Disease Prediction with Gene Expression Data

Eugine Kang

Sunday, August 30, 2015

Summary  
 Introduction  
 Steps that will follow

setwd("C:/Users/Golden/Desktop/Splunk/gene\_expression\_data")  
genex <- read.table("./genex.txt", header=FALSE, sep=" ")  
data <- read.table("./training.txt",header=TRUE, sep=" ")  
label <- read.table("./labels.txt", header=FALSE, sep=" ")  
dim(data)

## [1] 184 12066

# 1. Data Cleaning

dim(data)

## [1] 184 12066

The data is high-dimensional, meaning the number of variables far exceed the number of samples. Classicial approaches such as least squares linear regression are not appropriate in this setting, and other approaches have the danger of overfitting.

We can reduce the dimension by feature selection and dimension reduction techniques, but let's first see if we can discard variables with little information

require(caret)

## Loading required package: caret

## Warning: package 'caret' was built under R version 3.1.3

## Loading required package: lattice  
## Loading required package: ggplot2

var0 <- nearZeroVar(data)  
length(var0)

## [1] 3245

# Let's remove variables with no variance or near zero variance

require(caret) var0 <- nearZeroVar(data) length(var0) # 3245 features will be removed from data dataZero <- data[,-var0]

# Are there any features with missing values?

noNA <- sapply(dataZero, function(x) sum(is.na(x))) table(noNA) # 6 Columns have missing values # Not many but lets try to fill out the missing values with imputation

# Fill in NA values using the k Nearest Neighbours

require(DMwR) dataImpute <- knnImputation(dataZero, k=10)

# Let's try to remove the effects of Multicollinearity

# Remove features with more than 0.70 of correlation

require(corrplot) dataScale <- scale(dataImpute, center=TRUE, scale=TRUE) corMat <- cor(dataScale) require(caret) highlyCor <- findCorrelation(corMat, 0.70) dataFilter <- as.data.frame(dataScale[,-highlyCor])

# Add Label to our cleaner Data

dataFilterY <- as.factor(dataFilter$Y)