

Fundamentos de Análisis de Datos Distancias Estadísticas

Dra. Andrea Alejandra Rey

Especialización en Ciencia de Datos - ITBA



Permiten detectar presencia de outliers

Se usan como entrada de varios algoritmos de análisis multivariado

Describen la cercanía entre dos objetos estadísticos

Distancias estadísticas

El estudio de medidas de distancia y similaridad considera la estructura geométrica de los datos, la cual puede ser de gran importancia para resolver adecuadamente problemas relevantes del análisis de datos.

En particular, en el caso de aprendizaje automático (ML - *Machine Learning*) los datos utilizados para el entrenamiento del método pueden presentar fallas de recolección o generar desviaciones a largo plazo cuando se utilizan como entradas de ciertas funciones.

En este sentido, las distancias estadísticas brindan información para que los equipos puedan detectar cambios en los datos que pudieran afectar el rendimiento del modelo.

Distancias estadísticas

Problemas en la práctica real

- ➡ Un error de indexación de datos rompe el mapeo ascendente.
- ➡ Un mal manejo de texto causa símbolos desconocidos.
- ➡ Fuentes de características con diferentes coordenadas o indexación.
- ➡ La fuente de datos de terceros produce modificaciones al eliminar una función, cambiar el formato o mover datos.
- ➡ La recopilación periódica de datos falla provocando la falta de valores.
- ➡ La ingeniería de software cambia el significado de un campo.
- ➡ Cambios en la funcionalidad de la biblioteca por parte de terceros.
- ➡ Presunción de formato válido que cambia y de repente deja de ser válido.
- ➡ El mundo exterior cambia drásticamente (por ejemplo, la pandemia de COVID-19) y cada característica cambia.
- ➡ Un aumento drástico en el volumen sesga las estadísticas.

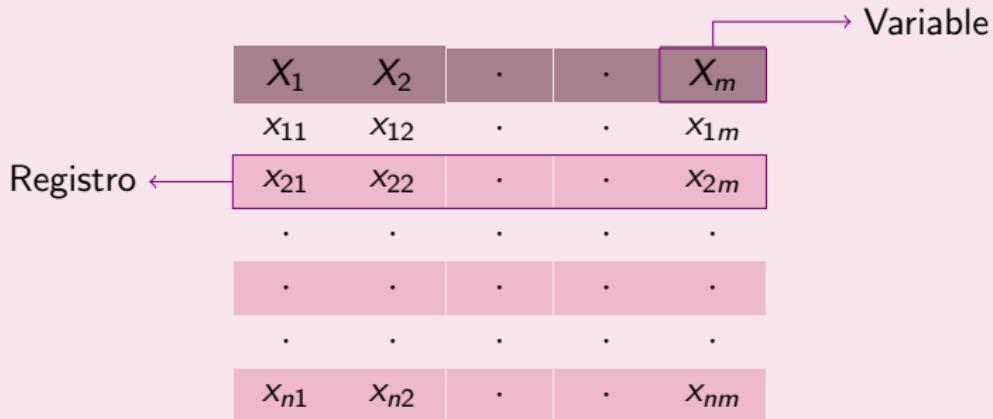
Distancias estadísticas

Pregunta

¿Dónde utilizar las comprobaciones de distancia estadística?

- ➡ Entradas del modelo
- ➡ Resultados del modelo
- ➡ Datos reales (*ground truth*)

Distancias estadísticas



Pregunta

¿Cómo podemos medir la diferencia existente entre dos registros de un conjunto de datos?

Distancia

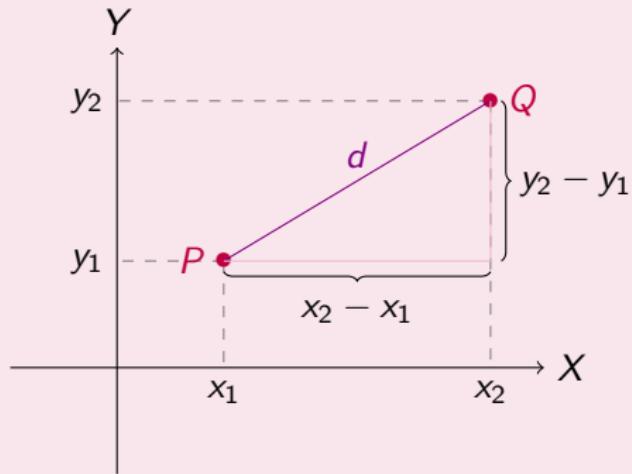
Definición

Dado un conjunto A , una **distancia** es una función $d : A \times A \rightarrow \mathbb{R}$ tal que para todo $x, y, z \in A$ se satisfacen las siguientes propiedades:

- 1 no negatividad: $d(x, y) \geq 0$,
- 2 identidad: $d(x, y) = 0$ si y sólo si $x = y$,
- 3 simetría: $d(x, y) = d(y, x)$,
- 4 desigualdad triangular: $d(x, z) \leq d(x, y) + d(y, z)$.

Distancia euclídea

Supongamos que tenemos dos registros de dos variables $P = (x_1, y_1)$ y $Q = (x_2, y_2)$.



Usando el Teorema de Pitágoras tenemos que:

$$d(P, Q) = d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Distancia euclídea

Definición

Sean $P = (x_1, x_2, \dots, x_m)$ y $Q = (y_1, y_2, \dots, y_m)$ dos puntos en \mathbb{R}^m .

La **distancia euclídea** entre P y Q se calcula como:

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2}.$$

Observación

- La distancia euclídea se usa para comparar dos puntos en un espacio m dimensional.
- Al trabajar con datos multivariados, la distancia euclídea tiene sentido cuando las variables comparten la unidad de medida.



Eligiendo el auto que más se acerca al de nuestros sueños...

Supongamos que queremos comprar un automóvil con tracción en las cuatro ruedas que nos llevará a las montañas.

Tenemos en vista el auto de nuestros sueños, pero sabemos que debemos comparar los gastos de consumo.

Para ello, vamos a observar otros autos 4x4 en el mercado* y comparar su millaje† de gasolina por galón‡ y desplazamiento en ruta (el volumen total de todos los cilindros del motor) para encontrar otros autos que nos puedan interesar.

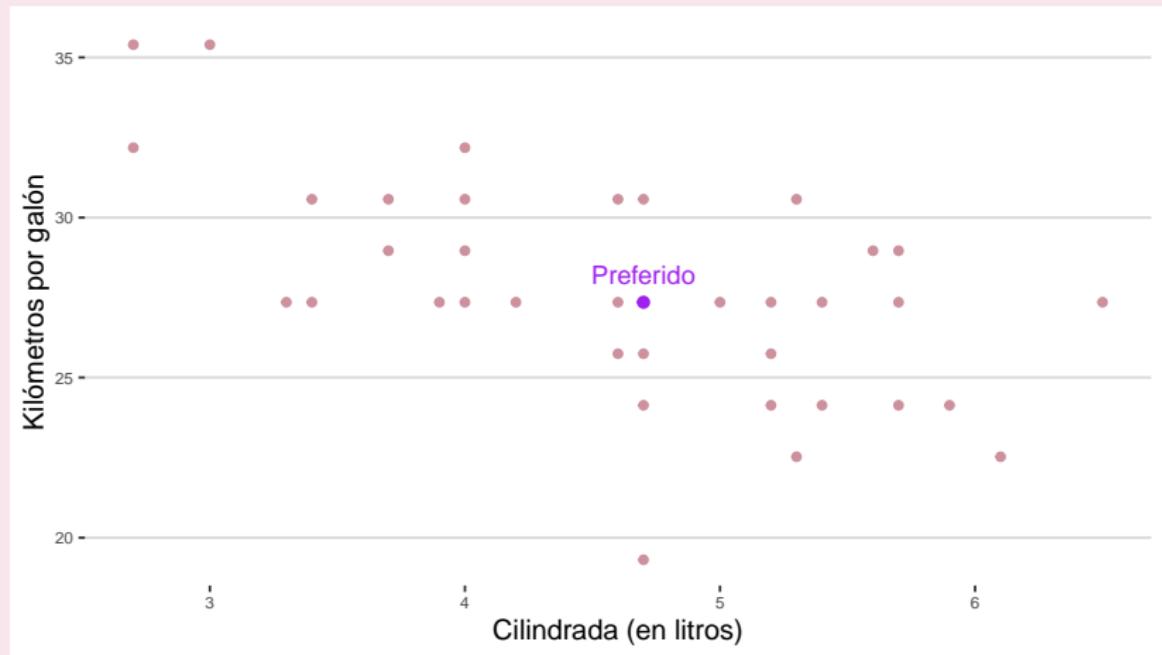
En otras palabras, estamos buscando los vecinos más cercanos del coche de ensueño, con respecto a esas dos medidas.

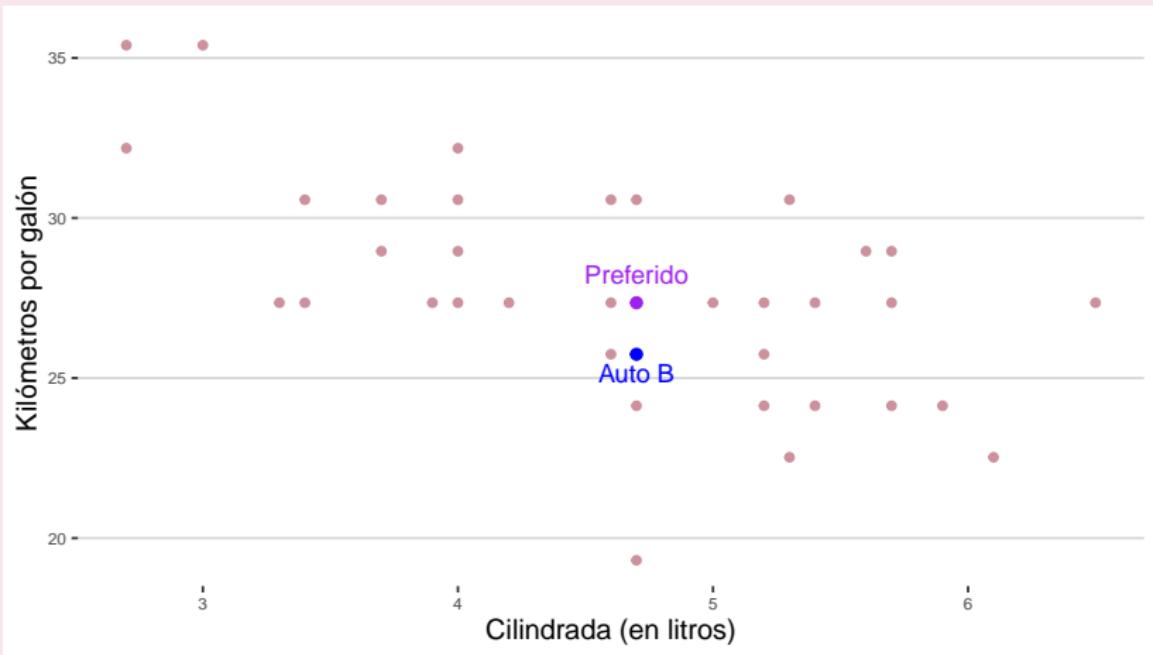
*Base de datos `mpg` disponible en R.

†Una milla equivale a 1.609 kilómetros.

‡Un galón equivale a 3.785 litros.

La ubicación de estas variables por cada auto es:





La distancia euclídea entre nuestro auto preferido y el auto B es:

$$d(\text{Preferido}, \text{Auto B}) = d((4.7, 27.353), (4.7, 25.744)) = \sqrt{0 + 1.609^2} = 1.609.$$

Pregunta

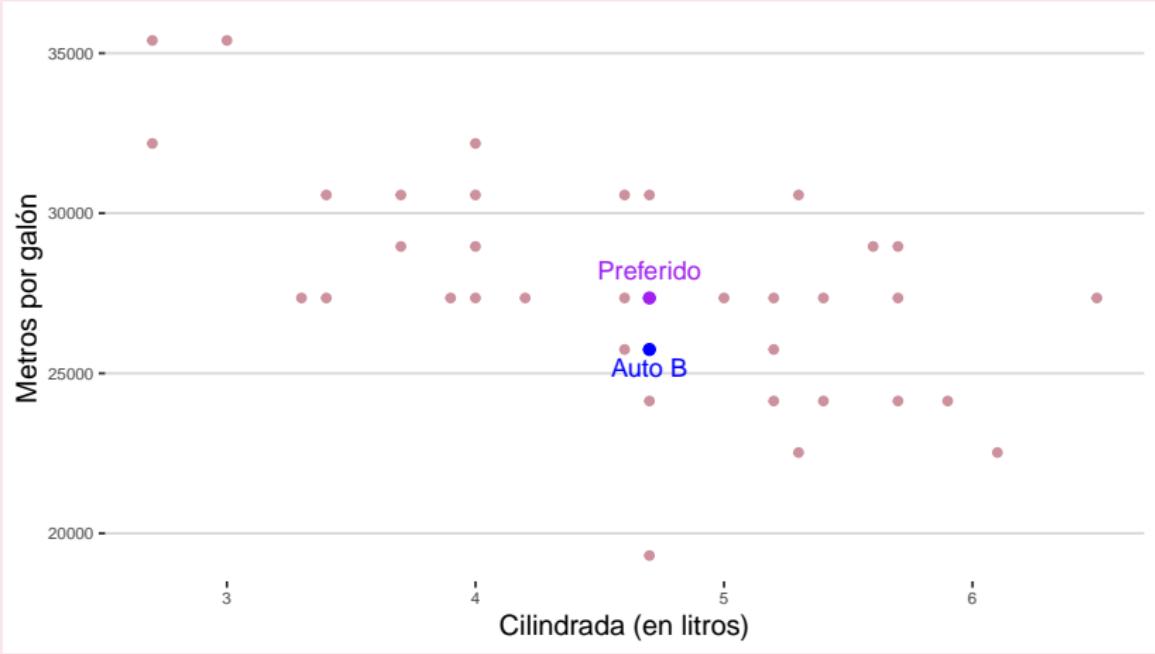
¿Qué sucede si en vez de tener la información de kilómetros por galón, tenemos la información de metros por galón?

La distancia euclídea entre nuestro auto preferido y el auto B es:

$$d(\text{Preferido}, \text{Auto B}) = d((4.7, 27353), (4.7, 25744)) = \sqrt{0 + 1609^2} = \boxed{1609}.$$

Vemos que la distancia creció considerablemente.

Sin embargo...



Visualmente no se registran cambios.

Observación

La distancia Euclídea depende de la unidad de medida de las variables involucradas.

Si estandarizamos la variable del recorrido por galón, entonces:

➡ usando kilómetros:

$$\frac{Y - \mu_{\text{km}}}{\sigma_{\text{km}}} = \frac{Y - 27.701}{3.395}.$$

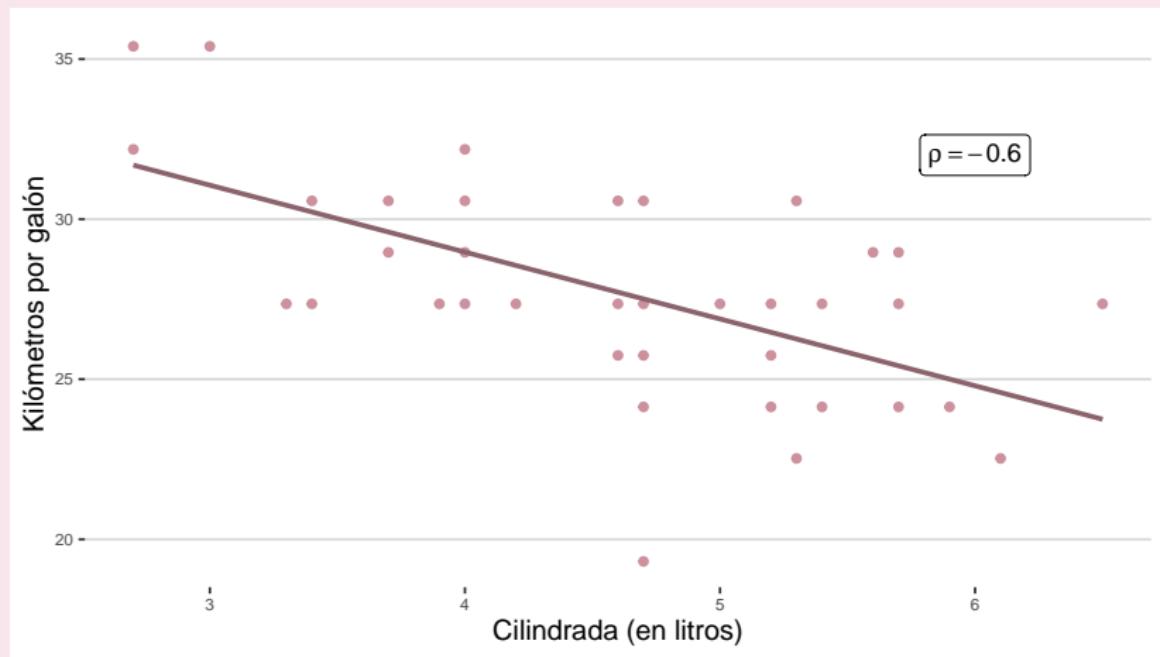
➡ usando metros:

$$\frac{Y - \mu_{\text{m}}}{\sigma_{\text{m}}} = \frac{Y - 27701}{3395}.$$

En ambos casos, la distancia euclídea entre nuestro auto preferido y el auto B es:

$$d(\text{Preferido}, \text{Auto B}) = d((4.7, -0.102), (4.7, -0.576)) = \sqrt{0 + 0.474^2} = \boxed{0.474}.$$

Además...





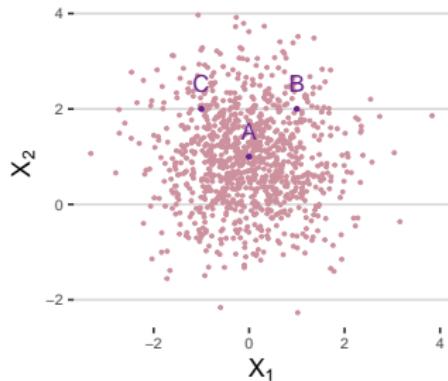
Mahalanobis (1893–1972)

Pregunta

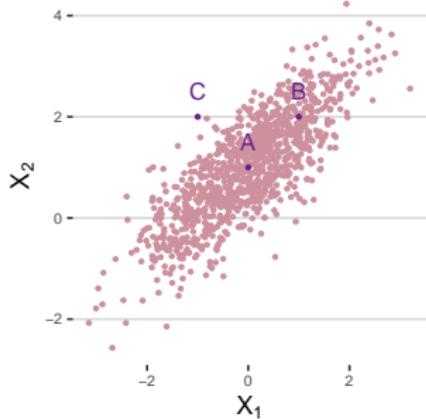
¿Cómo podemos considerar la correlación?

Distancia de Mahalanobis: Motivación

Datos no correlacionados



Datos correlacionados

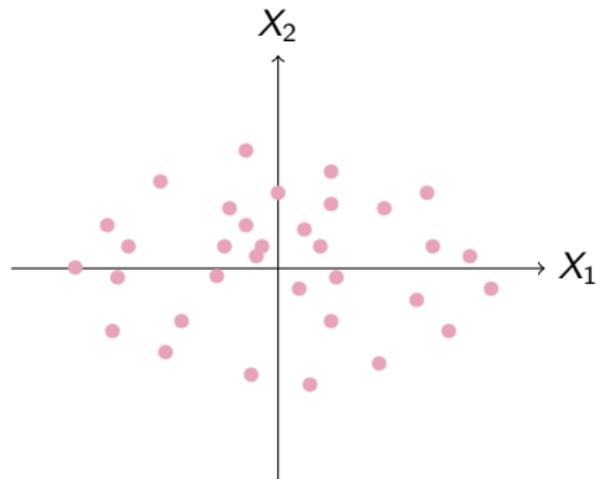


Si consideramos el punto A como el “centro de masa” de la distribución de los datos, vemos que B y C equidistan de A en términos de la distancia euclídea. Sin embargo, visualmente se aprecia que el punto C está más alejado de la nube de puntos que el punto B en el caso de datos correlacionados.

Distancia de Mahalanobis: Motivación

Primer caso

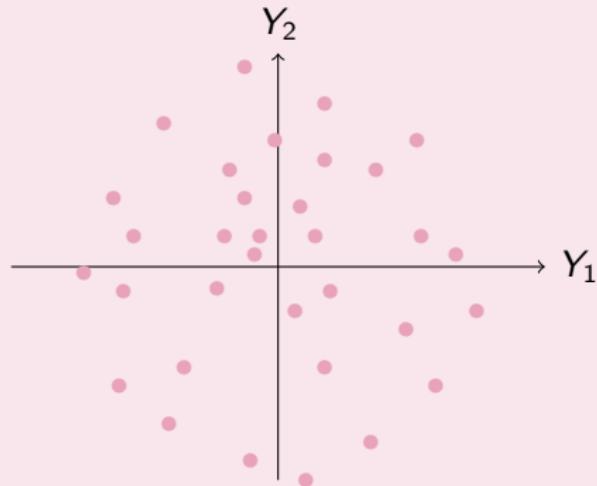
Las mediciones varían independientemente, pero las variabilidades son diferentes.



Distancia de Mahalanobis: Motivación

En este caso, una de la variable X_1 contribuye en la distancia euclídea mucho más que X_2 . Una manera de comparar estas variables en forma adecuada es estandarizando las coordenadas; es decir:

$$y_1 = \frac{x_1 - \bar{X}_1}{s_{X_1}} \quad \text{e} \quad y_2 = \frac{x_2 - \bar{X}_2}{s_{X_2}}.$$



Distancia de Mahalanobis: Motivación

Luego, si tenemos dos observaciones $\mathbf{x}_1 = (x_{11}, x_{12})$ y $\mathbf{x}_2 = (x_{21}, x_{22})$ podemos calcular la distancia euclídea en el espacio estandarizado $Y_1 \times Y_2$ de la siguiente manera:

$$\begin{aligned} d(\mathbf{y}_1, \mathbf{y}_2) &= \sqrt{(y_{11} - y_{21})^2 + (y_{12} - y_{22})^2} \\ &= \sqrt{\left(\frac{x_{11} - \bar{X}_1 - (x_{21} - \bar{X}_1)}{s_{X_1}} \right)^2 + \left(\frac{x_{12} - \bar{X}_2 - (x_{22} - \bar{X}_2)}{s_{X_2}} \right)^2} \\ &= \sqrt{\frac{(x_{11} - x_{21})^2}{s_{X_1}^2} + \frac{(x_{12} - x_{22})^2}{s_{X_2}^2}}. \end{aligned}$$

Distancia de Mahalanobis: Motivación

Observemos lo siguiente:

$$\begin{aligned} d(\mathbf{x}_1, \mathbf{x}_2) &= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2} \\ &= \sqrt{(x_{11} - x_{21} \quad x_{12} - x_{22}) \begin{pmatrix} x_{11} - x_{21} \\ x_{12} - x_{22} \end{pmatrix}} \\ &= \sqrt{(x_{11} - x_{21} \quad x_{12} - x_{22}) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{11} - x_{21} \\ x_{12} - x_{22} \end{pmatrix}} \\ &= \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)' I (\mathbf{x}_1 - \mathbf{x}_2)}. \end{aligned}$$

Distancia de Mahalanobis: Motivación

Por otro lado:

$$\begin{aligned} d(\mathbf{y}_1, \mathbf{y}_2) &= \sqrt{\frac{(x_{11} - x_{21})^2}{s_{X_1}^2} + \frac{(x_{12} - x_{22})^2}{s_{X_2}^2}} \\ &= \sqrt{\begin{pmatrix} \frac{x_{11}-x_{21}}{s_{X_1}} & \frac{x_{12}-x_{22}}{s_{X_2}} \end{pmatrix} \begin{pmatrix} \frac{x_{11}-x_{21}}{s_{X_1}} \\ \frac{x_{12}-x_{22}}{s_{X_2}} \end{pmatrix}^T} \\ &= \sqrt{(x_{11} - x_{21} \quad x_{12} - x_{22}) \begin{pmatrix} 1/s_{X_1}^2 & 0 \\ 0 & 1/s_{X_2}^2 \end{pmatrix} (x_{11} - x_{21} \quad x_{12} - x_{22})^T} \\ &= \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)' S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}, \end{aligned}$$

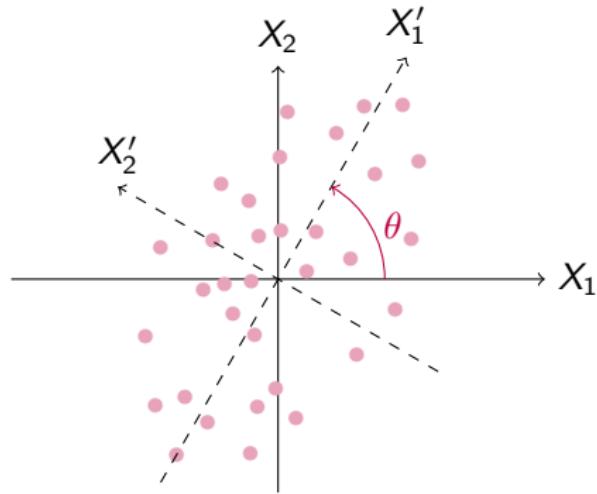
donde S es una matriz diagonal formada por las varianzas como componentes de su diagonal.

En general, si S es una matriz semi definida positiva la fórmula anterior define una distancia.

Distancia de Mahalanobis: Motivación

Segundo caso

Las mediciones no sólo tienen variabilidad diferente, sino que además están correlacionadas.



Distancia de Mahalanobis: Motivación

Vemos que si rotamos con un ángulo θ en sentido horario los ejes X'_1 y X'_2 , la situación se reduce al primer caso considerando los ejes X_1 y X_2 .

Las coordenadas en los ejes rotados son:

$$\begin{aligned}x'_1 &= \cos(\theta)x_1 + \sin(\theta)x_2, \\x'_2 &= -\sin(\theta)x_1 + \cos(\theta)x_2.\end{aligned}$$

El próximo paso es escalar.

Usando herramientas del Álgebra Lineal, se puede probar que la matriz de covarianzas se escribe como el producto de dos matrices, una de rotación y otra de escalamiento.

Distancia de Mahalanobis

Definición

Sea \mathcal{D} una distribución de probabilidad sobre \mathbb{R}^m con vector de medias $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_m)'$ y matriz de covarianza Σ .

Se define la **distancia de Mahalanobis** entre dos puntos $\mathbf{x} = (x_1, x_2, \dots, x_m)'$ e $\mathbf{y} = (y_1, y_2, \dots, y_m)'$ como:

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \Sigma^{-1} (\mathbf{x} - \mathbf{y})}.$$

Se define la **distancia de Mahalanobis** entre un punto $\mathbf{x} = (x_1, x_2, \dots, x_m)'$ y \mathcal{D} como:

$$d_M(\mathbf{x}, \mathcal{D}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}.$$

Distancia de Mahalanobis

Propiedades

- La distancia de Mahalanobis se puede usar para medir la distancia entre un punto P y una distribución de puntos \mathcal{D} .
- Fue introducida por Mahalanobis en 1936 como una generalización multidimensional de la idea de medir a cuántas desviaciones estándar se halla P de la media de \mathcal{D} .
- La distancia de Mahalanobis se anula si P coincide con la media de \mathcal{D} y aumenta a medida que P se aleja de esta media a lo largo de los ejes de las componentes principales.
- La distancia de Mahalanobis no tiene unidad de medida, es invariante por la escala elegida y tiene en cuenta la correlación de los datos.
- Uno de los usos más frecuentes de la distancia de Mahalanobis es la detección de outliers multivariados, los cuales indican una combinación inusual de dos o más variables.

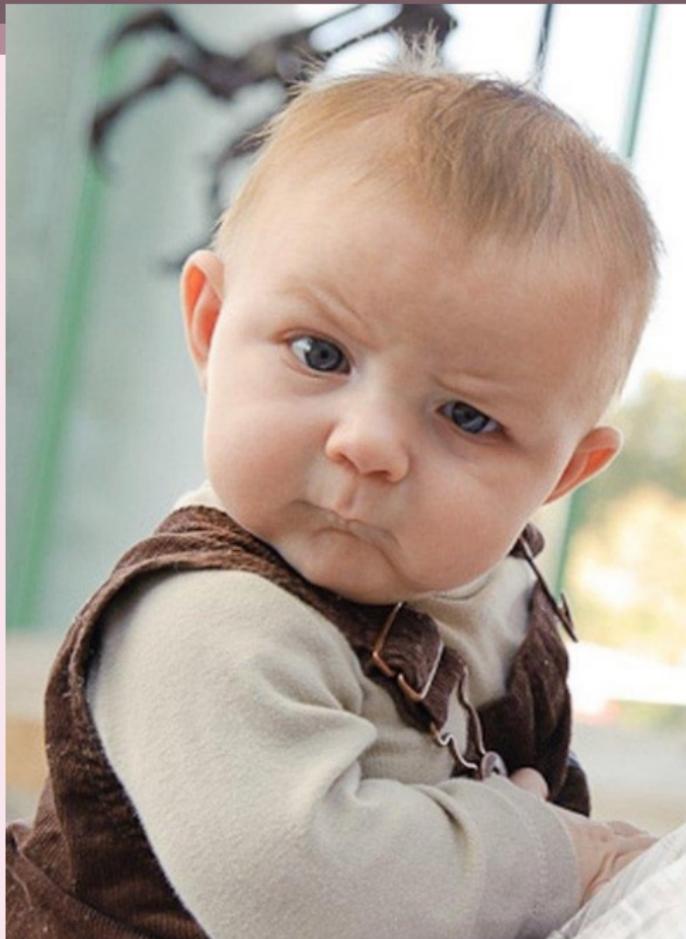


Distancia euclídea versus distancia de Mahalanobis

- ➡ La distancia euclídea tiene unidad de medida, pero la de Mahalanobis carece de unidad.
- ➡ Las operaciones empleadas para calcular la distancia euclídea son mucho más sencillas que las que se emplean para la distancia de Mahalanobis.
- ➡ La distancia euclídea no tiene en cuenta la correlación de los datos, mientras que la de Mahalanobis sí.
- ➡ La distancia de Mahalanobis puede pensarse como un índice que mide propiedades similares entre variables, incluso si las mismas son medidas con diferentes unidades.
- ➡ Uno de los mayores inconvenientes de la distancia de Mahalanobis es calcular la inversa de la matriz de covarianzas cuando las variables están fuertemente correlacionadas.

Distancia euclídea versus distancia de Mahalanobis

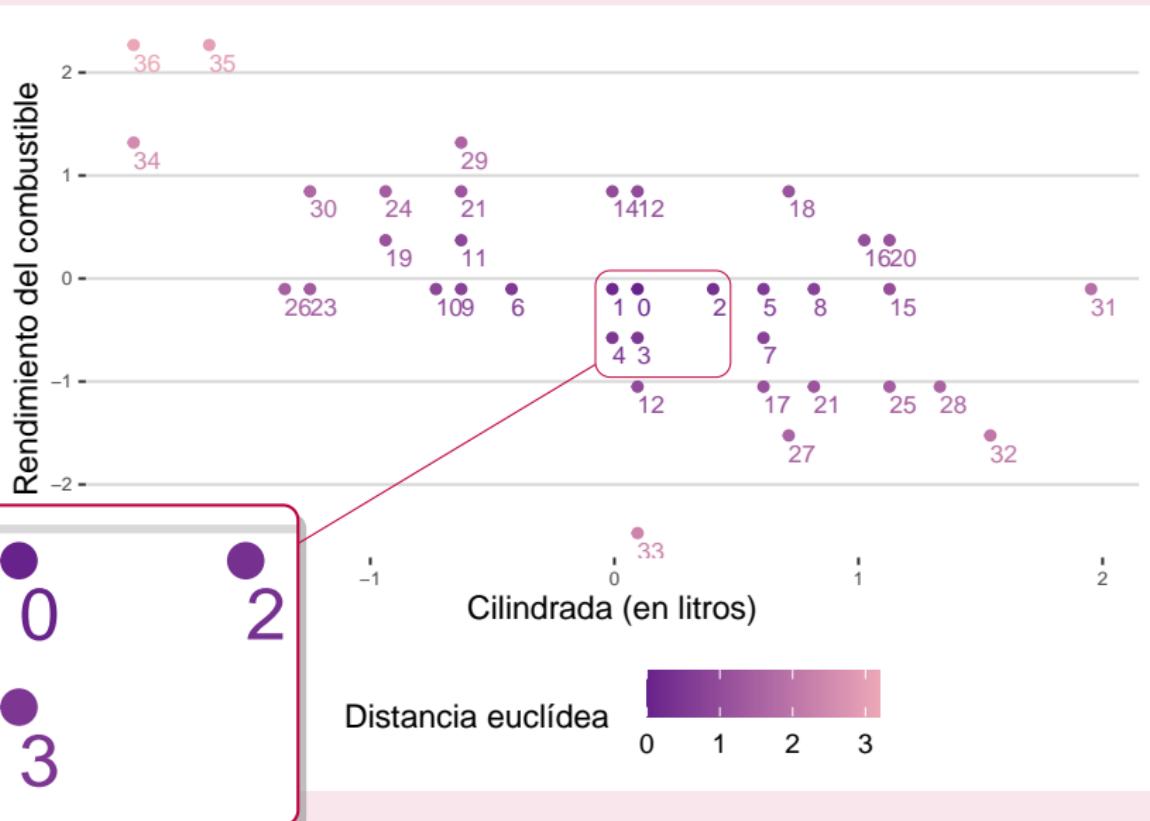
- Las distancias euclídea y de Mahalanobis coinciden en el caso de datos no correlacionados y de varianza unitaria.
- Si los ejes principales son re-escalados de manera tal de conseguir varianzas unitarias, la distancia de Mahalanobis coincide con la distancia euclídea en el espacio transformado.
- Si se desea usar la distancia euclídea por su simplicidad, se recomienda preprocesar los datos usando, entre otras, técnicas de estandarización, normalización, análisis de componentes principales. Sin embargo, como estas técnicas no consideran la dimensionalidad de los datos hacen que una de las ventajas de esta distancia se pierda.

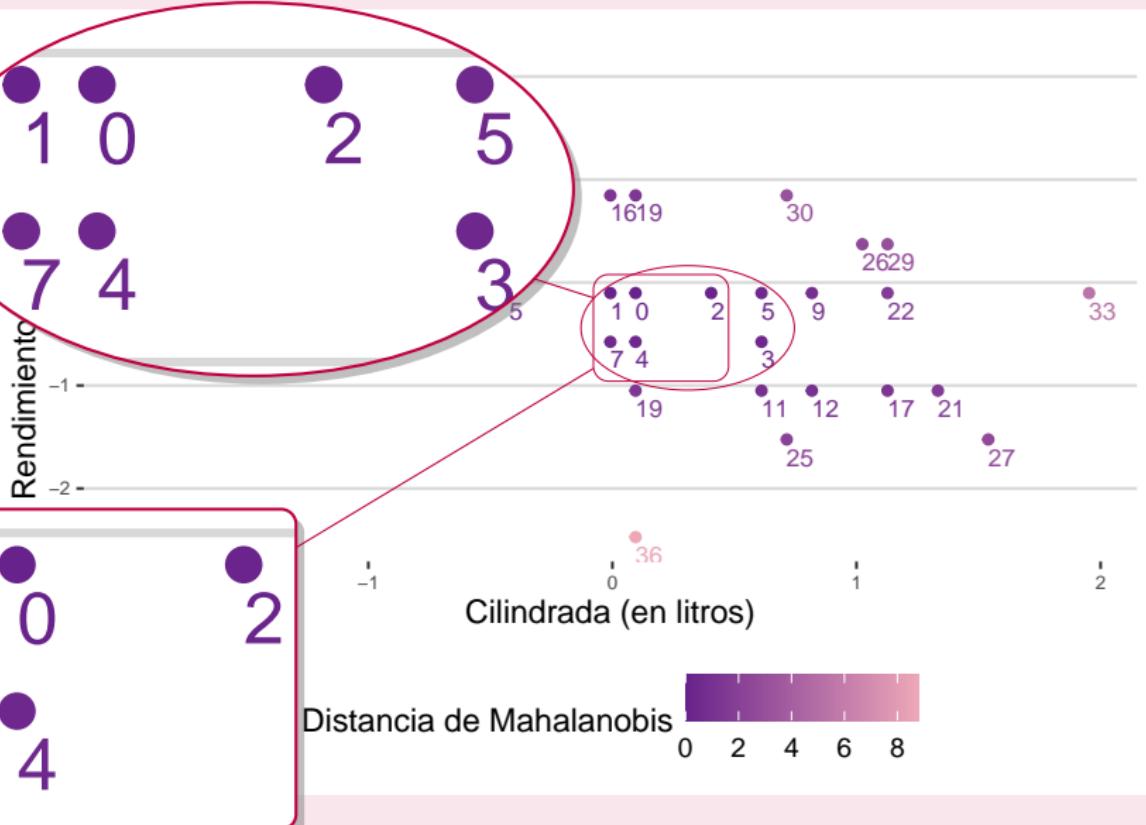


¿Y el auto de mis sueños??

Pregunta

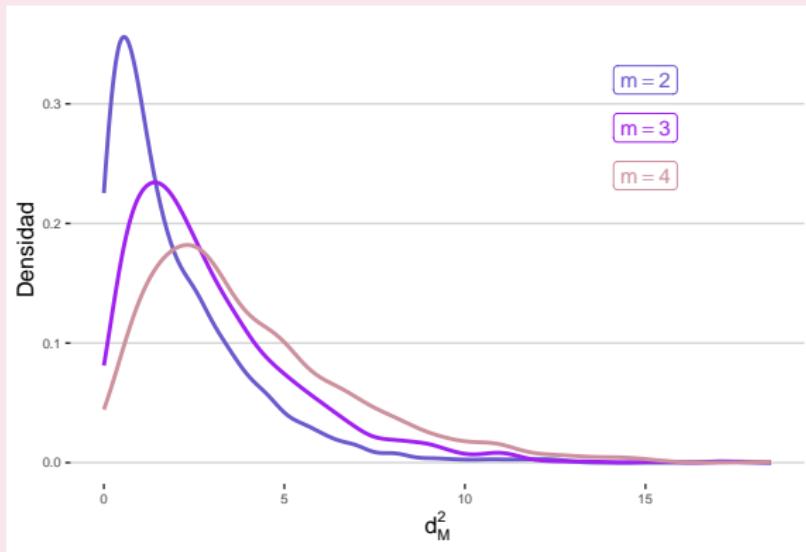
¿Cuáles son las 4 opciones más cercanas a nuestro auto favorito?





Distribución de la distancia de Mahalanobis

Generamos 1000 muestras de un vector aleatorio $\mathbf{X} = (X_1, X_2, \dots, X_m)$ con distribución Normal multivariada para $m = 2, 3, 4$. Luego, calculamos los cuadrados de las distancias de Mahalanobis a los correspondientes vectores de medias muestrales.



Distribución χ^2

Función de densidad de probabilidad

Si X es una variable aleatoria con distribución χ^2 con $k \in \mathbb{N}$ grados de libertad, su función de densidad de probabilidad para $x \geq 0$ es de la forma:

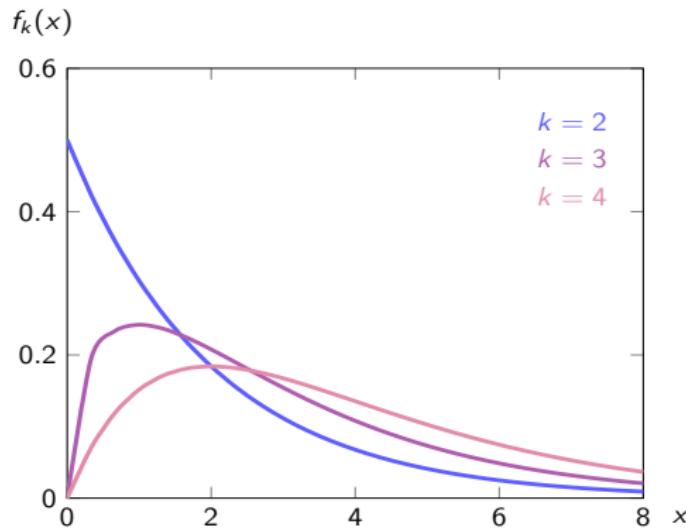
$$f_k(x) = \frac{x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right)}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)},$$

donde Γ denota la función gama definida como $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$.

En este caso notamos $X \sim \chi_k^2$.

Distribución χ^2

Gráfica

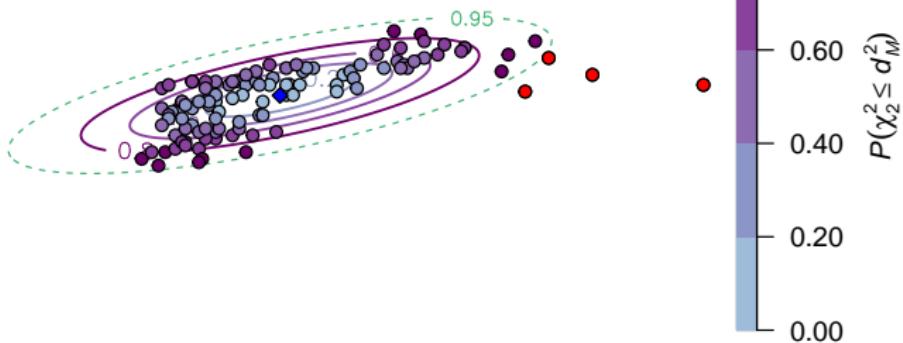


Detección de outliers multivariados

Sea $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ una muestra del vector aleatorio $\mathbf{X} = (X_1, X_2, \dots, X_m)$. Bajo el supuesto de normalidad, el algoritmo para detectar outliers consiste de los siguientes pasos:

- 1 Calcular el vector de medias muestral μ .
- 2 Calcular la matriz de covarianzas muestral Σ .
- 3 Calcular la distancia de Mahalanobis entre cada observación y μ .
- 4 Ordenar las distancias en orden creciente.
- 5 Calcular la probabilidad de los cuadrados de las distancias de Mahalanobis con la distribución χ_m^2 (los grados de libertad son la cantidad de variables).
- 6 Identificar los outliers con un nivel de significación α , como aquellas distancias d_M tales que $P(\chi_m^2 \leq d_M^2) \geq \alpha$.

$$\alpha = 0.95$$



Existen 4 outliers (puntos rojos) con una significación del 95%.

Ejemplo*



*Base de datos airquality de R

Temporary page!

\LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page, this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will disappear, because \LaTeX now knows how many pages to expect for this document.