Lecture 2b
Cleaning and Preprocessing Data

Breitzman 7/2/2018

Why Preprocess

- Garbage in Garbage Out (GIGO)
- Missing fields can crash our models
- When looking for patterns, Outliers and typos can draw attention to themselves
- Data Fields that are co-linear can overestimate or underestimate the importance of some variables
- We’ll talk about strategies for dealing with each of these cases

Data Preparation

- Dorian Pyle – Author of *Data Preparation for Data Mining* estimates that data preparation alone accounts for 60% of all the time and effort expended in the entire data mining process

- 60% sounds like a high estimate, but I can tell you from personal experience that data preparation accounts for at least a 1/3 and can rise to more than 1/2 of some projects that I’ve been involved in

Before we go to R

- We should talk about normalization
- For many of our models we need to first normalize the variables that get input into the models
- The 2 types of normalization I want to talk about tonight are min-max normalization and z-score normalization

Min-Max Normalization

- The advantage of Min-Max normalization is all values live between 0 and 1.
- Ex. Suppose I have the following frequencies:
  2, 17, 3, 102, 179
- Min = 2, Max = 179
- MinMaxNorm = (x – min)/(max – min)
- Normed frequencies become:
  0, 15/177, 1/177, 100/177, 1
- Suppose we have a model with multiple variables. Variable 1 has a range of 100 to 500 and variable 2 has a range of 1 to 10. We normalize so that variable 1 does not dominate the model

Z-Score Normalization

- Z-Score normalization will have values between -4 and 4 and values near the mean will be 0

- Let X be a sequence of values. Z-Score normalization transforms each x in X to x* in X* where
  x* = (x – mean(X))/StdDev(X)

- We’ll see an example in R shortly

Data Preparation in R

- Go to DataPreparationWithR.r
- Make sure we talk about
  - Missing Data
  - Wrong Data
  - Likely Outliers
  - Scatter Plots
  - Bar Charts
  - Z-Score Normalization
  - Min-Max Normalization
  - Compiling reports (html or pdf)