Aprendizaje No Supervisado: Clustering

Juan F. Pérez

Departamento MACC
Matemáticas Aplicadas y Ciencias de la Computación
Universidad del Rosario

juanferna.perez@urosario.edu.co

2019

Contenidos

1. Introducción
2. Clustering/Análisis de Conglomerados
3. K-means
4. K-Means en R
5. Clustering Jerárquico
6. Clustering con datos categóricos

Introducción

Introducción

- Búsqueda de patrones
- Estudio de fenómenos físicos
- Reconocimiento de patrones
- Descubrimiento automático de regularidades
- Algoritmos computacionales

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Aprendizaje supervisado

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Aprendizaje supervisado

- Resultado es una o varias variables continuas (no un número finito de categorías)

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Aprendizaje supervisado

- Resultado es una o varias variables continuas (no un número finito de categorías)

Regresión

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Aprendizaje supervisado

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Aprendizaje supervisado

- Resultado es una categoría (de un número finito de posibles categorías)

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i como las categorías/etiquetas t_i

Aprendizaje supervisado

- Resultado es una categoría (de un número finito de posibles categorías)

Clasificación

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Aprendizaje no supervisado

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Aprendizaje no supervisado

- Objetivo es descubrir grupos similares

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Aprendizaje no supervisado

- Objetivo es descubrir grupos similares

Clustering (análisis de conglomerados)

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Aprendizaje no supervisado

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Aprendizaje no supervisado

- Objetivo es determinar la distribución de los datos en el espacio de entrada

Algunos problemas de aprendizaje

- Datos de entrenamiento contienen las características x_i pero NO las categorías/etiquetas t_i

Aprendizaje no supervisado

- Objetivo es determinar la distribución de los datos en el espacio de entrada

Estimación de densidades

Clustering/Análisis de Conglomerados

Clustering/Análisis de Conglomerados

Objetivo:

- Dadas n observaciones cada una descrita como un vector x de dimensión D (características)

Clustering/Análisis de Conglomerados

Objetivo:

- Dadas n observaciones cada una descrita como un vector x de dimensión D (características)
- Descubrir grupos de observaciones similares

Clustering/Análisis de Conglomerados

Objetivo:

- Dadas n observaciones cada una descrita como un vector x de dimensión D (características)
- Descubrir grupos de observaciones similares
- Grupo = cluster = conglomerado

Clustering/Análisis de Conglomerados

Objetivo:

- Dadas n observaciones cada una descrita como un vector x de dimensión D (características)
- Descubrir grupos de observaciones similares
- Grupo = cluster = conglomerado
- **NO** hay etiquetas $\rightarrow$ no sabemos si efectivamente hay clusters o cuántos hay

Clustering/Análisis de Conglomerados

Objetivo:

- Dadas n observaciones cada una descrita como un vector x de dimensión D (características)
- Descubrir grupos de observaciones similares
- Grupo = cluster = conglomerado
- **NO** hay etiquetas $\rightarrow$ no sabemos si efectivamente hay clusters o cuántos hay
- $\rightarrow$ *diferente* al problema de clasificación

Ejemplos

- Se toman muestras de un tejido canceroso de n pacientes

Ejemplos

- Se toman muestras de un tejido canceroso de n pacientes
- Para cada muestra se obtienen D descriptores (características): medidas físicas, químicas, imágenes

Ejemplos

- Se toman muestras de un tejido canceroso de n pacientes
- Para cada muestra se obtienen D descriptores (características): medidas físicas, químicas, imágenes
- Se busca identificar casos/muestras similares

Ejemplos

- Se toman muestras de un tejido canceroso de n pacientes
- Para cada muestra se obtienen D descriptores (características): medidas físicas, químicas, imágenes
- Se busca identificar casos/muestras similares
- Podrían reflejar estados similares de avance de la enfermedad, respuesta similar a tratamiento, tipos diferentes de enfermedad/paciente

Ejemplos

- Se toman muestras de un tejido canceroso de n pacientes
- Para cada muestra se obtienen D descriptores (características): medidas físicas, químicas, imágenes
- Se busca identificar casos/muestras similares
- Podrían reflejar estados similares de avance de la enfermedad, respuesta similar a tratamiento, tipos diferentes de enfermedad/paciente
- No se sabe *a priori* pero se quiere explorar

Ejemplos

- Se tiene información de n clientes

Ejemplos

- Se tiene información de n clientes
- Para cada cliente se obtienen D descriptores (características): hábitos de compra, datos socio-demográficos

Ejemplos

- Se tiene información de n clientes
- Para cada cliente se obtienen D descriptores (características): hábitos de compra, datos socio-demográficos
- Se busca identificar clientes similares

Ejemplos

- Se tiene información de n clientes
- Para cada cliente se obtienen D descriptores (características): hábitos de compra, datos socio-demográficos
- Se busca identificar clientes similares
- Podrían reflejar potenciales clientes de nuevos productos, interés en ofertas de cierto tipo, capacidad/deseo de compra de ciertos artículos

Ejemplos

- Se tiene información de n clientes
- Para cada cliente se obtienen D descriptores (características): hábitos de compra, datos socio-demográficos
- Se busca identificar clientes similares
- Podrían reflejar potenciales clientes de nuevos productos, interés en ofertas de cierto tipo, capacidad/deseo de compra de ciertos artículos
- No se sabe *a priori* pero se quiere explorar

Ejemplos

- Se tiene información de n clientes
- Para cada cliente se obtienen D descriptores (características): hábitos de compra, datos socio-demográficos
- Se busca identificar clientes similares
- Podrían reflejar potenciales clientes de nuevos productos, interés en ofertas de cierto tipo, capacidad/deseo de compra de ciertos artículos
- No se sabe *a priori* pero se quiere explorar
- (*Segmentación del mercado*)

Resultado

- Clusters de observaciones similares

Resultado

- Clusters de observaciones similares
- Cada cluster puede reflejar un conjunto de interés a analizar

Resultado

- Clusters de observaciones similares
- Cada cluster puede reflejar un conjunto de interés a analizar
- Reducción o simplificación de información

Resultado

- Clusters de observaciones similares
- Cada cluster puede reflejar un conjunto de interés a analizar
- Reducción o simplificación de información
- Simplificar o posibilitar el análisis de grandes cantidades de información multi-dimensional

K-means

K-means

- Agrupar observaciones en K clusters

K-means

- Agrupar observaciones en K clusters
- Para cada observación se determina a cuál de los clusters pertenece (solo uno)

K-means

- Agrupar observaciones en K clusters
- Para cada observación se determina a cuál de los clusters pertenece (solo uno)
- Clusters $C_1, \ldots, C_K$

K-means

- Agrupar observaciones en K clusters
- Para cada observación se determina a cuál de los clusters pertenece (solo uno)
- Clusters $C_1, \ldots, C_K$
- Toda observación pertenece a un solo cluster

K-means

- Visión 1:

K-means

- Visión 1:
  - Observaciones en un mismo cluster deben ser parecidas entre sí

K-means

- Visión 1:
  - Observaciones en un mismo cluster deben ser parecidas entre sí
  - Observaciones en clusters diferentes deben ser relativamente diferentes

K-means

- Visión 1:
  - Observaciones en un mismo cluster deben ser parecidas entre sí
  - Observaciones en clusters diferentes deben ser relativamente diferentes
- Visión 2:

K-means

- Visión 1:
  - Observaciones en un mismo cluster deben ser parecidas entre sí
  - Observaciones en clusters diferentes deben ser relativamente diferentes
- Visión 2:
  - Una observación debe ser más parecida a otras observaciones en el mismo cluster que a observaciones en otros clusters

K-means

- Visión 1:
  - Observaciones en un mismo cluster deben ser parecidas entre sí
  - Observaciones en clusters diferentes deben ser relativamente diferentes
- Visión 2:
  - Una observación debe ser más parecida a otras observaciones en el mismo cluster que a observaciones en otros clusters
- Visión 3:

K-means

- Visión 1:
  - Observaciones en un mismo cluster deben ser parecidas entre sí
  - Observaciones en clusters diferentes deben ser relativamente diferentes
- Visión 2:
  - Una observación debe ser más parecida a otras observaciones en el mismo cluster que a observaciones en otros clusters
- Visión 3:
  - Minimizar la variabilidad al interior de cada cluster conformado

K-means

- Minimizar la variabilidad al interior de cada cluster conformado

K-means

- Minimizar la variabilidad al interior de cada cluster conformado
- Sea V_k una medida de la variabilidad en el cluster k

K-means

- Minimizar la variabilidad al interior de cada cluster conformado
- Sea V_k una medida de la variabilidad en el cluster k
- Objetivo de K-means:

$$\text{mín} \sum_{k=1}^{K} V_k$$

K-means

- Determinar V_k una medidad de la variabilidad en el cluster k

K-means

- Determinar V_k una medidad de la variabilidad en el cluster k
- Distancia (euclideana) entre todos los puntos del cluster

K-means

- Determinar V_k una medidad de la variabilidad en el cluster k
- Distancia (euclideana) entre todos los puntos del cluster
- Suponiendo una sola característica (x_i es un número)

$$V_k = \frac{1}{|C_k|} \sum_{i,j \in C_k} (x_i - x_j)^2$$

K-means

- Determinar V_k una medidad de la variabilidad en el cluster k
- Distancia (euclideana) entre todos los puntos del cluster
- Suponiendo una sola característica (x_i es un número)

$$V_k = \frac{1}{|C_k|} \sum_{i,j \in C_k} (x_i - x_j)^2$$

- Caso general con D características (x_i es un vector con entradas $x_{i,d}$):

$$V_k = \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2$$

K-means

- Objetivo de K-means:

$$\text{mín} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2$$

K-means

- Objetivo de K-means:

$$\text{mín} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2$$

- Problema de optimización

K-means

- Objetivo de K-means:

$$\text{mín} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2$$

- Problema de optimización
- K^n formas de asignar n observaciones a K clusters

K-means

- Objetivo de K-means:

$$\text{mín} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2$$

- Problema de optimización
- K^n formas de asignar n observaciones a K clusters
- ¿Cómo resolverlo?

K-means: Algoritmo

- Algoritmo sencillo que encuentra una muy buena solución

K-means: Algoritmo

- Algoritmo sencillo que encuentra una muy buena solución
- Algoritmo eficiente: ejecución veloz y escalable a muchos datos

K-means: Algoritmo

- Algoritmo sencillo que encuentra una muy buena solución
- Algoritmo eficiente: ejecución veloz y escalable a muchos datos
- No garantiza que se encuentre la mejor solución (óptimo)

K-means: Algoritmo

- Algoritmo sencillo que encuentra una muy buena solución
- Algoritmo eficiente: ejecución veloz y escalable a muchos datos
- No garantiza que se encuentre la mejor solución (óptimo)
- El centroide y_k del cluster k, :

$$y_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

K-means: Algoritmo

- Algoritmo sencillo que encuentra una muy buena solución
- Algoritmo eficiente: ejecución veloz y escalable a muchos datos
- No garantiza que se encuentre la mejor solución (óptimo)
- El centroide y_k del cluster k, :

$$y_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

- Centroide y_k: punto medio del cluster k

K-means: Algoritmo

1. Seleccione K centroides al azar, uno para cada cluster
2. Hasta que los centroides no cambien más, *repita*:
   1) Asigne cada observación al cluster más cercano
   2) Para cada cluster, re-calcule el centroide

K-Means en R

K-Means en R

```
set.seed (2)
x=matrix (rnorm (50*2) , ncol =2)
x[1:25 ,1]=x[1:25 ,1]+3
x[1:25 ,2]=x[1:25 ,2] -4
```

K-Means en R

```
km =kmeans (x,2, nstart =20)
km$cluster

plot(x, col =(km$cluster +1) ,
 main="Clustering con K-Means - K=2",
 xlab ="", ylab="", pch =20, cex =2)
```

K-Means en R

```
set.seed (4)
km =kmeans (x,3, nstart =20)
km
km$cluster

plot(x, col =(km$cluster +1) , main="Clustering con K-
 - K=3", xlab ="", ylab="", pch =20, cex =2)
```

K-Means en R

```
set.seed (3)
km =kmeans (x,3, nstart =1)
km$tot.withinss
```

Clustering Jerárquico

Clustering Jerárquico

Dendogramas

- Árboles
- Hojas: observaciones
- Bottom-up: se van fusionando ramas hasta llegar a la raiz
- En cada paso se fusionan las dos ramas más parecidas
- Similitud: distancia entre ramas/grupos
- Si las ramas/grupos están compuestas de una observación: distancia entre dos puntos

Clustering Jerárquico

- ¿Si las ramas/grupos contienen múltiples observaciones?
- Linkage
- Complete: máxima distancia entre los puntos de los grupos
- Single: mínima distancia entre los puntos de los grupos
- Average: distancia promedio entre los puntos de los grupos
- Centroid: distancia entre los centroide de los grupos

Clustering Jerárquico

```
hc.complete =
  hclust (dist(x), method ="complete")
hc.single =
  hclust (dist(x), method ="single")
hc.average =
  hclust (dist(x), method ="average")
```

Clustering Jerárquico

Dendogramas

```
hc.complete =
  hclust (dist(x), method ="complete")
hc.single =
  hclust (dist(x), method ="single")
hc.average =
  hclust (dist(x), method ="average")
```

Clustering Jerárquico

```
par(mfrow =c(1,3))
plot(hc.complete,
   main ="Complete Linkage", xlab="", sub ="",
   cex =.9)
plot(hc.average,
   main ="Average Linkage", xlab="", sub ="",
   cex =.9)
plot(hc.single ,
   main="Single Linkage", xlab="", sub ="",
   cex =.9)
```

Clustering Jerárquico

```
cutree(hc.complete, 2)
cutree(hc.average, 2)
cutree(hc.single, 2)
cutree(hc.single, 4)
```

Clustering Jerárquico

¿Distancia?

- Euclidiana
- Correlación
- Alerta: datos categóricos

Clustering Jerárquico

¿Escala?

- Escala original
- Escala estadarizada (desviación estándar $= 1$)
- Impacto en agrupamiento (jerárquico o k-means)

Clustering Jerárquico con características estandarizadas

```
y <- USArrests
summary(y)
apply(y, 2, sd)

ysc=scale (y)
summary(ysc)
apply(ysc, 2, sd)

ysc_clust = hclust (
   dist(ysc), method ="complete")

plot(ysc_clust ,
   main ="Clustering jerárquico")
```

Clustering con datos categóricos

Clustering con datos categóricos

- No es posible usar la misma medida de distancia
- Similitud: si son iguales 0, de lo contrario 1
- k-modes
- k-prototypes

Clustering con datos categóricos

```
?mtcars
x <- mtcars
names(x)
class(x)
sapply(x, typeof)
str(x)

x$vs <- factor(x$vs)
levels(x$vs)

x$am <- factor(x$am)
levels(x$am)
```

Clustering con datos categóricos

```
install.packages("klaR")
library(klaR)
km =kmodes(x[,c("vs","am")],2)
names(km)

km$cluster
km$withindiff
km$modes

x[km$cluster==2,c("vs","am")]
```

Clustering con datos categóricos

```
install.packages("clustMixType")
library(clustMixType)
km =kproto(x,2)
names(km)

km$centers
km$withinss
km$tot.withinss
km$size
```

Clustering con datos categóricos

```
pairs(x, col =(km$cluster +1),
  pch =20, cex =2)
```