Machine Learning - 1100-ML0ENG (Ćwiczenia informatyczne Z-23/24)

Home > My courses > Machine Learning - 1100-ML0ENG (Ćwiczenia informatyczne Z-23/24) > Clustering >

Clustering Mixed Data Types

Clustering Mixed Data Types

Dataset College

```
install.packages("ISLR")
library(ISLR)# for college dataset
library(cluster)
library(dplyr)
```

```
coll<-College
```

It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10% of high school class
- Top25perc : New students from top 25% of high school class
- F.Undergrad : Number of full-time undergraduates

- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Let's insert the names of the colleges as a column and create two new columns: acceptance rate is created by diving the number of acceptances by the number of applications and isElite is created by labeling colleges with more than 50% of their new students who were in the top 10% of their high school class as elite.

```
data=coll%>%
 mutate(name= row.names(coll),
 accept_rate = coll$Accept/coll$Apps,
 isElite = cut(coll$Top10perc,
               breaks = c(0, 50, 100),
               labels = c("Not Elite", "Elite")))%>%
select(name, accept_rate, Outstate, Enroll, Grad.Rate, Private, isElite)
```

```
rownames(data)<-1:nrow(coll)
```

```
gower_dist <- daisy(data[, -1],metric = "gower")
summary(gower_dist)
```

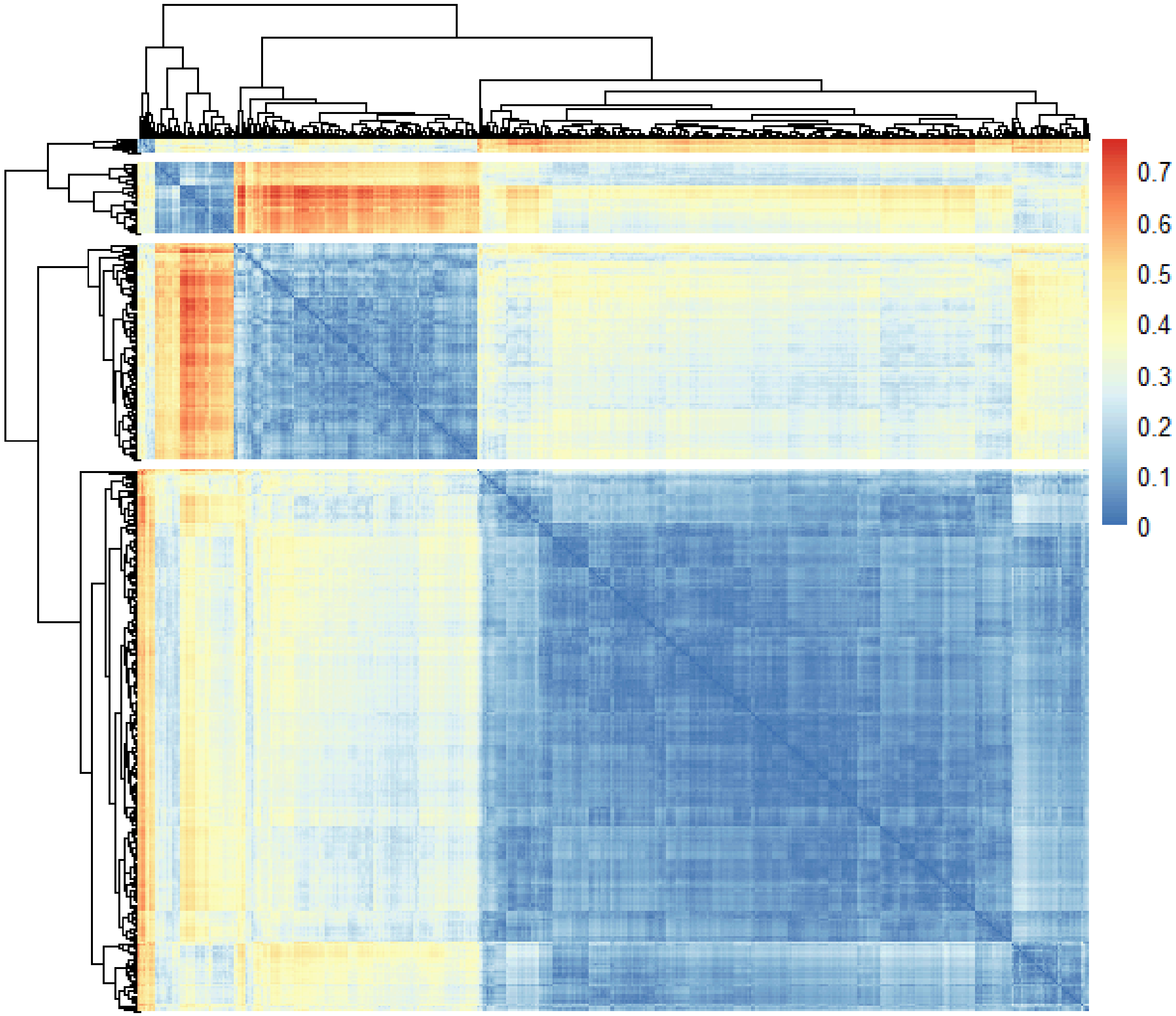
Let's find the most similar and the most dissimilar observations (obviously not required in the analysis :)

```
gower_mat <- as.matrix(gower_dist)
#The most similar
data[ which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind =
TRUE)[1, ], ]
#The most dissimilar
data[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]), arr.ind = TRUE)
[1, ], ]
```

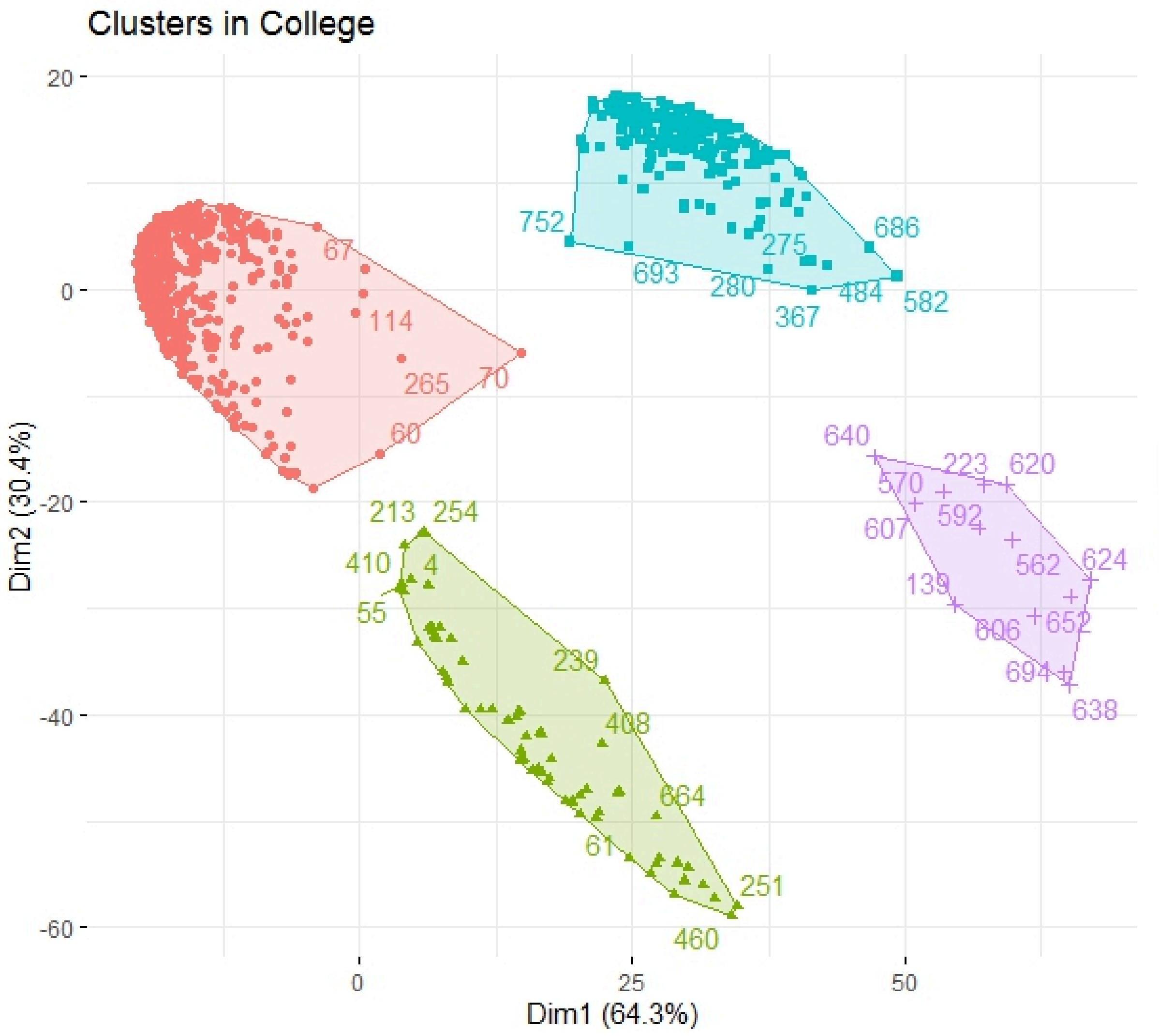
```
tree.coll <- hclust(gower_dist)
```

```
fviz_dend(tree.coll, cex = 0.5, k=4, main = "College tree ")
clust.coll <- cutree(tree.coll, 4)
```

```
pheatmap(gower_dist, clustering_method="complete", cutree_rows = 4)
heatmap(gower_mat, scale = "none")
```



```
fviz_cluster(list(data = gower_mat, cluster = clust.coll),
 ellipse.type = "convex",
 repel = TRUE,
 main = "Clusters in College",
 show.clust.cent = FALSE, ggtheme = theme_minimal())
```



Description of clusters

```
data.cl=cbind(data, clust.coll)
```

```
levels(data.cl$isElite)=c(0,1)
data.cl$isElite=as.numeric(levels(data.cl$isElite))[data.cl$isElite]
levels(data.cl$Private)=c(0,1)
data.cl$Private=as.numeric(levels(data.cl$Private))[data.cl$Private]
data.cl %>%
 group_by(clust.coll) %>%
 summarise(
 number.of.obs= n(),
 average.threshold.of.admission=mean(accept_rate),
 av.number.enroll=mean(Enroll),
 elite.coll=sum(isElite),
 private.coll=sum(Private),
 state.coll=n()-private.coll
 )
```

Cluster Validation

```
clust.coll.eclust<- eclust(gower_mat, "hclust", k = 4, stand = TRUE, nstart = 25,
graph = FALSE)
```

```
fviz_cluster(clust.coll.eclust, gower_mat, geom = "point", repel = FALSE,
 ggtheme = theme_classic())
```

```
fviz_silhouette(clust.coll.eclust, palette = "jco", ggtheme = theme_classic())
# observations in the wrong cluster
silinfo <- clust.coll.eclust$silinfo
silinfo
sil <- clust.coll.eclust$silinfo$widths
neg <- which(sil$sil_width < 0)
sil[neg, , drop = FALSE]
rn<-rownames(sil[neg, , drop = FALSE] )
data[rn,]
```

Last modified: czwartek, 30 listopada 2023, 11:42

Accessibility settings

Przetwarzanie danych osobowych

Platformą administruje Komisja ds. Doskonalenia Dydaktyki wraz z Centrum Informatyki Uniwersytetu Łódzkiego Więcej

Informacje na temat logowania

Na platformie jest wykorzystywana metoda logowania za pośrednictwem Centralnego Systemu Logowania.

Studentów i pracowników Uniwersytetu Łódzkiego obowiązuje nazwa użytkownika i hasło wykorzystywane podczas logowania się do systemu USOSweb.

Deklaracja dostępności