Clustering

Modelos estadísticos para la regresión y la clasificación

27 de Mayo, 2021

Matías Carrasco

Instituto de Matemática y Estadística – Facultad de Ingeniería – UdelaR

Índice

1/ *k*-Medias

Introducción

Breve descripción

K-medias (MacQueen 1967) es uno de los algoritmos de aprendizaje automático (no supervisado) más utilizados para dividir un conjunto de datos determinado en K grupos (es decir, K clusters), en donde K representa el número de grupos pre-especificados por el analista.

Clasifica a los objetos en múltiples grupos (o clusters), de modo que los objetos dentro del mismo grupo sean tan similares como sea posible (es decir, alta homogeneidad intraclase), mientras que los objetos de diferentes grupos sean lo más diferentes posible.

En el agrupamiento de K-medias, cada grupo está representado por su centro que corresponde a la media de los puntos asignados al grupo.

Introducción

Vagamente: la idea principal

La idea básica detrás de K-medias consiste en definir grupos de modo que se minimice la variación total dentro de cada grupo.

Concretamente

Hay varios algoritmos de K-medias disponibles. El algoritmo estándar es el algoritmo de Hartigan-Wong (1979), que define la variación total dentro del cluster como la suma de las distancias al cuadrado (las distancias euclídeas) entre las observaciones y el centro correspondiente:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

en donde x_i denota una observación perteneciente al cluster C_k y μ_k es el promedio de los puntos asignados al cluster C_k.

Introducción

Concretamente

Cada observación x_i se asigna a un grupo dado de modo que la suma de cuadrados de la distancia de la observación a su centro asignado es un mínimo.

Definimos la variación total de la siguiente manera:

$$\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

La suma de distancias al cuadrado dentro del grupo mide la compacidad del agrupamiento y queremos que sea lo más pequeño posible (en promedio).

El algoritmo

- El primer paso de K-medias es indicar el número de clústers (K) que se generarán en la solución final.
- El algoritmo comienza seleccionando aleatoriamente K observaciones del conjunto de datos para que sirvan como centros iniciales para los grupos.
- A continuación, cada una de las observaciones restantes se asigna a su centro más cercano, donde más cercano se define utilizando la distancia euclídea entre la observación y la media del grupo. Este paso se denomina “paso de asignación de clúster”.
- Después del paso de asignación, el algoritmo calcula el nuevo promedio de cada grupo. Se utiliza el nombre “actualización del centro” para este paso.
- Ahora que se han recalculado los centros, se vuelve a comprobar cada observación para ver si podría estar más cerca de un centro nuevo. Todas las observaciones se reasignan nuevamente utilizando los centros de clúster actualizados.
- Los pasos de actualización y asignación se repiten iterativamente hasta que las asignaciones dejan de cambiar (es decir, hasta que se logra la convergencia).

El algoritmo en resumen

1. Especificar el número de clústeres (K).
2. Seleccionar aleatoriamente K observaciones del conjunto de datos como centros iniciales.
3. Asignar cada observación a su centro más cercano.
4. Para cada uno de los K clústers, actualizar el centro calculando los nuevos promedios de todas las observaciones del clúster. El centro del grupo k es un vector de longitud p que contiene los promedios de todas las variables para las observaciones en el grupo; p es el número de variables.
5. Repetir los pasos 3 y 4 hasta que las asignaciones dejen de cambiar o se alcance el número máximo de iteraciones. Por defecto, R utiliza 10 como valor predeterminado para el número máximo de iteraciones.

Ejemplo: el conjunto de datos USArrests

Descripción

Usaremos los conjuntos de datos de demostración “USArrests”. Este conjunto de datos contiene estadísticas, en arrestos por cada 100mil residentes por asalto, asesinato y violación en cada uno de los 50 estados de Estados Unidos, en 1973. También se da el porcentaje de la población que vive en áreas urbanas.
Un data frame con 50 observaciones sobre 4 variables:

1. `Murder` arrestos por asesinato (por 100mil)
2. `Assault` arrestos por asalto (por 100mil)
3. UrbanPop porcentaje de población urbana
4. Rape arrestos por violación (por 100mil)

Ejemplo: el conjunto de datos USArrests

Observación: solo variables continuas

Los datos deben contener solo variables continuas, ya que el algoritmo de K-medias calcula promedios.

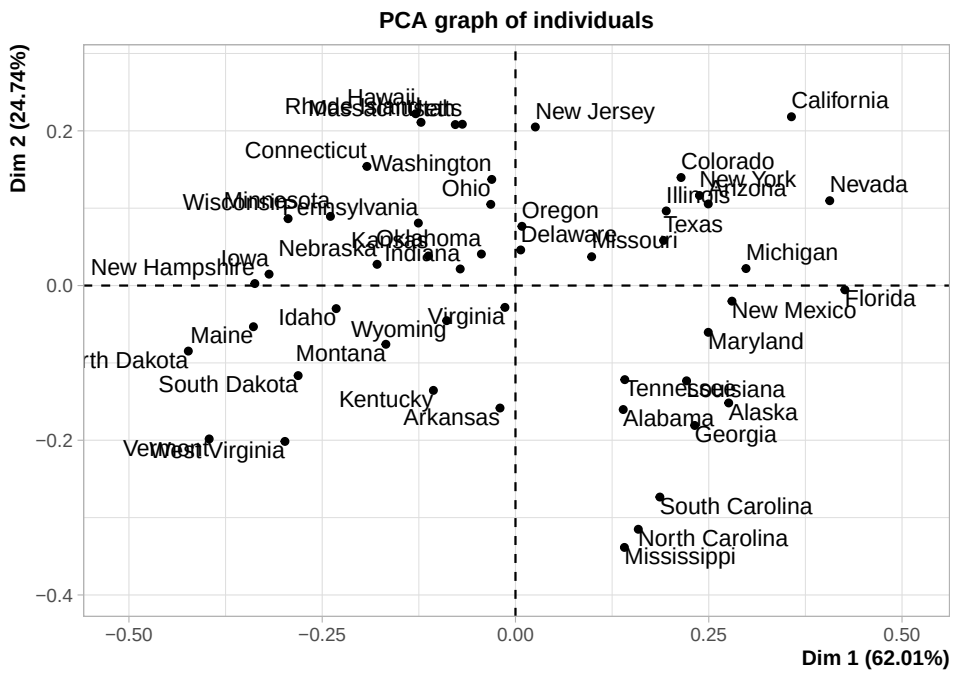
Cargamos los datos

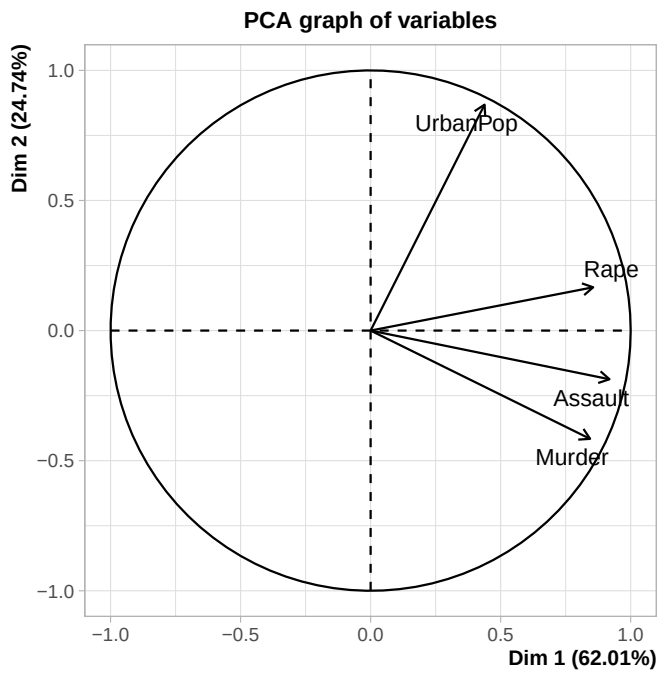
```
data("USArrests")
```

Estandarizamos los datos con la función *scale()*

Estandarizamos los datos de la siguiente manera:

```
M <- apply(USArrests,MARGIN = 2,FUN = mean)
desvio <- function(x){sqrt(sum((x-mean(x))^2))}
D <- apply(USArrests,MARGIN = 2, FUN = desvio)
datos <- scale(USArrests,center = M,scale = D)
```


PCA graph of individuals
Dim 2 (24.74%)
Dim 1 (62.01%)
0.2
0.0
−0.2
−0.4
−0.50
−0.25
0.00
0.25
0.50
Hawaii
New Jersey
California
Connecticut
Washington
Colorado
New York
Ohio
Arizona
Nevada
Minnesota
Pennsylvania
Oregon
Illinois
Texas
Delaware
Missouri
Nebraska
Oklahoma
Indiana
Michigan
Iowa
New Hampshire
Florida
New Mexico
Idaho
Virginia
Wyoming
Maine
Maryland
Montana
South Dakota
Tennessee
Louisiana
Kentucky
Alaska
Arkansas
Alabama
Georgia
West Virginia
South Carolina
North Carolina
Mississippi


PCA graph of variables
Dim 2 (24.74%)
1.0
0.5
0.0
−0.5
−1.0
UrbanPop
Rape
Assault
Murder
−1.0
−0.5
0.0
0.5
1.0
Dim 1 (62.01%)

2/ *K*-medias difuso

Introducción

Clustering exclusivo

En el clustering exclusivo, cada observación se asigna de forma exclusiva a un solo clúster. Es decir, cada observación puede pertenecer completamente a un clúster o no. El método de K-Medias es un algoritmo de clustering exclusivo.

Clustering difuso

En el clustering difuso, en lugar de poner cada observación en grupos separados, se asigna una "probabilidad" o "grado de pertenencia" de que una observación esté en un determinado grupo. Es decir, cada observación puede pertenecer a múltiples grupos junto con su puntuación de pertenencia. Uno de los algoritmos de clustering difuso más utilizados es el algoritmo de K-medias difuso (o FCM por sus siglas en inglés: Fuzzy C-Means).

Notación y terminología básica

La función objetivo

Al igual que K-medias, su versión difusa pertenece a la clase de algoritmos basados en funciones objetivo. Estos algoritmos definen un criterio de agrupamiento usando una función objetivo que depende de la partición difusa. El procedimiento consiste en minimizar iterativamente esta función hasta obtener una partición difusa óptima.
En este caso la función objetivo es:

$$J_m(P, \mu) = \sum_{k=1}^{K} \sum_{i=1}^{n} p_{ki}^m d_{ki}^2$$

en donde d_{ki} indica la distancia entre la observación i y el centro del grupo k:

$$d_{ki} = \|x_i - \mu_k\|$$

Notación y terminología básica

La matriz de pesos P

La matriz de pesos

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & p_{ki} & \vdots \\ p_{K1} & \cdots & p_{Kn} \end{pmatrix}$$

es una matriz $K \times n$ cuya entrada ki indica el grado de pertenencia del individuo i al grupo k.

El caso de K-medias exclusivo

A modo de ejemplo, cuando hacemos un clustering exclusivo, la matriz de pesos o pertenencia tendrá entradas

$$p_{ki} = \begin{cases} 1 & \text{si } x_i \in C_k \\ 0 & \text{si } x_i \notin C_k \end{cases}$$

En el caso difuso vamos a permitir que los pesos tomen valores intermedios entre 0 y 1.

Notación y terminología básica

Restricciones sobre la matriz P

Las entradas de la matriz P están sujetas a las siguientes restricciones:

- $p_{ki} \in [0,1] \ ; \ 1in \ , 1 \leq k \leq K$
- $0 \sum_{i=1}^{n} p_{ki} \leq n \ ; \ 1 \leq k \leq K$
- $\sum_{k=1}^{K} p_{ki} = 1 \ ; \ 1 \leq i \leq n$

El parámetro difusor m

En la función objetivo el peso asociado a cada distancia (al cuadrado), p_{ki}^{m}, es la m-ésima potencia del grado de pertenencia del i-ésimo dato al grupo k. Cuando $m \longrightarrow 1$ la partición óptima es cada vez más cercana a una partición exclusiva, mientras que cuando $m \longrightarrow \infty$ la partición óptima se aproxima a la matriz con todos sus valores iguales a $1/K$.

El algoritmo

El procedimiento consiste de los siguientes pasos:

1. Fijar K y m. Elegir una matriz inicial $P^{(0)}$.
2. Paso de actualización: calcular los centros de los grupos con la fórmula

$$\mu_k = \sum_{i=1}^{n} p_{ki}^m x_i / \sum_{i=1}^{n} p_{ki}^m; \quad 1 \leq k \leq K.$$

3. Paso de asignación: calcular la matriz de pesos P con

$$p_{ki} = 1 / \sum_{j=1}^{K} \left(\frac{d_{ki}}{d_{ji}} \right)^{\frac{2}{m-1}}; 1 \leq i \leq n; 1 \leq k \leq K.$$

4. Si se alcanzó el criterio de parada, terminar. En caso contrario, regresar al paso 2.

Por ejemplo, el criterio de parada más utilizado es:

- Un número máximo de iteraciones
- Que la variación en la matriz P sea muy pequeña: $||P^{k+1} - P^k|| < \epsilon$.

3/ Expectation-Maximization para mezclas de normales

Normales independientes

Simetría rotacional

Sean X e Y independientes con distribución normal estándar

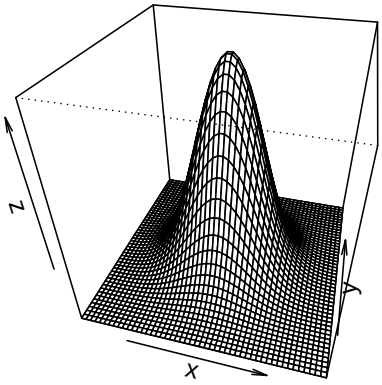
$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (z \in \mathbb{R}).$$

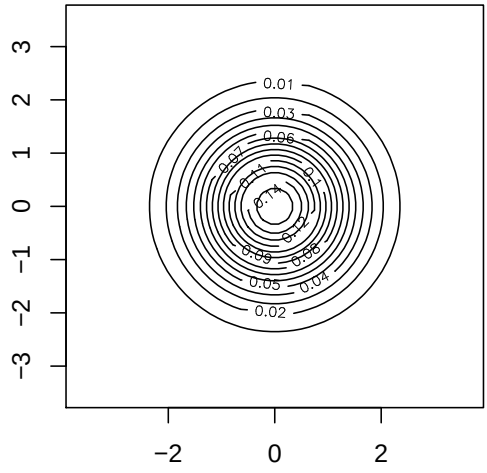
La densidad conjunta de X e Y está dada por

$$p(x, y) = \varphi(x)\varphi(y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} \quad ((x, y) \in \mathbb{R}^2).$$

La propiedad clave de esta densidad es que es una función de $r^2 = x^2 + y^2$, en donde r es la distancia del punto (x, y) al origen. Esto hace que la gráfica de la función $p(x, y)$ se obtenga haciendo girar la campana de Gauss sobre el eje z (las secciones son de hecho proporcionales a la campana de Gauss).

2 normales independientes





Combinaciones lineales de normales independientes

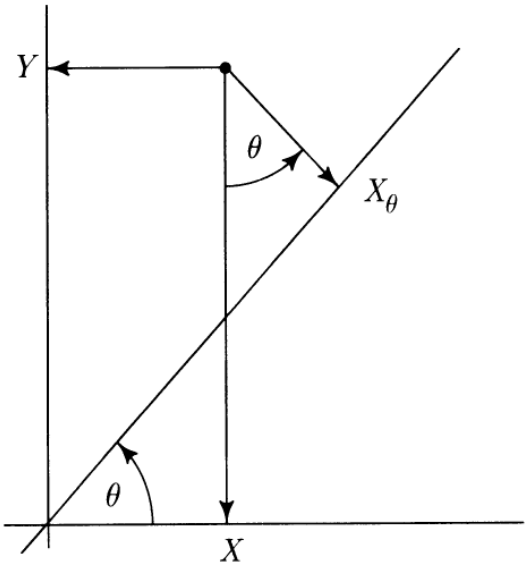
Combinaciones lineales de variables normales independientes son siempre normales. Este hecho importante es otra consecuencia de la simetría rotacional de la distribución conjunta de variables independientes X e Y con distribución normal estándar.

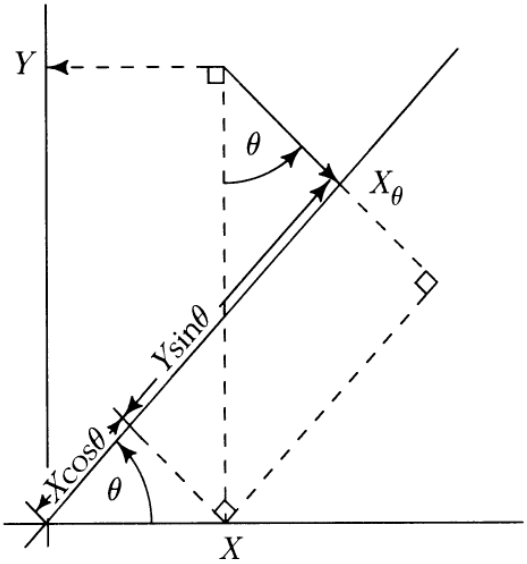
Proyección sobre una recta

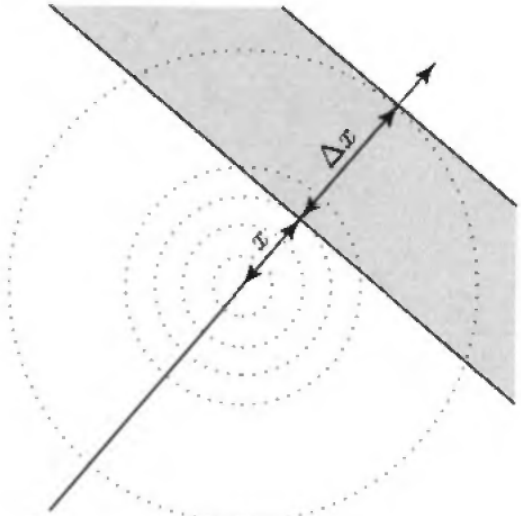
Para ver esto, sea X_θ la primer coordenada de (X, Y) relativa a un nuevo sistema de ejes que forma un ángulo θ con los ejes originales. Vemos que

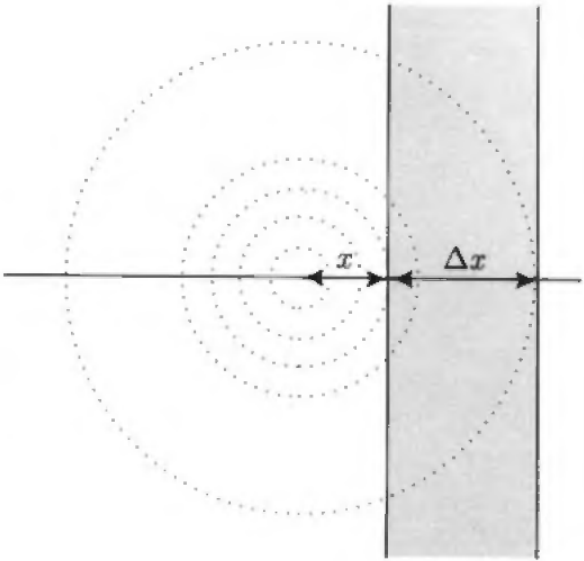
$$X_\theta = (\cos\theta, \sin\theta) \cdot (X, Y) = X\cos\theta + Y\sin\theta$$

Pero en vista de la simetría rotacional de la densidad conjunta de (X, Y), la distribución de X_θ debe ser la misma que la distribución de X, que es normal estándar, cualquiera sea el ángulo θ. Entonces $X_\theta \sim N(0, 1)$.


Y
θ
X_θ
θ
X


Y
θ
X_θ
Ysinθ
Xcosθ
θ
X


Δx
x


x
Δx

Combinaciones lineales de normales independientes

Producto escalar con cualquier vector

Hemos visto que al hacer el producto escalar $X_\theta = (\cos\theta, \sin\theta) \cdot (X, Y)$ obtenemos una normal estándar. ¿Qué ocurre si hacemos el producto con cualquier vector (a, b)?

En ese caso podemos escribir

$$(a, b) \cdot (X, Y) = \|(a, b)\| \frac{1}{\|(a, b)\|} (a, b) \cdot (X, Y) \sim \|(a, b)\| N(0, 1) = N(0, a^2 + b^2)$$

Es decir, $aX + bY$ es una normal de media o y varianza $a^2 + b^2$.

Más aún, $aX + bY + c$ es una normal de media c y varianza $a^2 + b^2$.

Combinaciones lineales de normales independientes

Si $X \sim N(\mu, \sigma^2)$ e $Y \sim N(\lambda, \tau^2)$ son independientes, entonces

$$aX + bY \sim N(a\mu + b\lambda, a^2\sigma^2 + b^2\tau^2).$$

Para probarlo, observar que

$$U = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad V = \frac{Y - \lambda}{\tau} \sim N(0, 1),$$

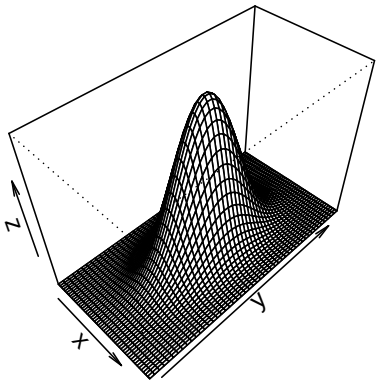
y son independientes.

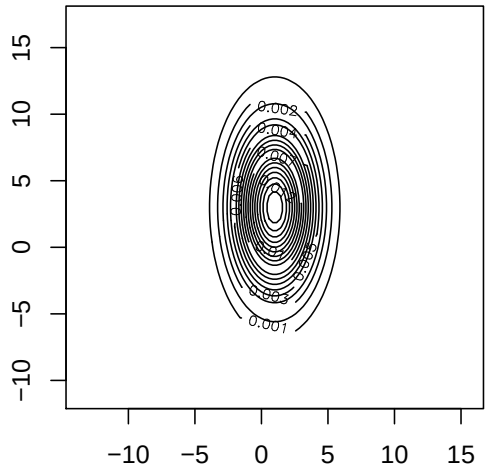
Entonces

$$aX + bY = a(\sigma U + \mu) + b(\tau V + \lambda) = (a\sigma)U + (b\tau)V + (a\mu + b\lambda)$$

La afirmación se deduce de la diapositiva anterior.

Con desvios diferentes





Normal bi-variada con correlación

Recordar el coeficiente de correlación ρ

El coeficiente de correlación es

$$\rho = \rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} \in [-1, 1].$$

Normales correlacionadas

Para obtener un par de variables normales correlacionadas X e Y, podemos comenzar con variables independientes X y U normales estándar. Luego tomamos la proyección del par (X, U) sobre la recta que forma un ángulo θ con el eje X.

Normales correlacionadas

De la misma forma que vimos antes, la proyección Y se escribe como

$$Y = X \cos\theta + U \sin\theta,$$

y sabemos que Y tiene distribución normal estándar.

Calculo de la correlación

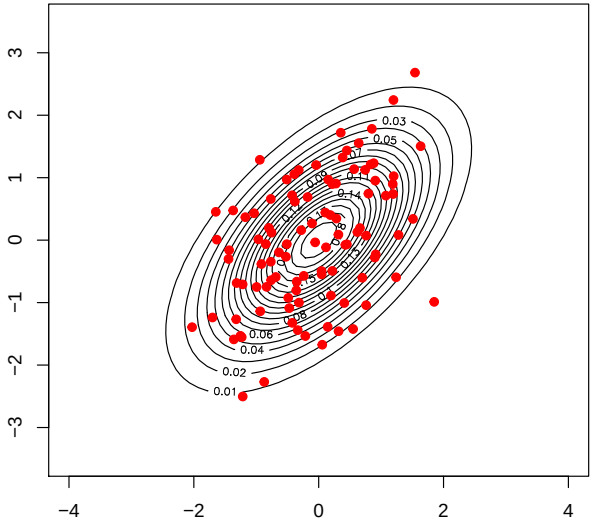
Por otro lado, la correlación entre X e Y es entonces

$$\begin{aligned}\rho_{XY} = \mathrm{E}(XY) &= \mathrm{E}(X(X\cos\theta + U\sin\theta)) = \mathrm{E}(X^2\cos\theta + XU\sin\theta) \\ &= \underbrace{\mathrm{E}(X^2)}_{1}\cos\theta + \underbrace{\mathrm{E}(XU)}_{0}\sin\theta = \cos\theta.\end{aligned}$$

Para resumir, $\rho = \cos\theta$. Como $\sin\theta = \sqrt{1-\rho^2}$, podemos escribir

$$Y = \rho X + \sqrt{1-\rho^2}\, U$$

Normales correlacionadas



Observaciones

- Se puede ver que la densidad conjunta de X e Y en el ejemplo anterior viene dada por

$$p(x,y) = \frac{1}{2\pi\sqrt{\det(S)}} \exp\left\{-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^t S^{-1} \begin{pmatrix} x \\ y \end{pmatrix}\right\}$$

  en donde

$$S = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

  es la matriz de covarianzas (coincidente con la de correlaciones en este caso).

- Este ejemplo podemos pensarlo como un producto matricial

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X \\ \rho X + \sqrt{1-\rho^2}U \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} X \\ U \end{pmatrix}$$

Observaciones

- Notar que

$$\begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}^t = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = S$$

- El ejemplo de normales independientes con desvíos y medias diferentes puede pensarse como un producto matricial

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & \tau \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} + \begin{pmatrix} \mu \\ \lambda \end{pmatrix}$$

  En este caso también resulta que

$$\begin{pmatrix} \sigma & 0 \\ 0 & \tau \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & \tau \end{pmatrix}^t = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix} = S$$

Definición de normal bi-variada

Primera definición

Decimos que el par de variables (X, Y) tiene distribución normal bivariada si se puede escribir como

$$\begin{pmatrix} X \\ Y \end{pmatrix} = A \begin{pmatrix} U \\ V \end{pmatrix} + b$$

en donde U y V son normales estándar independientes, A es una matriz de 2×2 y b es un vector columna. En este caso la matriz de covarianzas es $S = AA^t$ y el vector de promedios es b.

Segunda definición

Decimos que el par de variables (X, Y) tiene distribución normal bivariada si toda combinación lineal $aX + bY$ tiene distribución normal.

Equivalencia

Ambas definiciones son equivalentes.

Definición de normal multi-variada

Primera definición

Decimos que la n-upla de variables $(X_1, \ldots, X_n)$ tienen distribución normal multivariada si se puede escribir como

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = A \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} + b$$

en donde $U_1, \ldots, U_n$ son normales estándar independientes, A es una matriz de $n \times n$ y b es un vector columna. En este caso la matriz de covarianzas es $S = AA^t$ y el vector de promedios es b.

Segunda definición

Decimos que la n-upla de variables $(X_1, \ldots, X_n)$ tienen distribución normal multivariada si toda combinación lineal $a_1 X_1 + \cdots + a_n X_n$ tiene distribución normal.

Ambas definiciones son equivalentes.

Clustering con mezcla de normales

Ejemplo: Descripción de los datos *Faithful*

Los Geysers son fuentes naturales de agua que se eyectan al aire, a intervalos más o menos regulares, como una columna de agua caliente y vapor. Old Faithful es uno de esos Geysers y es la atracción más popular del Parque Nacional de Yellowstone. Old Faithful puede variar en altura de 100 a 180 pies con un promedio de 130 a 140 pies. Las erupciones normalmente duran entre 1.5 y 5 minutos.

Desde el 1 de agosto al 15 de agosto de 1985, se observó Old Faithful y se anotaron los tiempos de espera entre erupciones sucesivas. Se observaron 300 erupciones, por lo que se registraron 299 tiempos de espera (en minutos).

El algoritmo EM para mezcla de normales

El modelo

La distribución que consiste en una mezcla de dos componentes normales fue considerada por primera vez por Karl Pearson hace más de 100 años (en 1894) y está dada explícitamente por

$$p(x) = p\phi(x, \mu_1, \sigma_1^2) + (1 - p)\phi(x, \mu_2, \sigma_2^2)$$

donde $\phi(x, \mu, \sigma_2)$ denota una densidad normal con media μ y varianza σ^2.

Objetivo

El algoritmo (EM) es un método iterativo para calcular estimaciones de máxima verosimilitud (máximos locales en realidad) de parámetros en modelos estadísticos, donde el modelo depende de variables latentes no observadas. En nuestro ejemplo la variable latente es el cluster al que pertenece cada observación.

Empecemos en 1D

Las variables latentes

La variable que medimos es X que representa el tiempo de espera entre dos erupciones consecutivas. Disponemos de una muestra $x = (x_1, x_2, \ldots, x_n)$ con n observaciones independientes de una mezcla de dos distribuciones normales, y sea $z = (z_1, z_2, \ldots, z_n)$ las variables latentes que determinan la componente de la que se origina cada observación.

Podemos pensar a la mezcla de la siguiente forma:

$$X \mid (Z = 1) \sim N(\mu_1, \sigma_1^2) \quad \text{y} \quad X \mid (Z = 2) \sim N(\mu_2, \sigma_2^2)$$

dónde

$$P(Z = 1) = p, \quad \text{y} \quad P(Z = 2) = 1 - p.$$

El objetivo es estimar los parámetros desconocidos:

$$\theta = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$$