Intro to Data Analytics and Visualizations

Lecture 33 – Decision Trees; ROC curves and AUC
Fall 2014, November 10th

Outline

1. CART Decision Trees
2. KDD Cup Data
3. ROC curve and AUC
4. Inclass 11_1

Practice Problem

- We are still in the modelling part of the data science process
- We will focus again on supervised learning
- We will focus on classification
- How can we classify a group of customers into buying vs non-buying or ‘staying’ vs ‘leaving’ with the aid of a set of features (collected data)?

Classification and Regression (CART) Decision Trees

- Nodes and branches
- At each node we make a split based on a rule (feature < value)

- The end branch is called a leaf; it ends with the prediction of the probability of the subject to be in the "positive" category (for 0/1 classification)

- Alternative to logistic regression

Note*: Full code and steps explanation for protein example in Dec_trees_Lec33.R*

Potential issues with tree implementation

- Many features
- Many factor features with many levels
- Many missing values

- Should first do wrangling/transformations on the data; feature engineering and selection

CART Decision Trees with KDD Cup 2009 data

- Predict “churn” rate based on information on customers
- Vocabulary from customer relationship management field, a rich field for data science application:

  - “Churn” = customer drops (we will use this as response of interest)

  - “upsell”= customer responds well to marketing pitch

  -”cross-sell” = customer responds well to switching to another product

  -”appetency”= customer has tendency to use new products

Note*: Full code and steps explanation for KDD cup example in Dec_trees_Lec33.R*

Validation and Performance: ROC and AUC

- Using the training, and at least a test set, calculate the typical performance measure;
- ***Another performance measure: Receiver Operating Curve (ROC curve) and Area Under the Curve(AUC)***
- ***AUC should be close to 1 for a good model.***

- **Note*: Full code and steps explanation for KDD cup example in Dec_trees_Lec33.R***

Inclass Assignment 11_3

- Use mtcars R dataset provided by the:

data(mtcars) command;

- Display the first 6 observations (*head(mtcars)*) and the variable names (*names(mtcars)*);
- Describe each variable in the dataset with a few words (data manual); you can find that information by using the command *?mtcars*;
- Classify cars into “automatic” vs “manual” transmission using a CART decision tree and the other variables as features; interpret the tree.
- Plot the tree; what do you see?
- Calculate the AUC (note: we are using the same data for training and calculating AUC).
- Submit your R file IN11_1.R by next Monday.