SI2020_clustering_exercise_v2

PFR

August 2020

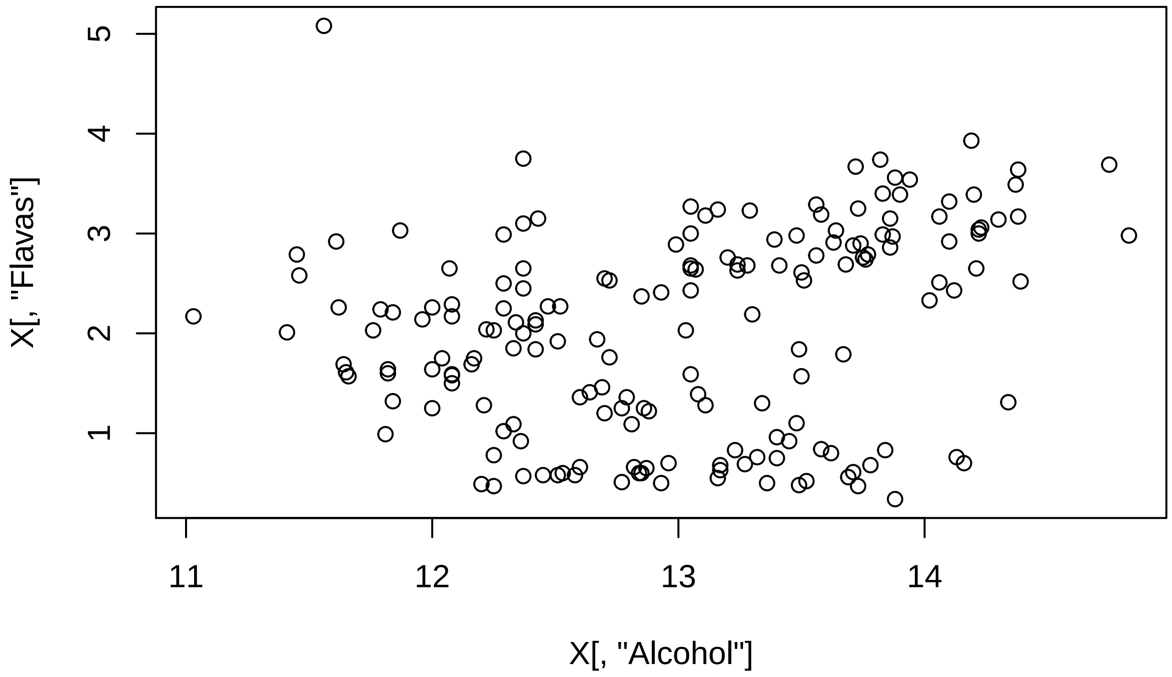
PFR data prep exercises for clustering using UCI Data Repository winedata

---

```
#use setwd("c:/your-path-to-your-project/data/")

X = read.table('winedata.csv',
                         header=TRUE,sep=",",
                         stringsAsFactors = TRUE)   #try TRUE
Y = X[,'Class']
X = subset(X,select=-c(Class))

plot(X[,'Alcohol'],X[,'Flavas'])
```



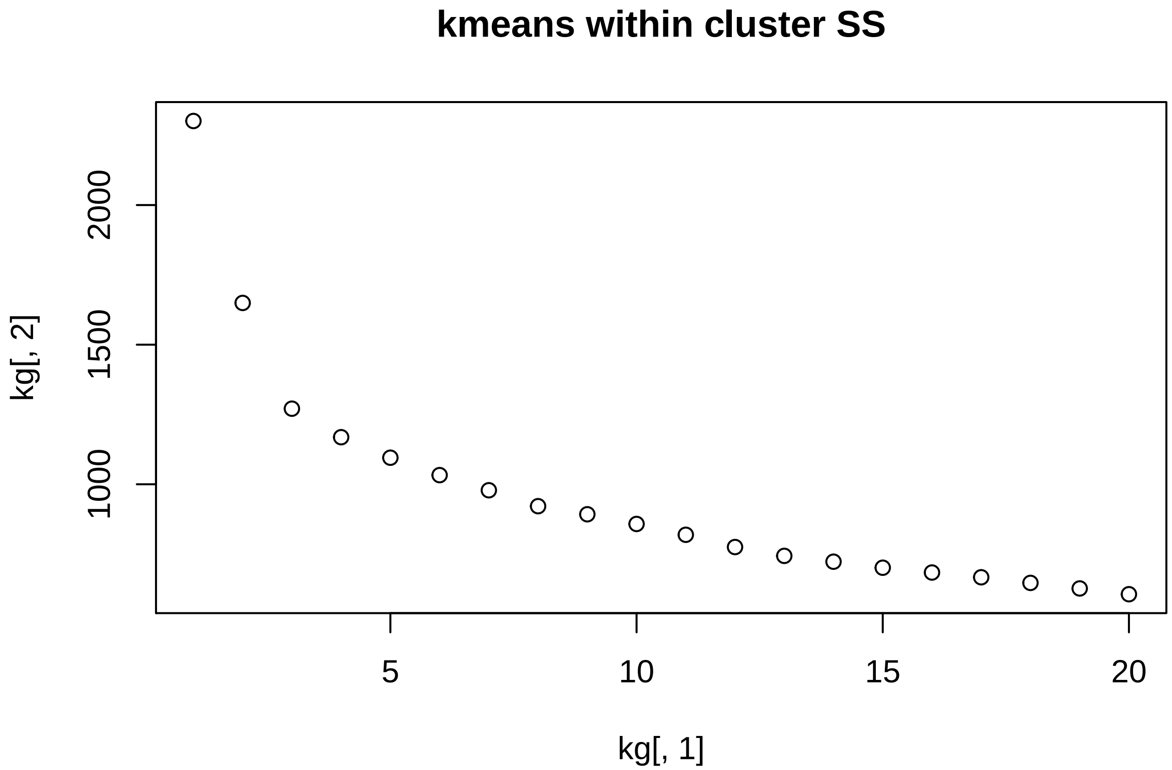
Normalize data

```
X_mncntr = scale(X, center=TRUE, scale= TRUE) #normalize to z-scores
head(X_mncntr,2)
```

```
##     Alcohol      Malic        Ash      Alcal        Mag   Phenols    Flavas
## 1 1.5143408 -0.5606682  0.2313998 -1.166303 1.90852151 0.8067217 1.0319081
## 2 0.2455968 -0.4980086 -0.8256672 -2.483841 0.01809398 0.5670481 0.7315653
##     NonFlavs    ProAnth      Color       Hue       OD   Proline
## 1 -0.6577078  1.2214385  0.2510088 0.3611585 1.842721 1.0101594
## 2 -0.8184106 -0.5431887 -0.2924962 0.4049085 1.110317 0.9625263
```

Run Kmeans for several K values

```
# Run kmeans for 20 values of K
kg=matrix(0,20,2)
for (i in 1:20){
  ktest=kmeans(X_mncntr,i,20,5);
  kg[i,1]=i;
  kg[i,2]=ktest$tot.withinss;  #total withn SSE
  }
plot(kg[,1],kg[,2],main='kmeans within cluster SS')
```



```
#try
#str(ktest)
```

Now, let's plot the cluster mean values

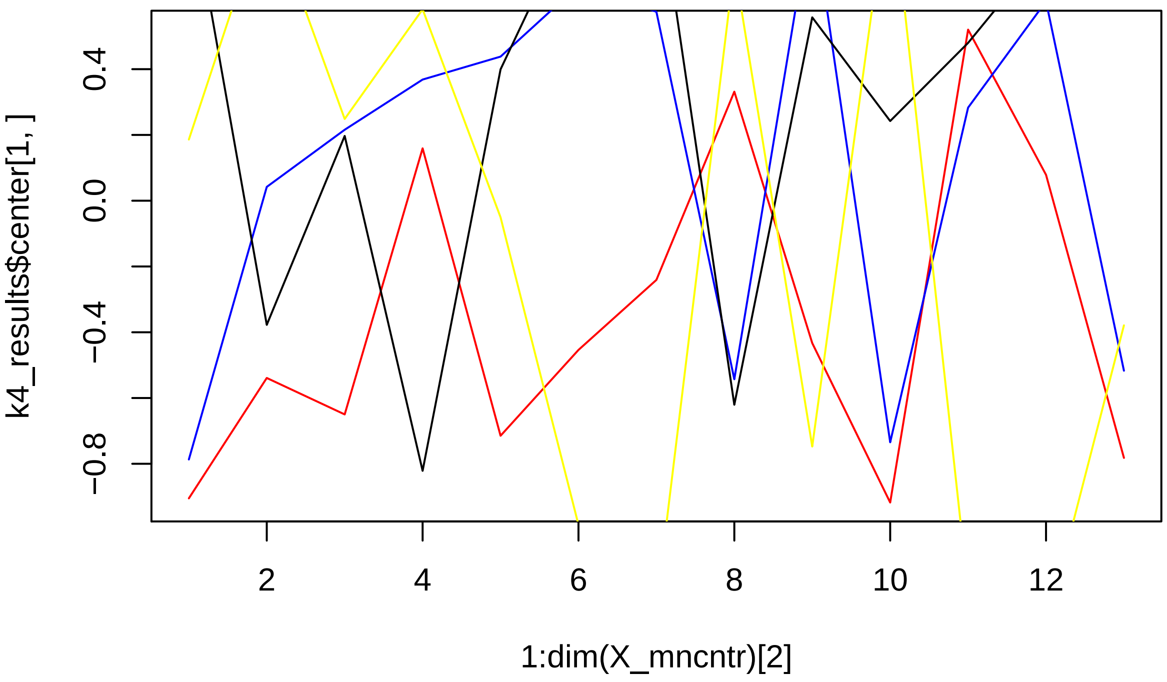
```
#get Kmeans for 4 clusters
k4_results     = kmeans(X_mncntr,4,20,5)

#get color scheme
col2use        = c('red','blue','black','yellow')

#get cluster assignment in colors
colassignments = col2use[k4_results$cluster]

plot(1:dim(X_mncntr)[2],k4_results$center[1,],type='l',col=col2use[1])

for (i in seq(2,4)){
  points(1:dim(X_mncntr)[2],k4_results$center[i,],type='l',col=col2use[i])
  }
```



Now, let's combine the Kmeans cluster information with SVD

Exercise: rerun with different components - what does it say about the data and predictability?

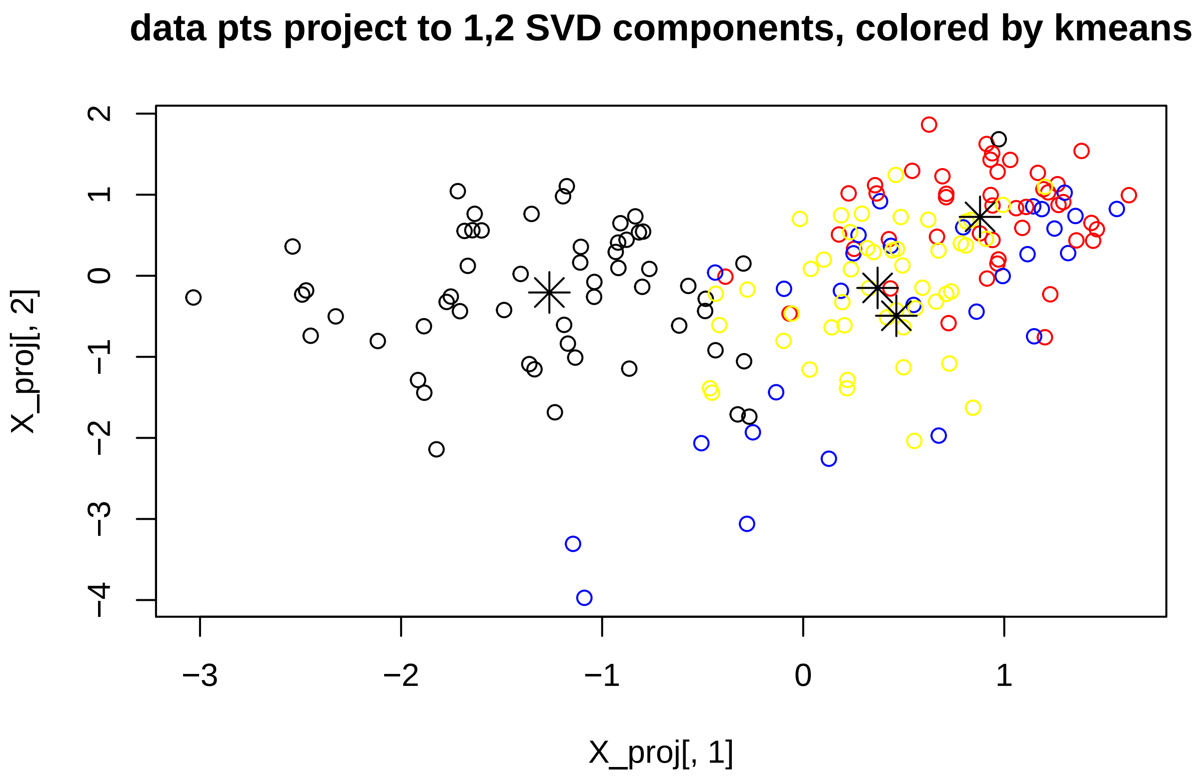
```
#get SVD of X_mncntr
Xsvd = svd(X)

X_proj = as.matrix(X_mncntr) %*% Xsvd$v[,c(1,2)]
```

```
#or just use Xsvd$u[,c(1,2)]
#or use different components beside 1 and 2

plot(X_proj[,1],X_proj[,2],col=colassignments,main='data pts project to 1,2 SVD components, colored by

# to plot center points, first project them into components
c3 = k4_results$centers%*% Xsvd$v[,1:3]
points(c3[,1],c3[,2],pch=8,cex=2)
```



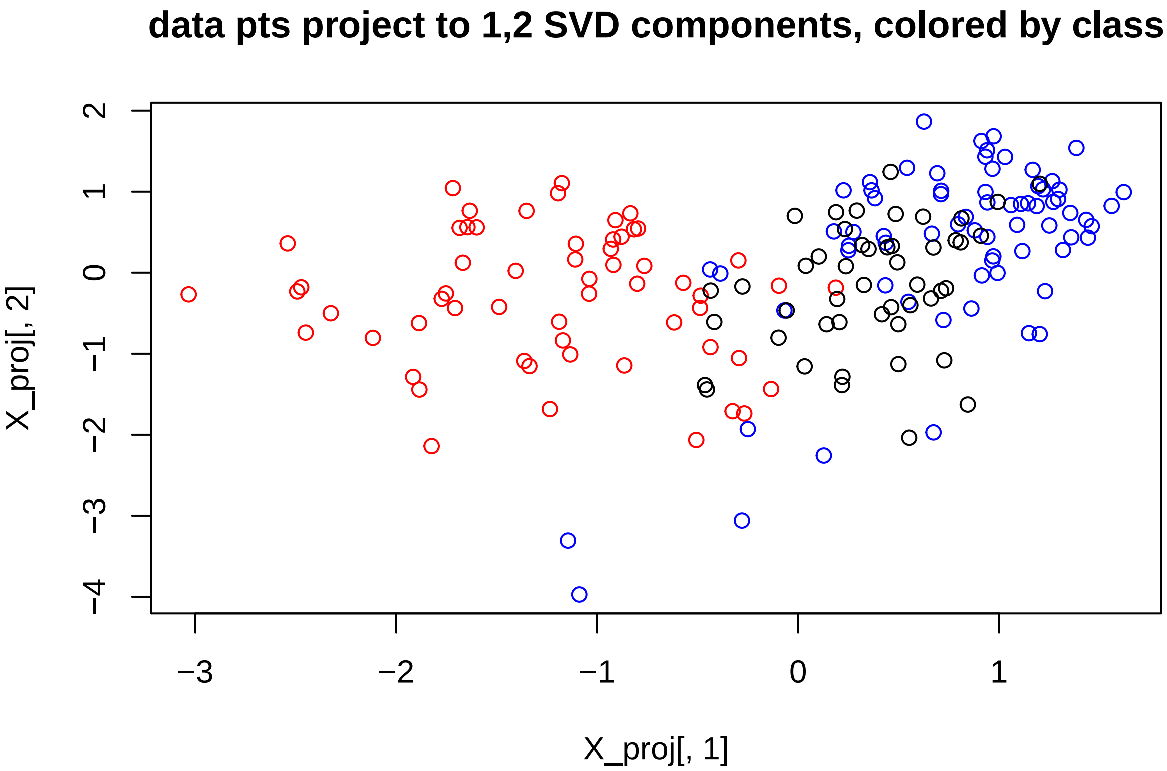
```
#Try different components?
```

Now try coloring by class (using Y mean centered)

```
#Y was created in SVD exercise, use it to select 2 colors

#get class assignment in colors
colassignments = col2use[Y]

plot(X_proj[,1],X_proj[,2],col=colassignments,main='data pts project to 1,2 SVD components, colored by
```



## Extra: Run a classification model, and color the above scatter plots by each class and correct predictions, or each class and incorrect predictions - what do you think you'll see