forests

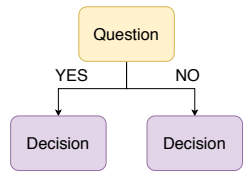
A **Random Forest** is a set of decision trees.

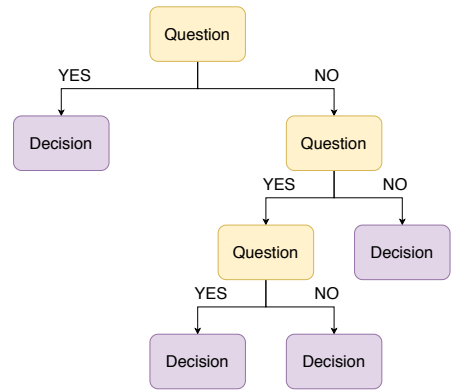
The final prediction is the most common prediction from the trees.

Random Forests









Random Forests

For classification, the prediction from a Random Forest is given by

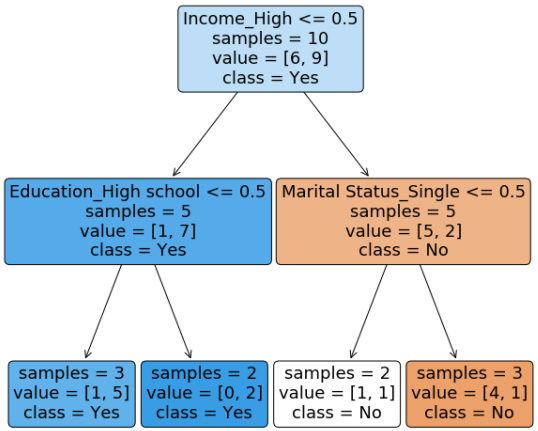
$$\overline{F}(\mathrm{x}, \beta) = \text{Mode}\left\{\widehat{F}_b(\mathrm{x}, \beta^{(b)})\right\}_{b=1}^{B}$$

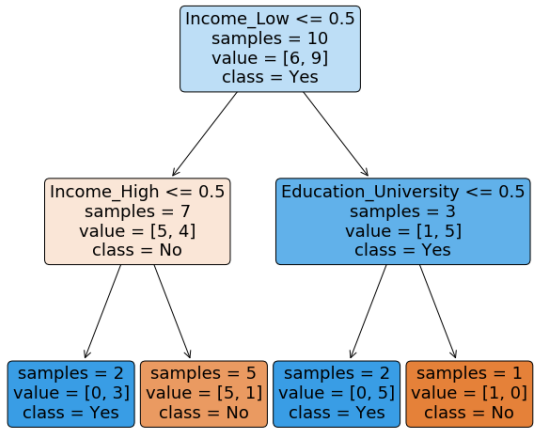
where

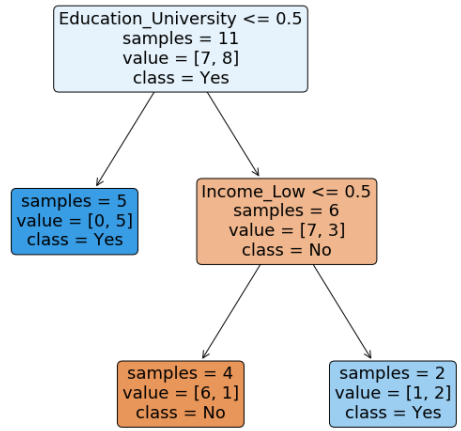
- B is the number of trees
- $F_b(\mathrm{x})$ is the prediction from tree b

Random Forests - Example

A random forest with three trees has been trained on the purchases dataset







Predict the class of a new customer:

- married
- high income
- high school educated

Answer

- Tree 1: No (third leaf)
- Tree 2: No (second leaf)
- Tree 3: Yes (first leaf)

Final: No

Why Do Forests Outperform Trees?

Modelling Mistakes

Suppose that B is the number of trees in a Random Forest.

Since we take the mode as our final prediction, a Random Forest makes a mistake when $\geq \lceil \frac{B}{2} \rceil$ trees make a mistake.

Example: There are three trees in a Random Forest, i.e. $B = 3$. A mistake is made when 2 or more trees are incorrect.

Modelling Mistakes

We will use

$$P(X \geq \left\lceil \frac{B}{2} \right\rceil)$$

to denote the probability that a Random Forest makes a mistake, where X is a random variable that counts the number of mistakes from the trees.

Modelling Mistakes

If we make the following **assumptions:**

- each tree produces a correct or incorrect prediction
- the predictions from each tree are independent

Therefore:

- The predictions from each tree are binary, i.e. Bernoulli trials.
- The sum of incorrect predictions, X, is a **Binomial** random variable.

Modelling Mistakes

The cumulative distribution function of a Binomial distribution is

$$P(X \leq k) = \sum_{i=0}^{k} \binom{B}{i} p^i (1-p)^{B-i}$$

where k is the number of successes, B is the number of trials and p is the probability of "success".

Modelling Mistakes

Therefore $P(X \geq k) = 1 - P(X \leq k-1)$ and

$$P(X \geq \left\lceil \frac{B}{2} \right\rceil) = 1 - \sum_{i=0}^{\left\lceil \frac{B}{2} \right\rceil - 1} \binom{B}{i} p^i (1-p)^{B-i}$$

Which we can simplify to

$$P(X \geq \left\lceil \frac{B}{2} \right\rceil) = \sum_{i=\left\lceil \frac{B}{2} \right\rceil}^{B} \binom{B}{i} p^i (1-p)^{B-i}$$

Modelling Mistakes

In the case of a tree, the probability of an incorrect prediction is given by its misclassification rate, which we denote as ϵ.

We use the average misclassification rate of all trees $\bar{\epsilon}$ in our model.

Updating our formula

$$P(X \geq \left\lceil \frac{B}{2} \right\rceil) = \sum_{i=\left\lceil \frac{B}{2} \right\rceil}^{B} \binom{B}{i} \bar{\epsilon}^{i} (1 - \bar{\epsilon})^{B-i}$$

Modelling Mistakes - Example

Suppose we have three trees in our forest where the average misclassification rate is $\bar{\epsilon} = 25\%$.

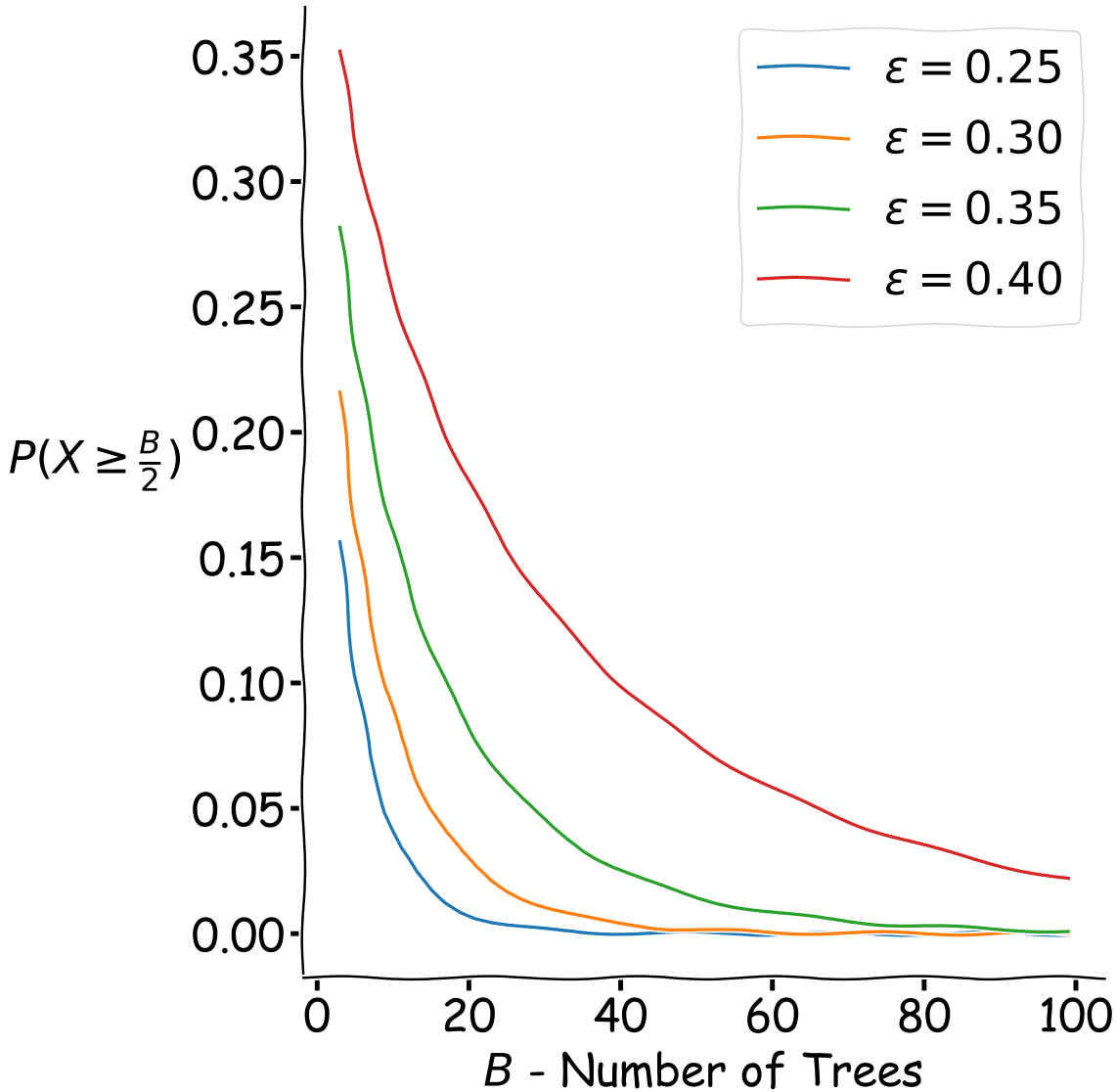
Since $B = 3$ and $\left\lceil \frac{B}{2} \right\rceil = 2$ the probability of an incorrect classification is

$$\begin{aligned} P(X \geq 2) &= \sum_{i=2}^{3} \binom{3}{i} \bar{\epsilon}^i (1 - \bar{\epsilon})^{3-i} \\ &= \sum_{i=2}^{3} \binom{3}{i} 0.25^i (1 - 0.25)^{3-i} \\ &= 0.156 \end{aligned}$$

The random forest is less likely to make an incorrect prediction than a single tree since $0.156 < 0.25$.

Modelling Mistakes - Properties

Generally, the more trees the lower the probability of errors.



Training a Random Forest

Assumptions

Earlier we made a very important assumption:

- the predictions from each tree are independent

Our training procedure must ensure the assumption is met.

Independent Trees

What are independent trees?

We have independent trees when the prediction of one tree is not correlated with any other. For example this means that the prediction from tree 1 shouldn't influence tree 3.

How do we get independent trees?

With a combination of:

- bootstrap aggregating
- random feature selection

Intuition

If we train each tree on the same dataset then the trees will be identical since training trees is deterministic.

We want the trees to bring different perspectives from the data.

Otherwise there's no benefit to using a forest over a tree.

Bootstrap

Definition (Bootstrap Dataset)

Let $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ *is our dataset.*

A random sample of size N from this dataset with replacement *is called a Bootstrap dataset.*

A Boostrap dataset is denoted by

$$\mathcal{D}^* = \{(x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_N^*, y_N^*)\}$$

Bootstrap Example

Consider a data set with five items, i.e., $N = 5$, shown below with three bootstrapped samples

Original Data

Customer	Income	Education	Marital Status	Purchase
1	Medium	University	Single	Yes
2	High	University	Single	No
3	High	University	Married	No
4	Low	University	Single	Yes
5	Low	High school	Single	Yes
6	Low	High school	Married	No
7	Medium	High school	Married	Yes
8	High	University	Single	No
9	High	High school	Single	Yes
10	Low	High school	Single	Yes
11	High	High school	Married	Yes
12	Low	University	Married	No
13	High	University	Single	No
14	Medium	University	Married	Yes
15	Medium	High school	Single	Yes

Bootstrap Sample 1

Customer	Income	Education	Marital Status	Purchase
13	High	University	Single	No
6	Low	High school	Married	No
1	Medium	University	Single	Yes
4	Low	University	Single	Yes
12	Low	University	Married	No

Bootstrap Sample 2

Customer	Income	Education	Marital Status	Purchase
4	Low	University	Single	Yes
8	High	University	Single	No
10	Low	High school	Single	Yes
4	Low	University	Single	Yes
6	Low	High school	Married	No

Bootstrap Sample 3

Customer	Income	Education	Marital Status	Purchase
3	High	University	Married	No
5	Low	High school	Single	Yes
8	High	University	Single	No
7	Medium	High school	Married	Yes
9	High	High school	Single	Yes

Since the sampling is with replacement some data might be repeated.

Bagging: Bootstrap Aggregating

Bagging is an ensemble method where individual models are trained on different bootstrap samples.

Bagging: Bootstrap Aggregating

The generic Bagging procedure is:

1. Sample B bootstrap samples from $\mathcal{D}$
2. Train B different classifiers on their respective bootstrap sample
3. For a new example, let all classifiers predict and take a majority vote (or average), i.e., for new example x, we calculate

$$\overline{F}(\mathrm{x}, \boldsymbol{\beta}) = \text{Mode}\left\{\widehat{F}_b(\mathrm{x}, \beta^{(b)})\right\}_{b=1}^{B}$$

Random Features

When we train each tree, we restrict ourselves to p randomly chosen features at each decision node.

This adds another layer of randomisation and decreases the correlation amongst trees.

This is where the name **Random Forests** comes from.

Training Algorithm

1. For tree $b = 1$ to B:
   - 1.1 Generate a bootstrap sample of size N from the training data
   - 1.2 Train tree T_b using the bootstrapped data, by recursively repeating the following steps for each leaf node, until stopping criteria is reached:
     - 1.2.1 Select p variables at random from the d variables ($p \leq d$)
     - 1.2.2 Pick the best variable/split-point among the p
     - 1.2.3 Split the node into two decision nodes
2. Output the ensemble of trees $\{T_b\}_{b=1}^{B}$.

Analysis

Independence

Does our training algorithm create independent trees?

Independence

You can view each bootstrap sample as if we were sampling from the original population. We take a subset of the population and return it.

Since each of these bootstrap samples are independent and identically distributed[1], we can expect our trees to be i.i.d.

[1] in the same way that our original sample is i.i.d.

Independence

However some features might dominate in terms of information gain and would be selected at decision nodes more often than others.

This would make our trees correlated.

However since we restrict each decision node to a random subset of features we can minimise the correlation amongst trees.

Independence - In Practice

It is likely that there will be some dependency or correlation between trees.

Usually the correlation is so low that it doesn't impact the model.

Bias and Variance

Let Y_b be the random variable that indicates the misclassification rate of tree b and $\overline{Y}$ be the average rate for a forest.

Since each tree generated in bagging is identically distributed (i.d.), the expectation of the average misclassification rate $E[\overline{Y}]$ is the same as the expectation of any one of them $E[Y_b]$.

This reveals that the bias of a forest is the same as that of the individual trees, and the only hope of improvement is through variance reduction.

Bias and Variance

Assume that:

- each Y_b is identically distributed with variance σ^2
- the average correlation between each tree is ρ

Then the variance of $\overline{Y}$ is

$$\mathrm{Var}(\overline{Y}) = \frac{\sigma^2}{B} + \frac{B-1}{B}\rho\sigma^2$$

As we increase the number of trees, we can reduce the variance of the forest, to less than the variance of a single tree.

This provides another perspective showing that a forest should outperform a single tree.

Connection to our Mistake Model

Earlier we built a model for $P(X \geq \lceil \frac{B}{2} \rceil)$ that required us to use an average misclassification rate for our trees.

Hopefully the previous slides convinced you that this is a reasonable assumption in theory since $E[\overline{Y}] = E[Y_b]$.

In practice we require a large number of trees for this to hold.

Practical Considerations and Extensions

Avoiding Ties

Since we use the mode of our trees in the case of classification we might end up with ties if we use an even number of trees.

You can avoid this with an odd number of trees.

Regression with Random Forests

Regression with a Random Forest is similar, except the trees are trained as regression trees and the final prediction is given by the mean of trees

$$\overline{F}(\mathrm{x}, \boldsymbol{\beta}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{F}_b(\mathrm{x}, \boldsymbol{\beta}^{(b)})$$

Suggested Values for p

Intuitively, reducing p will reduce the correlation between any pair of trees in the ensemble.

For classification, the suggested value for p is $\sqrt{d}$, for regression, the suggested value is $\frac{d}{3}$.

Hyperparameters

A Random Forest has many hyper-parameters to set

- number of trees
- depth of each tree
- other stopping criteria of the trees such as min entropy
- number of features at each decision node

As always, the best way to do this is with **cross validation**.

Wrapping Up

Advantages of RF

Advantages of RF

- It is very robust and rarely overfits.
- It is empirically one of the best performing algorithms available.
- It runs efficiently on large datasets.
- It can handle thousands of input variables without prior variable selection.

Disadvantages of RF

- It is harder to interpret than a single decision tree
- There are more parameters to tune