Clustering

Jeff B

April 4, 2019

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE) knitr::opts_chunk$set(echo
= TRUE) library(FNN) library(mvtnorm) library(mclust) library(cluster) library(fpc)
library(boot) library(tree) library(MASS) library(randomForest) library(MLmetrics)
```

```
{r, echo=FALSE} admissionsData <- read.csv("Admission_Predict_Ver1.1.csv") #summary
(admissionsData) attach(admissionsData)
```

Clustering

We begin by computing the respective pairwise distances in our data, and plotting the output.

```
{r, echo=FALSE} dg<-daisy(admissionsData[,-1], metric="gower") pdist <- cmdscale(d=dg)
plot(pdist)
```

We quickly see that two clear groups appear. We can isolate these two groups using hierarchical clustering with single-linkage chaining.

```
{r, , echo=FALSE, eval=FALSE} set.seed(413) km <- kmeans(pdist, centers = 2) #plotcluster(pdist,
km$cluster)
```

```
{r, echo=FALSE} hms <- hclust(na.omit(dg), method="single") #plot(hms) plot(pdist, col=cutree(hms,2))
#plot(pdist)
```

We can then use scatterplots to show the entirety of the data, while still keeping the groups intact, to see if we can determine which predictors most affect these clusters.

```
{r, echo=FALSE} pairs(admissionsData[,-1], col=cutree(hms,2))
```

We notice that, using the single linkage chaining from above, we can predict whether or not a student performs research almost perfectly.

So, by applying Gower's Distance on all predictors and using single-linkage chaining, we have two clear clusters directly coinciding with the presence of a research variable. This tells us that we should use Research as a response variable in models, in addition to Chance of Admit.

We can now perform analyses on the data to attempt to predict a candidate's Chance of Admission, as well as the presence of Research Experience.