Análise Exploratória

KMeans

Delermando Branquinho Filho

Podemos encontrar coisas que estão próximas?

- Como podemos definir próximo?
- Como agrupamos as coisas?
- Como visualizamos o agrupamento?
- Como interpretamos o agrupamento?

Como definimos close?

- Etapa mais importante
- Lixo em $ longrightarrow $ lixo para fora
- Distância ou similaridade
- Distância contínua - euclidiana
- Semelhança de correlação contínua
- Binário - distância manhattan
- Escolha uma distância / semelhança que faz sentido para o seu problema

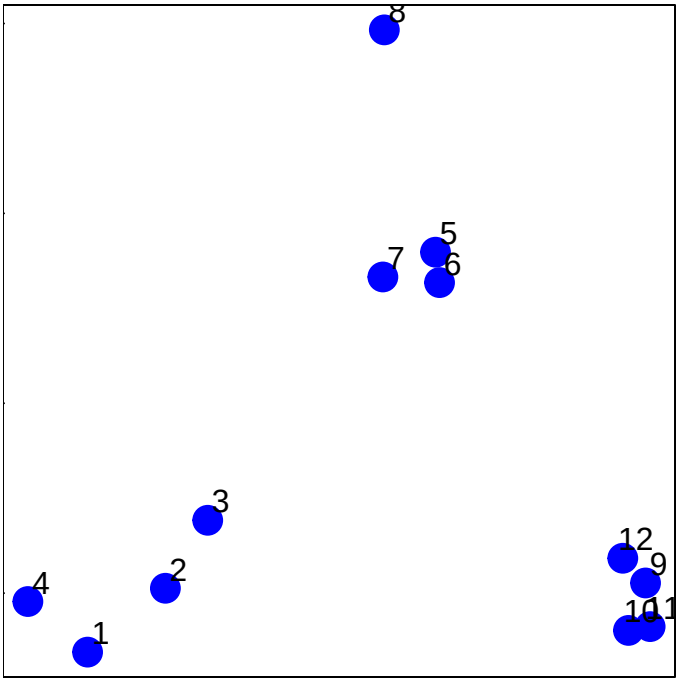
---

K-significa agrupamento

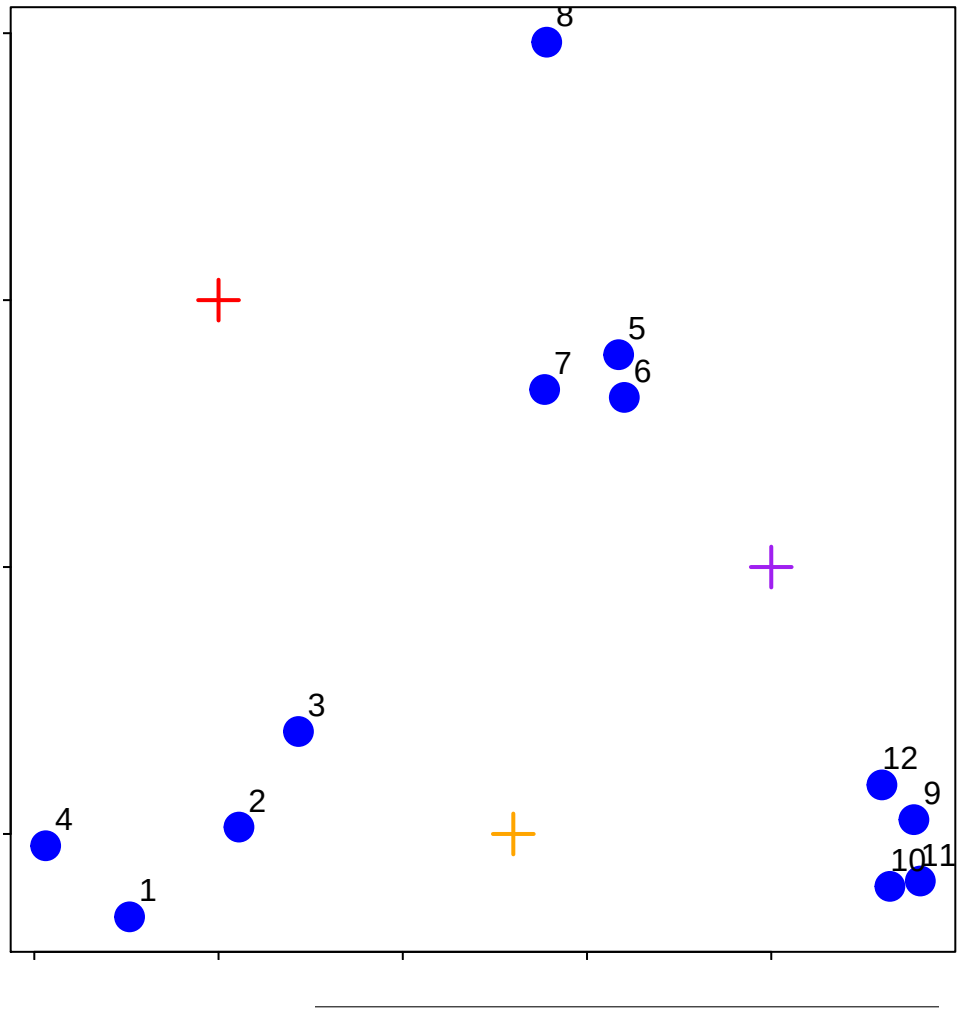
- Uma abordagem de partilha
- Corrigir uma série de clusters
- Obter “centroids” de cada cluster
- Atribuir coisas ao centróide mais próximo
- Reclacular centróides
- Requer
- Uma métrica de distância definida
- Uma série de clusters
- Uma adivinhação inicial para centróides de cluster
- Produz
- Estimativa final de centróides de cluster
- Uma atribuição de cada ponto a clusters

K-means clustering - exemplo

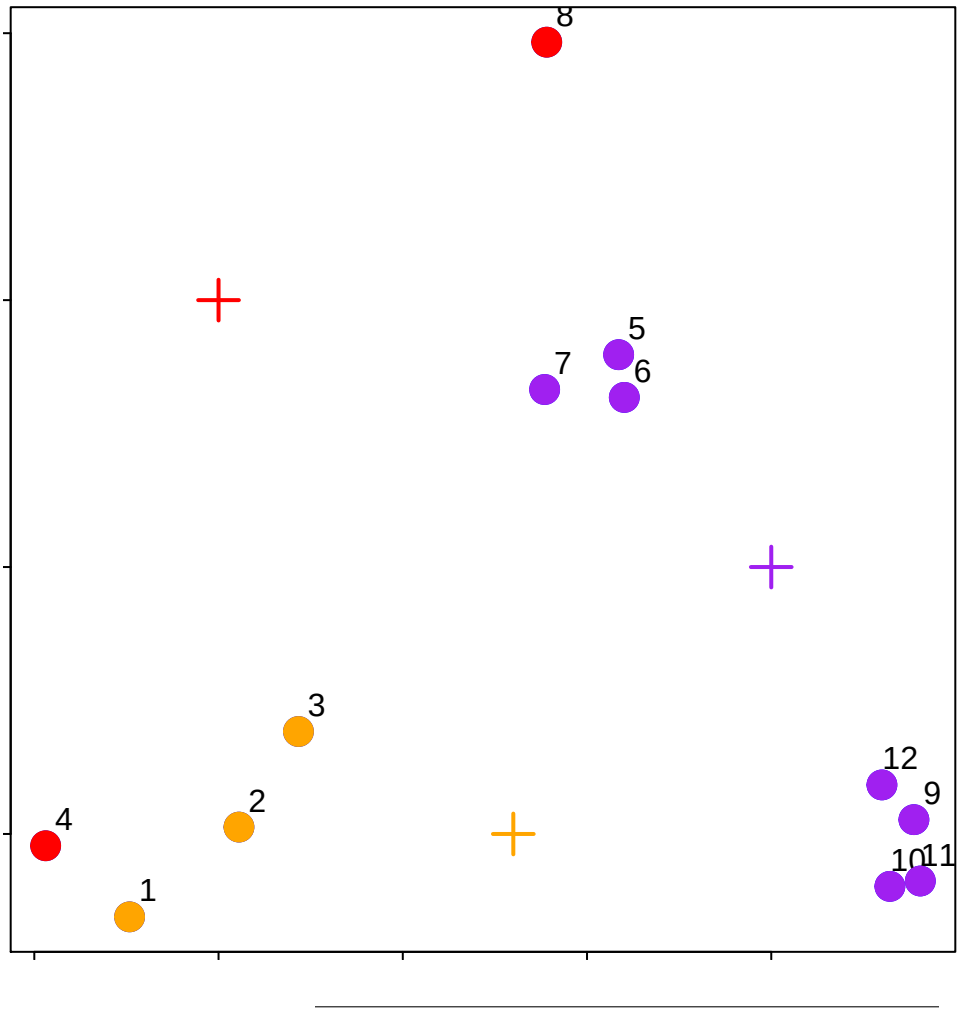
```
set.seed(1234); par(mar=c(0,0,0,0))
x <- rnorm(12,mean=rep(1:3,each=4),sd=0.2)
y <- rnorm(12,mean=rep(c(1,2,1),each=4),sd=0.2)
plot(x,y,col="blue",pch=19,cex=2)
text(x+0.05,y+0.05,labels=as.character(1:12))
```


8
5
7
6
3
12
9
2
4
10
11
1

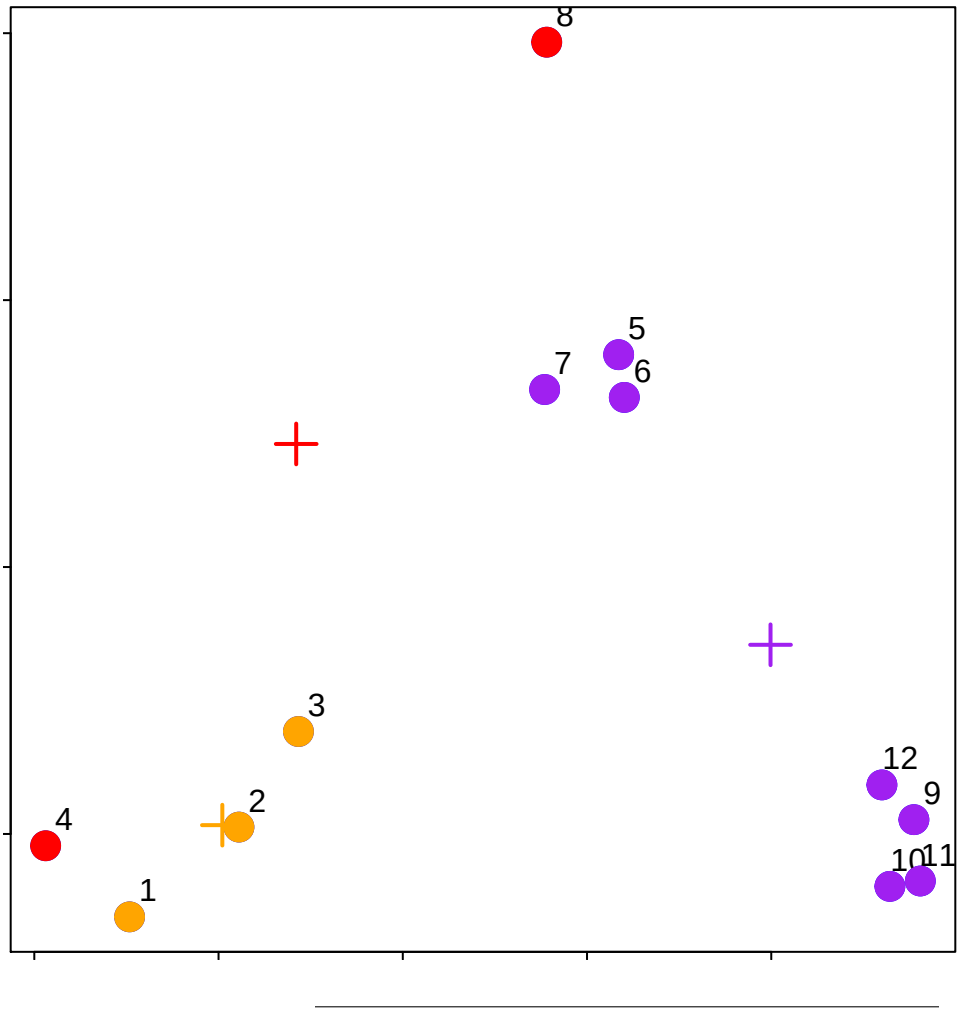
K-means clustering - Iniciando centróides



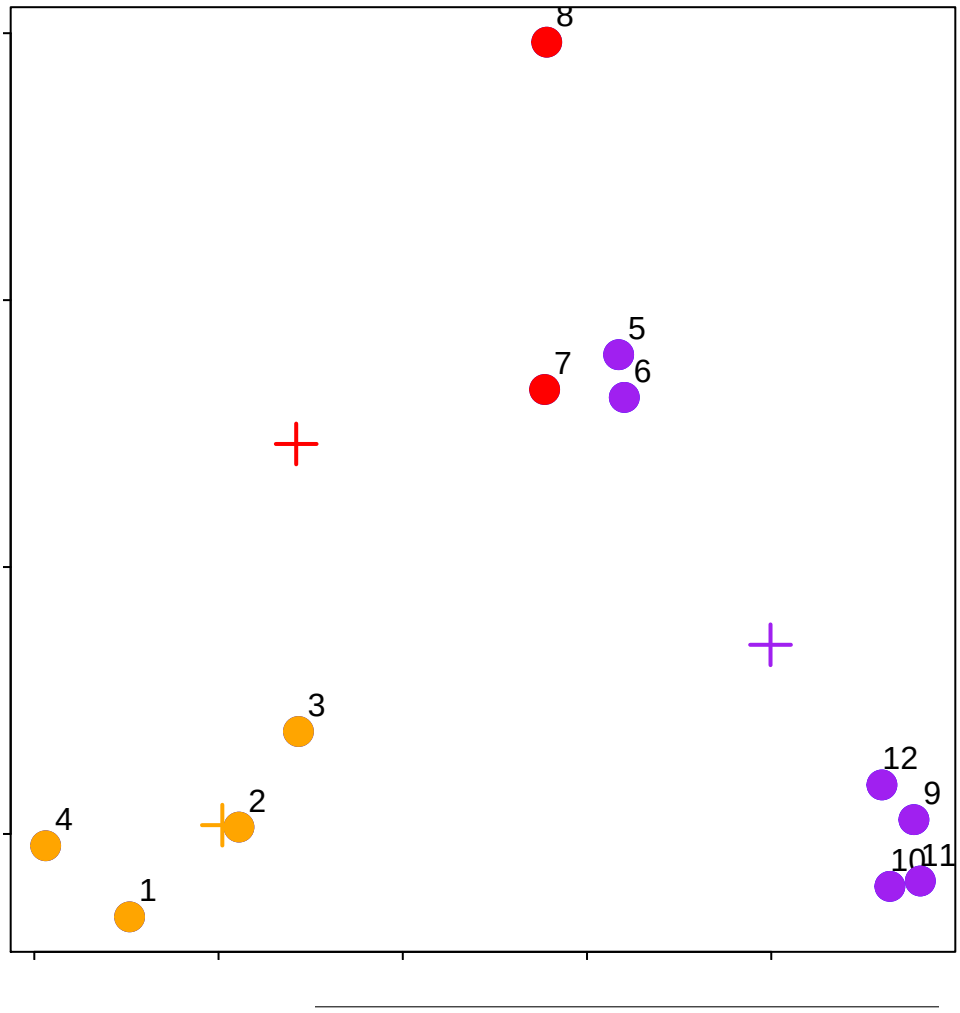
K-means clustering - Atribuir ao centróide mais próximo



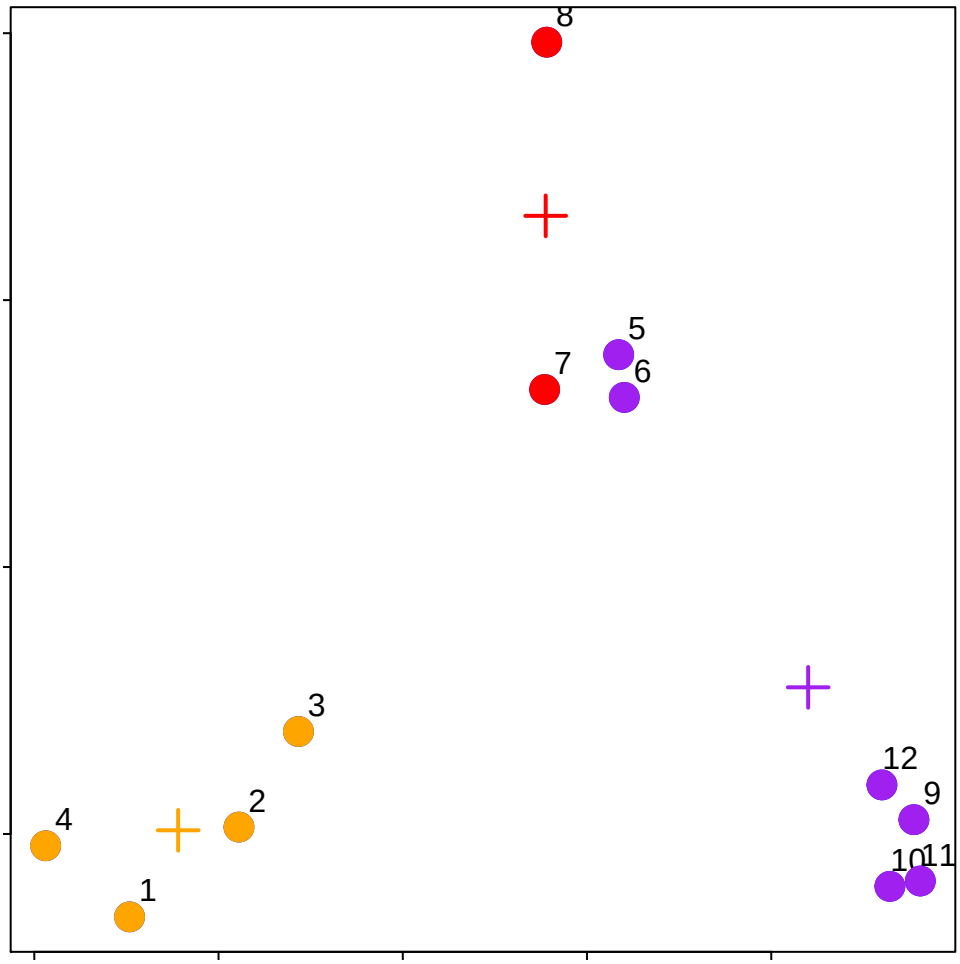
K-means clustering - Recalcular centróides



K-means clustering - Reatribuir valores



K-means clustering - atualização dos Centróides



`kmeans()`

- Important parameters: *x*, *centers*, *iter.max*, *nstart*

```
dataFrame <- data.frame(x,y)
kmeansObj <- kmeans(dataFrame,centers=4)
names(kmeansObj)
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

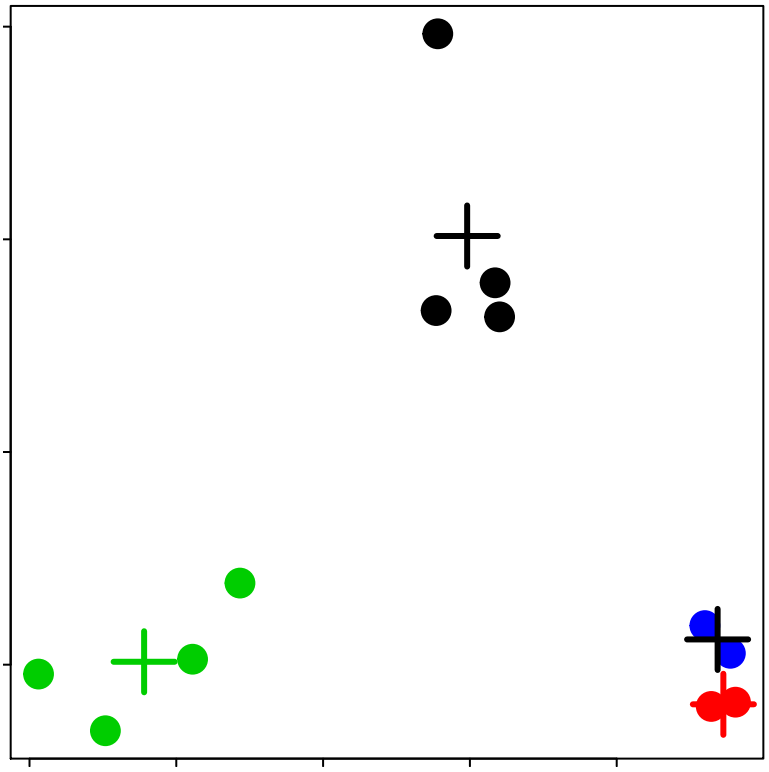
```
kmeansObj$cluster
```

```
##  [1] 3 3 3 3 1 1 1 1 4 2 2 4
```

---

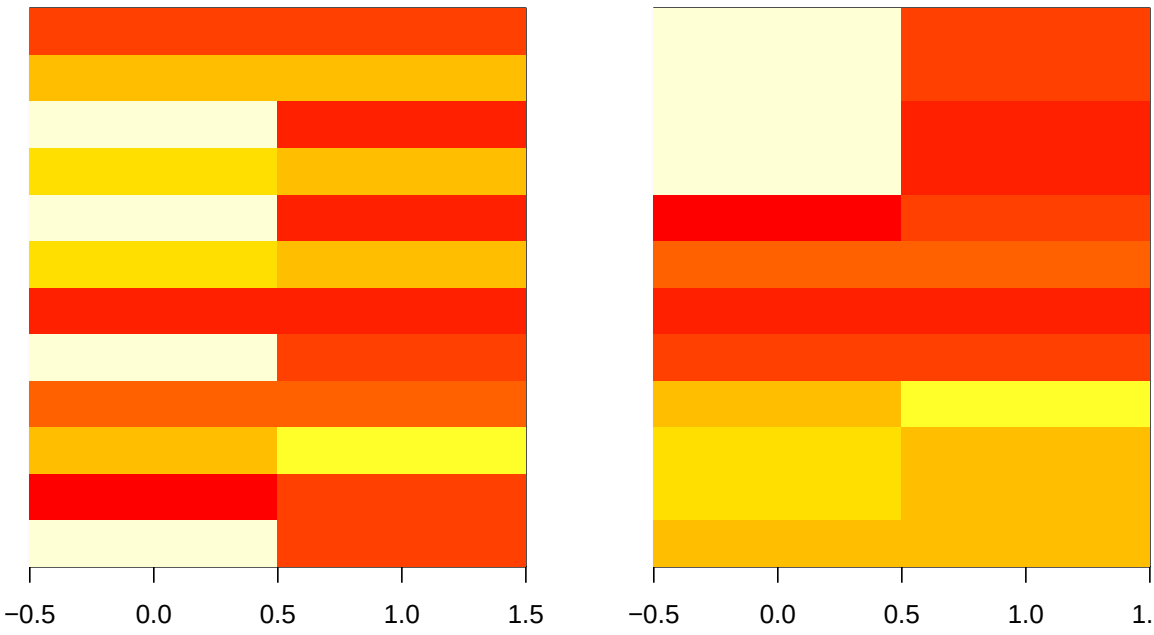
kmeans()

```
par(mar=rep(0.2,4))
plot(x,y,col=kmeansObj$cluster,pch=19,cex=2)
points(kmeansObj$centers,col=1:3,pch=3,cex=3,lwd=3)
```



Heatmaps

```
set.seed(1234)
dataMatrix <- as.matrix(dataFrame)[sample(1:12),]
kmeansObj <- kmeans(dataMatrix,centers=3)
par(mfrow=c(1,2), mar = c(2, 4, 0.1, 0.1))
# Um caractere que especifica o tipo de eixo y. Especificar "n" suprime o traçado.
image(t(dataMatrix)[,nrow(dataMatrix):1],yaxt="n")
image(t(dataMatrix)[,order(kmeansObj$cluster)],yaxt="n")
```


−0.5
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0
1.