ML Commando Course 2018

Session 2b - Titanic

Russell Moore

ALTA / Computer Laboratory

August 2018

Get the latest notebook

`session_2b_titanic.ipynb`

Titanic - a survival analysis

"I'm the King of the World!"

-Leonardo DiCaprio, Titanic (1997)

For those aboard the *RMS Titanic*, such euphoria was short-lived. On April 1912, four days into her maiden voyage, she hit an iceberg and sank, with the loss of over 1490 of 2224 passengers and crew.

Would you have survived?

Titanic - what we'll do

In this example the tools we'll be looking at are:

- Label Encoding
- One Hot Encoding
- Leave-one-out cross-validation
- Decision Tree Classification
- Random Forests

Titanic - data

The passenger data we'll use is pretty simple:

- Age (float)
- Passenger class (1st, 2nd, 3rd)
- Gender (male, female)
- Target y_i is binary 'survived' flag $\{0,1\}$

Titanic - preprocess the data

We will pre-process the data as follows:

- Age: fill in 'NA' values with mean age of passengers
- Passenger class: convert 1st, 2nd, 3rd to one-hot encoding
  - 1st $\rightarrow$ [1,0,0]
  - 2nd $\rightarrow$ [0,1,0]
  - 3rd $\rightarrow$ [0,0,1]
  - Note this means we have three new columns in our data
- Gender: label encode 'male' and 'female' to binary $\{0,1\}$

Decision Tree Classification

- Decision tree approach: what questions lead us to a particular outcome?
- Divide and conquer:
  - Use data attributes to split the data into subsets
  - If each subset is "pure" (all the same outcome) then stop
  - Else, pick new attribute, and split the subset again
- It's recursive, we perform the same type of splitting operation each time
- The tricky part is picking the right attribute at each stage...

Decision Tree Classification

- How to choose an attribute to split on?
- We want to measure the 'purity' (certainty) of the resulting subsets of outcomes, e.g:
  - (40 lived / 0 died) = completely pure/certain (100%)
  - (20 lived / 20 died) = completely impure/uncertain (50%)
- From Information Theory, we can use *entropy* as a measure of certainty.

Entropy

- Devised by Claude Shannon (1948)
- Clever idea: unlikely events give us more info than commonplace ones.
- For a single outcome s, $Info(s) \stackrel{def}{=} -log(Pr(s))$
- $Info(s) \to \infty$ as $Pr(s) \to 0$
- $Info(s) \to 0$ as $Pr(s) \to 1$
- Also, we can sum info together in an intuitive way.
- For a set of outcomes S, the uncertainty or *entropy* is defined as the expected value of the info contained in S:

$$H(S) \stackrel{def}{=} \mathbb{E}(Info(S))$$

Entropy cont'd

$$H(S) \stackrel{def}{=} \mathbb{E}(Info(S))$$

- To get the expected value for the set, we sum the expected info for each outcome.
- In our *Titanic* case, the outcomes are (L)ived or (D)ied:

$$\begin{aligned} H(S) &= Pr(L)Info(L) + Pr(D)Info(D) \\ &= -Pr(L)log(Pr(L)) + -Pr(D)log(Pr(D)) \end{aligned}$$

- The probabilities are just %age of that outcome in the set. Hence as we'd hope, the purity of each division comes from the proportion of outcomes it contains.
- Greedily choose the split which gives the greatest reduction in uncertainty (aka Gain): $G(S, A) = H(S) - \sum_{a \in A} \frac{|S_a|}{|S|} H(S_a)$

In Brief: Random Forests

An 'ensemble' type classifier

- Select random subsets of data
- Select random subsets of features
- Build many decision trees with these choices
- In classifying new data, each tree may have a different prediction
- Use voting to decide which prediction prevails

In general it's computationally intractable to build an optimal decision tree and greedy algos are not guaranteed to do so. An ensemble looks for agreement between different approaches as a indicator of correctness.

Titanic - Ticket prices

Where would you have been on board?

- First Class (parlour suite) £870/$4,350
- First Class (berth) £30/$150
- Second Class £12/$60
- Third Class £3-£8/$15-$40

£1 in 1912 is approximately equivalent to £108 today!
Use the decision tree or random forest to classify your survival probability...