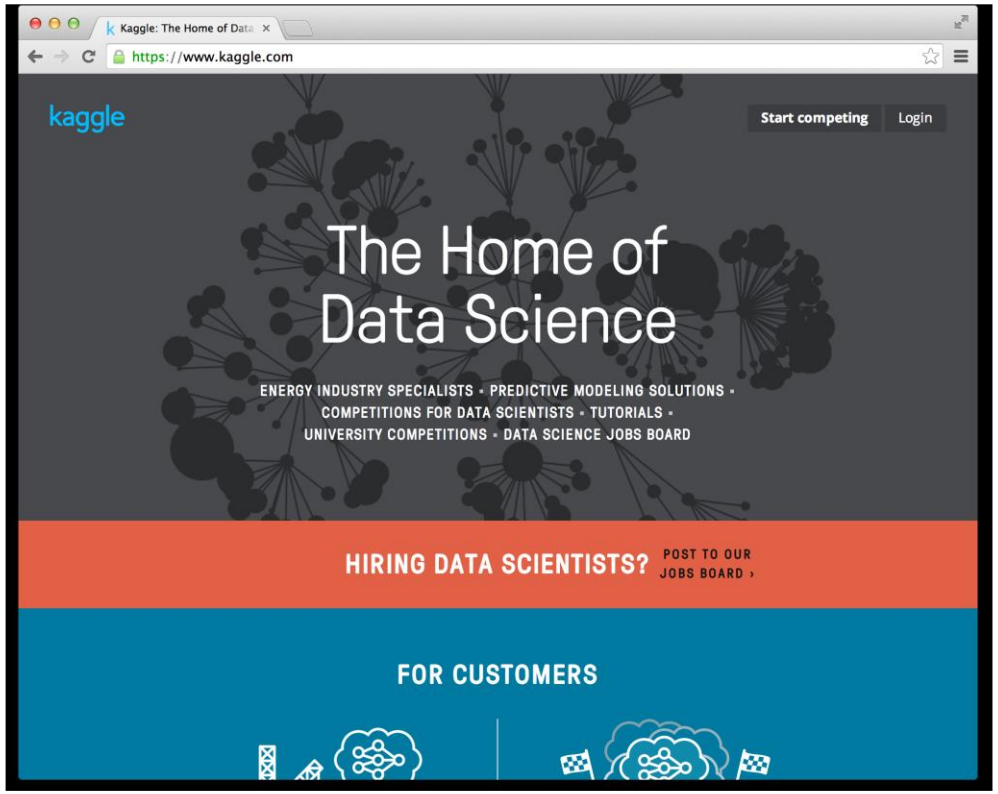
UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Features & Trees

2019

Feature extraction and selection

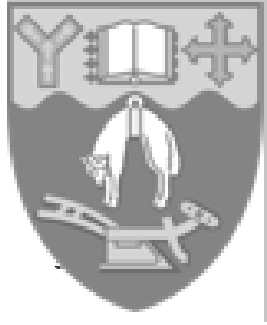
- *"Feature extraction and selection are the most important but underrated steps in machine learning. Better features are better than better algorithms."*, Will Cukierski, Kaggle



Example Twitter data

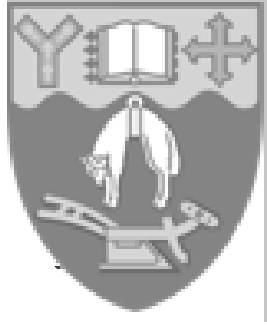
- N = 14,252
- Senators = 79
- Timeframe = over nine weeks around the 2018 US midterm elections
- Features that may be useful to predict retweet:
  - number of friends
  - number of followers
  - URLs
  - emoticons
  - sentiment score (e.g., positive/negative/neutral)
  - activity levels (e.g., number of posted tweets)
  - date/time the tweet posted
  - hashtags
  - @username

Large numbers of features



- In late 1990s few domains explored more than 40 features
- Now it is not uncommon for the number of features to be very large:
  - Gene expression:
    - Microarray data says what genes are expressed in a sample (e.g. cancer biopsy)
      - 6,000 to 60,000 genes (features)
      - 100 patients in each category (cancer/non-cancer)
  - Text classification:
    - Bag of words
    - 15,000 effective words (features)
    - 50,000 to 800,000 documents

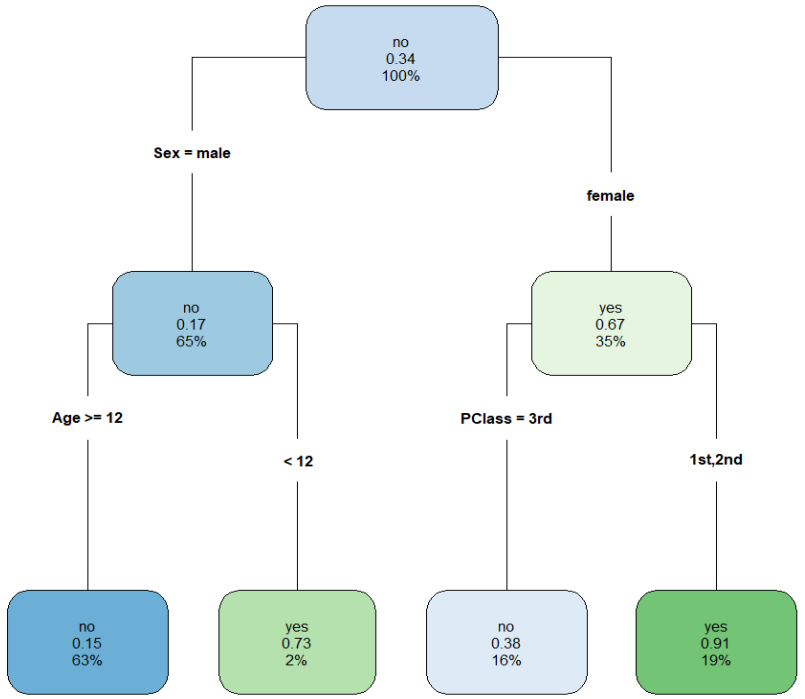
Feature selection



- Feature selection: choosing features to include in a model
- Why do feature selection?
  - Reduce measurement and storage requirements
  - Reduce time for both training and model utilisation
  - Facilitate visualisation and improve understanding
  - Defy the curse of dimensionality and improve performance
- Not looking to rank individual features but instead find *useful* subsets for building good predictors
  - Relevant features that are highly correlated with features already in a subset would therefore be excluded.

Decision Trees

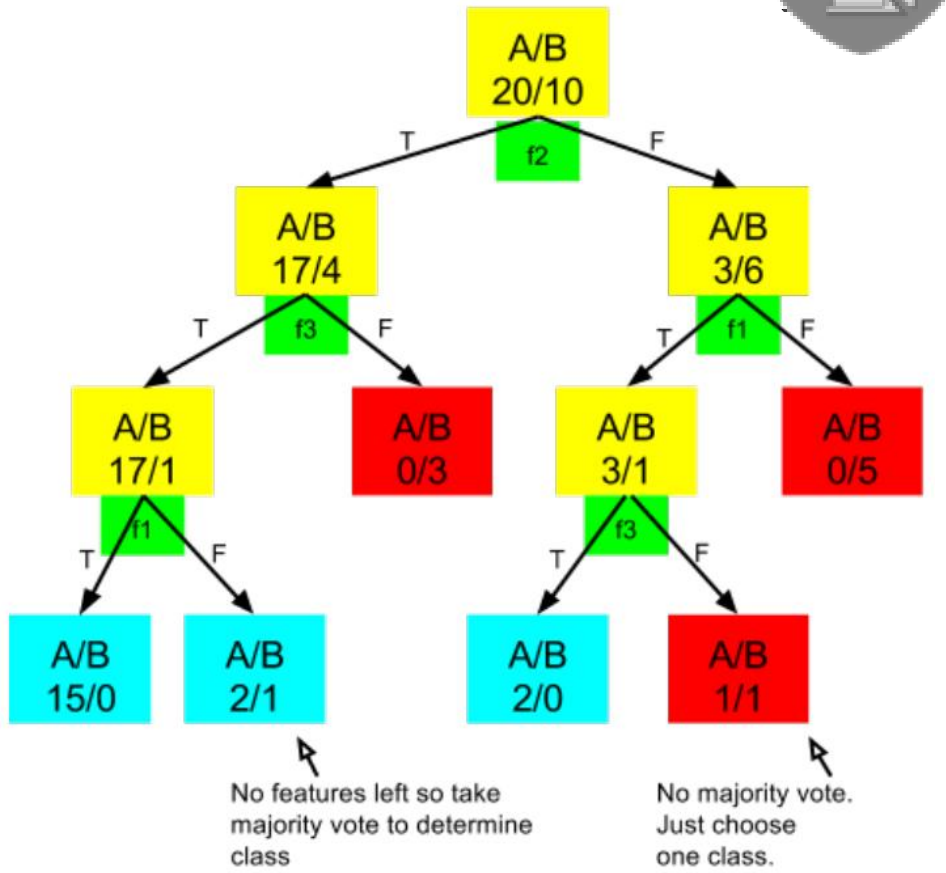
- Embedded selection:
  - Learning algorithms explicitly selects features
  - creates easy to visualize decision rules for predicting a categorical(classification tree) or continuous(regression tree) data



- Titanic Data
- The passenger list from the Titanic records the Age, Gender, Passenger Class and Survival of each passenger on the boat.
- We can use decision tree to look at the survival of passengers on the boat.

Decision tree algorithm

- A glass-box model as we can see the patterns in the data
- The algorithm successively splits the data into groups to maximize the difference of each group
- Repeat process until:
  - All entries in the node are the same class, or
  - All features are used: take a majority vote
- Final tree to often pruned to avoid over-fitting (check ‘prune’ function)



Exercise

- Create a decision tree using data on breast cancer diagnosis
- The dataset contains 699 observations, where 458 (65.5%) are benign and 241 (34.5%) are malignan

  - ‘’Class’’ 2: Benign 4: Malignant

- How good is the model? Train with 75% of the data and test with the remainder

More Practices - Prune a Tree

- Trees tend to be too large and need pruning
- pfit<- prune(fit, cp=fit$cptable[which.min(fit$cptable[, "xerror"]),"CP"])
- # cp is the cost complexity factor
- rpart.plot(pfit,type=4)

More understandings – Tree-based model

- Recursive partitioning to create a tree is a fundamental tool in data mining/machine learning
- Reveals the structure of a set of data
- *It does not evaluate all possible trees *
- Further learning Random Forest