Clustering Overview

Finds observations that are similar to one another and groups them together. Unsupervised technique: don't have defined/known groups. Must decide to scale data, choose how to calculate distance between observations, and choose how many groups to have.

Calculating Distance

Consider 3 points on a grid: (1,4), (3,2), (7,8) and calculate the distance between them

```
points = data.frame(x=c(1,3,7), y=c(4,2,8))
dist(points) #defaults to euclidean
```

```
##          1        2
## 2 2.828427
## 3 7.211103 7.211103
```

```
dist(points, method="manhattan")
```

```
##    1  2
## 2  4
## 3 10 10
```

Importance of Scaling

Consider calculating the distance between 3 height (inches) and weight (pounds) pairs.

```
height_weight = data.frame(height = c(65, 68, 60), weight = c(140, 175, 115))
dist(height_weight)
```

```
##          1        2
## 2 35.12834
## 3 25.49510 60.53098
```

Now scale (normalize) the height and weight pairs to be centered at 0 with standard deviation 1 and calculate distance

```
height_weight_scaled = scale(height_weight)
dist(height_weight_scaled)
```

```
##          1        2
## 2 1.378276
## 3 1.489525 2.807431
```

With the unscaled data, observations 1 and 3 are the closest, while with the scaled data observations 1 and 2 are the closest.

Calculating Distance for Categorical Variables

First convert to dummy variables

```
college_gender = data.frame(college = as.factor(c('ASC', 'ASC', 'BUS', 'BUS', 'BUS', 'ENG')), gender = a
#install.packages('dummies')
```

```
library(dummies)
college_gender_dummy = dummy.data.frame(college_gender)
print(college_gender)
```

```
##   college gender
## 1     ASC female
## 2     ASC   male
## 3     BUS   male
## 4     BUS   male
## 5     BUS female
## 6     ENG   male
```

```
print(college_gender_dummy)
```

```
##   collegeASC collegeBUS collegeENG genderfemale gendermale
## 1          1          0          0            1          0
## 2          1          0          0            0          1
## 3          0          1          0            0          1
## 4          0          1          0            0          1
## 5          0          1          0            1          0
## 6          0          0          1            0          1
```

This is just one example of creating dummy variables in R

Then calculate distance

```
dist(college_gender_dummy, method='binary')
```

```
##           1         2         3         4         5
## 2 0.6666667
## 3 1.0000000 0.6666667
## 4 1.0000000 0.6666667 0.0000000
## 5 0.6666667 1.0000000 0.6666667 0.6666667
## 6 1.0000000 0.6666667 0.6666667 0.6666667 1.0000000
```

Clustering Techniques

Hierarchical, K-means, PAM, Principal Component Analysis (PCA)