Decision Trees and Random Forests

1 Decision Trees

There might be multiple decision trees for deciding the same thing from different conditions. To decide which is best, we use Gini Impurity

$$\text{Gini Impurity} = 1 - (\text{the probability of Yes})^2 - (\text{the Probability of No})^2$$

The lower the value the better

From a raw table of data to a decision tree:

1. Calculate all of the Gini Impurity values
2. If a node itself has the lowest value, leave it as a Leaf node
3. If separating the data results in an improvement, then pick the separation with the lowest Gini impurity value

1.1 Numeric Data

To get impurities

1. Sort the values lowest to highest
2. Calculate the average for adjacent values
3. Calculate the impurity values for each average weight

To Build a tree:

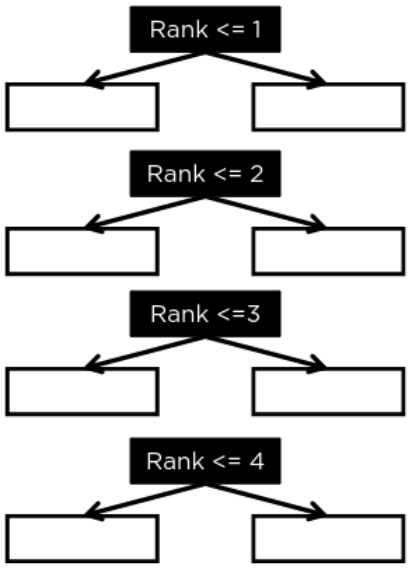
1. Yes/no questions at each step
2. Numeric data, like patient weight

1.2 Ranked Data and Multiple Choice Data

Ranked Data

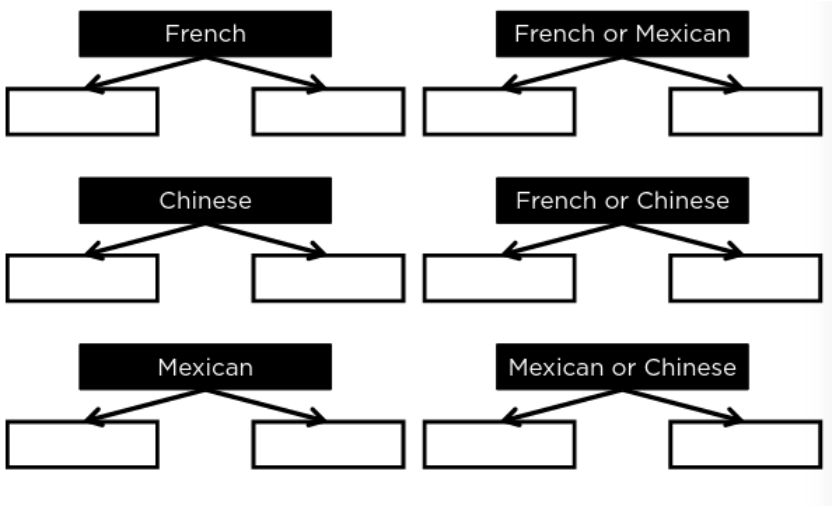
- Built tree from **Ranked Data**

Rank my ML lectures	Pass the exam
2	No
5	Yes
4	Yes
3	No
...	...



Multiple Choices Data

Cuisine Choice	Pass the exam
Mexican	Yes
Chinese	Yes
French	No
Chinese	Yes
...	...



2 Random Forests

Why Random Forests:

- Decision Trees are easy to build, use and interpret, but not flexible when classifying new samples
- Random forests combine the simplicity of decision trees with flexibility for better accuracy

2.1 How to build a random forest

Step 1 - Create a "bootstrapped" dataset:

- Same size as the original dataset
- Randomly selected samples from the original dataset
- Samples can be selected more than once

Step 2 - Build a decision tree using "bootstrapped" dataset, but only use a random subset of variables, e.g. 2

Step 3 - Go back to step 1 and repeat: make a new bootstrap dataset and build a tree considering a subset of variables at each step (ideally 100's of times)

- Using a bootstrapped sample and considering only a subset of the variables at each steps results in a wide variety of trees
- The variety makes random forests more effective than individual Decision Trees

2.2 How to use a random forest

- Take the data and run it down the first tree we built
- Keep track of the result

Definition: Bagging

Bootstrapping the data plus using the aggregate to make a decision