Deconstruction Data Science

Anubhav Gupta - HW3 - Part 1

Decision Trees - One of the reasons why decision tress are very popular is that they are extremely easy to interpret. Unlike other models like SVM and Neural networks that are black boxes, decision tress function as white boxes allowing the user to see exactly what is happening. Decision trees are considered a relatively “interpretable” model, since they can be post processed in a sequence of decisions

To interpret a decision tree, the analyst starts from the root of the tree and reads through it until a leaf node is reached. For example a rule can be interpreted like this:

- If self-reported location = Berkeley and “benghazi”= false, then y = Democrat

Logistic Regression - Logistic Regression comes in at third place after Naive Bayes and is more difficult to interpret and understand.

$ln[p/(1-p)] = b0 + b1X1 + b2X2 + \ldots + bkXk$ (logistic)

The logistic model is less interpretable. In the logistic model, if b1 is .05, that means that a one-unit increase in X1 is associated with a .05 increase in the log odds that Y is 1. And what does that mean? I’ve never met anyone with any intuition for log odds.

Naive Bayes - Naive Bayes comes in a close second after Decision trees when it comes to users understanding the model.

$p = a0 + a1X1 + a2X2 + \ldots + akXk$ (linear)

In the linear model, if a1 is (say) .05, that means that a one-unit increase in X1 is associated with a 5 percentage point increase in the probability that Y is 1. Just about everyone has some understanding of what it would mean to increase by 5 percentage points their probability of, say, voting, or dying, or becoming obese.

Topic Models - Topic models are the least interpretable when it comes to understanding them.