Intro to Data Analytics and Visualizations

Lecture 34 – k Nearest Neighbors
Fall 2014, November 12th

Outline

1. K Nearest Neighbors (KNN) Algorithm
2. Implementation on KDD Cup data and validation with AUC
3. Comparison to CART trees and Logistic algo for classification
4. Inclass 11_2

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; KNN algorithm can be used for regression as well (scoring algorithm)
- 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)?

kNN

- The kNearest Neighbor (kNN) method makes predictions by locating similar cases to a given data instance (using a similarity function) and returning the average or majority of the most similar data instances.
- K=how many neighbors we consider for classification of each point.
- Similarity is defined by the Euclidian distance (“nearest”)
- Should first scale features, so that each feature carries the same weight in distance calculations

Note*: Full code and steps explanation for protein example in knn_Lect34.R*

Potential issues with knn implementation

- Unbalanced data (very few pos and many neg in training data) leads to poor predictions
- Need to use high k; the more rare the pos, the higher the k needed)
- Expensive (takes time and resources); important for big data
- Better results in this case with logistic

kNN 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 knn_Lec34.R***

Validation and Performance

- Will compare the system time taken by knn, logistic regression and decision trees for classification on KDD data

- ***Will compare the classification performance of decision trees, logistic regression and knn on KDD cup training data using AUC measure***

- **Note*: Full code and steps explanation for KDD cup example in knn_Lect34.R***

Inclass Assignment 11_2

- 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*;
- Train a knn, decision tree and logistic model on this data. Get predictions on the same training data.
- Calculate and compare the AUC and system time for each algorithm. Which one does better? Do you run into any problems?
- Submit your R file IN11_2.R by next Monday.