UC | Chile

UC | Chile

Algebra Lineal Aplicada para Ciencia de Datos

Clase 2.

1. Introducción
2. Tema 1: Producto interior
3. Tema 2: Normas y distancias
4. Tema 3: Clustering

UC | Chile

Introducción

Introducción

Los conceptos de normas y distancias son fundamentales para problemas de análisis de datos, ya sea si queremos agrupar conjuntos de datos o si queremos determinar que tanto se parecen o no a otro determinado conjunto.
Nuestro objetivo en esta clases es definir matemáticamente los conceptos de norma y distnacia para vectores reales y ver una aplicación directa para un problema de clustering usando el algoritmo de clustering K-means.

UC | Chile

Tema 1: Producto interior

Productor interior

Comenzamos definiendo el productor interior o también llamado producto punto de dos vectores reales de tamaño n es el escalar para $x, y \in \mathbb{R}^n$ por

$$x \cdot y = x^\top y = x_1 y_1 + x_2 y_2 + ... + x_n y_n = \sum_{i=1}^{n} x_i y_i$$

Observe que la notación usando el superíndice $\top$ se refiere a la operación matricial transpuesta, la cual en este caso transforma el vector columna a vector fila.

Definición matemática

Las propiedades del producto interior son las siguientes.

Propiedades del producto interior	Representación matemática
Conmutatividad del producto interior	$x \cdot y = x^\top y = y^\top x = y \cdot x$
Asociatividad del producto interior respecto a la multiplicación por escalar	$(\alpha x)^\top y = \alpha(x^\top y)$
Distributividad del producto interior con la adición de vectores	$(x + y)^\top z = x^\top z + y^\top z$

Ejemplo producto interior

Promedio y promedio ponderado. Primerp vemos como ejemplo ue podemos escribir el promedio de las componenetes de un vector $x$de tamano n como el producto interior de x con el vector de tamano n que contiene solo 1s y multiplicado por $1/n$,

$$\frac{1}{n}\left(x_1 + x_2 + ... + x_n\right) = \left(\frac{1}{n}[1, ..., 1]^\top\right)^\top x$$

Similarmente podemos escribir el promedio ponderado de un vector con pesos $\omega_1, \omega_2, ..., \omega_n$ como el producto interior de x con el vector ω cuyas componentes son los pesos y multiplicado por $1/n$

$$\frac{1}{n}\left(\omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n\right) = \left(\frac{1}{n}[\omega_1, \omega_2, ..., \omega_n]^\top\right)^\top x$$

Ejemplo producto interior

Lista de compras item-precio. Consideramos una lista de comprar donde el vecto de x y c de valores reales de tamano n representan el número de articulos i comprados a precio c_i. Entonces, el producto interior entre x y c represrntan el precio total de la compra asociada al vector x.

UC | Chile

Tema 2: Normas y distancias

Definición matemática

Una norma en el espacio de vectores reales de tamaño n es una funcin que toma vectores y los lleva a valores no negativos. Esta función satisface:

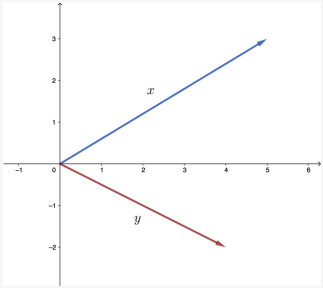
Propiedades de norma	Representación matemática
Positividad	$\|x\| \geq 0$ y $\|x\| = 0 \iff x = 0$
Homogeneidad	$\|\alpha\| = \|\alpha\|\|x\|$
Desigualdad triangular	$\|x + y\| \leq \|x\| + \|y\|$

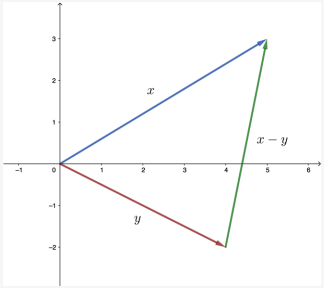
Ejemplo: Norma Euclideana o norma 2

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2} = \sqrt{x^\top x}$$

Distancia Euclideana entre dos vectores

$$x = \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \quad y = \begin{bmatrix} 4 \\ -2 \end{bmatrix}, \quad x - y = \begin{bmatrix} 1 \\ 5 \end{bmatrix}. \qquad \|x - y\|_2 = \sqrt{26}.$$



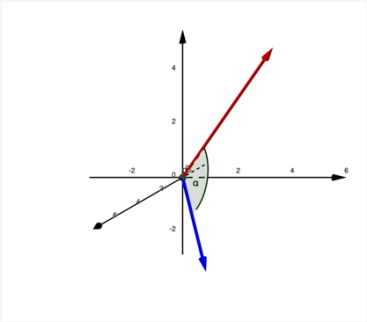


Ejemplo. Ángulo entre vectores

Considere los siguientes vectores $x = \begin{bmatrix} 2 \\ \sqrt{3} \\ -2 \end{bmatrix}, \quad y = \begin{bmatrix} 3/2 \\ 4 \\ 3\sqrt{3} \end{bmatrix}$

Calculamos las normas Euclideanas respectivas y el producto interior

$$\|x\|_2 = \sqrt{2^2 + (\sqrt{3})^2 + (-3)^2} = 4$$
$$\|y\|_2 = \sqrt{(3/2)^2 + (4)^2 + (3\sqrt{3})^2} = 5$$
$$x^\top y = 3 - 2\sqrt{3} \approx -0.46410$$



Entonces, tenemos

$$x^\top y = \cos(\alpha)\|x\|_2\|y\|_2 \quad \rightarrow \quad \alpha \approx 1.85773 \text{ rad}$$

Norma de la suma y la diferencia de vectores

Sean $x, y \in \mathbb{R}^n$, entonces

$$\begin{aligned} \|x+y\|_2^2 &= (x+y)^\top (x+y) \\ &= x^\top x + 2x^\top y + y^\top \\ &= \|x\|_2^2 + 2x^\top y + \|y\|_2^2 \end{aligned}$$

$$\begin{aligned} \|x-y\|_2^2 &= (x-y)^\top (x-y) \\ &= x^\top x - 2x^\top y + y^\top \\ &= \|x\|_2^2 - 2x^\top y + \|y\|_2^2 \end{aligned}$$

Desigualdad de Cauchy-Schwarz

Sean $x, y \in \mathbb{R}^n$. Entonces,

$$|x^\top y| \leq \|x\|_2 \|y\|_2$$

Desigualdad de Chebyshev

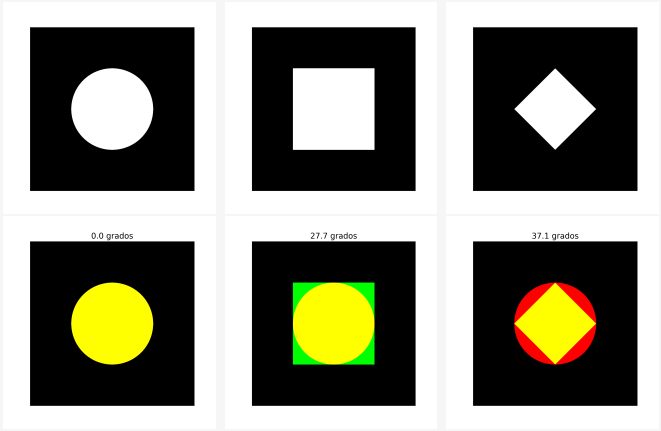
Sea un vector $x \in \mathbb{R}^n$ tal que un número k del valor absoluto de sus componentes son mayores que un número $a > 0$, esto es

$$|x_i| \geq a > 0, \quad \text{para } i \in \mathcal{I}_k$$

donde $\mathcal{I}_k \subset \{1, 2, ..., n\}$ y card$(\mathcal{I}_k) = k$. Entonces,

$$\|x\|_2^2 \leq ka^2$$

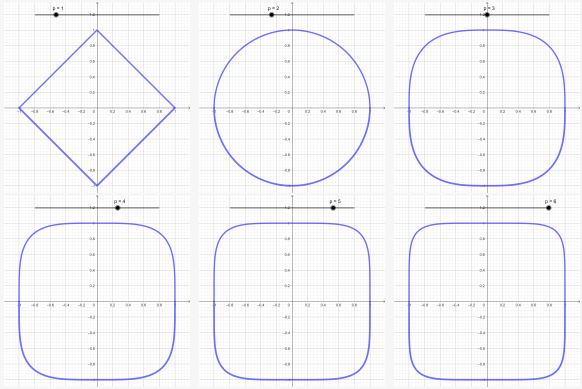
Ilustración. Producto interior entre imágenes.



Ejemplo de normas

Sea $x \in \mathbb{R}^n$ y $1 \leq p < \infty$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$



Definiciones de media cuadrática y distancia

La **media cuadrática** (o root mean square) de un vector $x \in \mathbb{R}^n$ se define por

$$\mathrm{rms}(x) := \sqrt{\frac{x_1^2 + x_2^2 + ... + x_n^2}{n}} = \frac{\|x\|_2}{\sqrt{n}}$$

Definimos la **distancia Euclidiana** entre dos vectores $x, y \in \mathbb{R}^n$ por:

$$\mathrm{dist}(x, y) := \|x - y\|_2$$

Además, definimos la **desviación cuadrática media** por

$$\mathrm{rms}(x - y) = \frac{\|x - y\|_2}{\sqrt{n}}$$

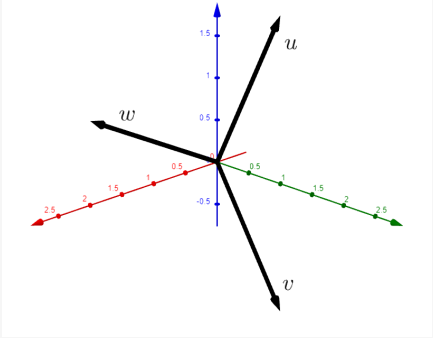
Ejemplo. Distancia entre vectores

Considere los vectores

$$u = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad v = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \quad w = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

Las respectivas distancias son

- $\text{dist}(u, v) = \|u - v\|_2 = \sqrt{11}$
- $\text{dist}(u, w) = \|u - w\|_2 = \sqrt{6}$
- $\text{dist}(w, v) = \|w - v\|_2 = \sqrt{9}$



Ejemplo. Interpretación de distancia

Distancia de features o atributos. Si $x, y \in \mathbb{R}^n$ son vectores que representan atributos, entonces dist(x, y) indica que tan similares o que tan cercanos son los vectores de atributos.
Vecino mas cercano. Suponga que $z_1, ..., z_m \in \mathbb{R}^n$. Decimos que z_j es el vecino mas cercano de $x \in \mathbb{R}^n$ dentro de la colección de vectores $\{z_i\}$, $1 \leq i \leq m$, si

$$\|x - z_j\|_2 \leq \|x - z_i\|_2, \quad \text{para } 1 \leq i \leq n.$$

(Di)Similitud de documentos. Si $x, y \in \mathbb{R}^n$ son vectores que representan histogramas de ocurrencias de palabras en dos documentos. Entonces, dist(x, y) representa una medida de disimilitud de los documentos.

Observación

Si las componentes de un vector representan distintos tipos de cantidades debemos tener cuidado al calcular la distancia entre estos vectores. Si queremos que las componentes de un vector tengan el mismo peso al determinar la distancia, entonces sus valores numéricos deben ser aproximadamente de la misma magnitud.

Ejemplo

Sean $x, y, z \in \mathbb{R}^2$, vectores que represetan atributos de 3 casas vendidas (área, número de dormitorios). Primero, expresemos el área en miles de sq.ft, así los vectores son:

$$x = \begin{bmatrix} 1.62 \\ 2 \end{bmatrix}, \quad y = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}, \quad z = \begin{bmatrix} 1.6 \\ 4 \end{bmatrix}$$

de donde podemos calcular $\|x\|_2 \approx 2.57, \quad \|x - y\|_2 = 0.12, \quad \|x - z\|_2 \approx 2$
Desde aquí concluiríamos que el vector x es mas cercano al vector y que al vector z, y así los atributos de las casa x e y son mas parecidos que los de x y z. Por otro lado, si expresamos el área de las casa en sq. ft. tenemos que

$$\|x\| \approx 1600, \quad \|x - y\|_2 = 1000, \quad \|x - z\|_2 = 2$$

de donde concluímos exactamente lo opuesto.

Definición promedio y desviación estándar.

Se define el **promedio** de un vector $x \in \mathbb{R}^n$ por

$$\mathsf{avg}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La **desviación estándar** del vector x se define por:

$$\mathsf{std}(x) = \mathsf{rms}(x - \mathsf{avg}(x)[1, ..., 1]^\top) = \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(x_i - \mathsf{avg(x)})^2}$$

Propiedad (Ejercicio). Muestre que, para $x \in \mathbb{R}^n$,

$$\mathrm{rms}(x)^2 = \mathrm{avg}(x)^2 + \mathrm{std}(x)^2.$$

UC | Chile

Tema 3: Clustering

Clustering

Supongamos que estamos interesados en agrupar una colección de datos representados por vectores reales de tamano n en grupos o clusters que esten cerca los unos de los otros. A esto le llamamos clustering.

Objetivo: dados N vectores reales de tamaño n queremos asignarlos a k grupos o cluster con k obviamente mucho menor que N

$$\{x_1, x_2, x_3, ..., x_{N-2}, x_{N-1}, x_N\} \quad \Longrightarrow \quad \begin{cases} \{x_1, x_{12}, x_{40}, ...\} \\ \{x_2, x_6, x_{14}, ...\} \\ \dots \\ \{x_N, x_5, x_{10}, ...\} \end{cases} \quad k - \text{clusters}$$

Ilustración gráfica clustering

Presentamos una interpretación grafica usando puntos en dos dimensiones para representar determinados datos.

Aplicaciones

Descubrimiento de un tema central en un texto. Si los vectores con los que contamos representan histogramas de palabras que aparecen en n documentos. Entonces, podemos interpretar el agrupar los documentos como descubrir temas en común entre estos.

Clasificación de pacientes de un hospital. Si los vectores representan los datos clínicos de los pacientes de un hospital, nos puede interesar clasificarlos para asignarlos a alguna unidad de cuidado.

Determinar patrones climáticos. Identificar variaciones en las condiciones del clima, temperatura, velocidad de viento, dirección del viento, presión, humedad, etc.

Conceptos de clustering

Asignaciones. Las asignaciones de cada vector x_i a un cluster la registramos mediante un vector de etiquetas c de tamaño N. Este vector registra en su componente i-esima el número de cluster j al cual pertenece el vector x_i. Además, definimos el conjunto de los índices correspondientes a cada cluster por:

$$G_i = \{1 \leq i \leq N : \quad c_i = j\}, \quad \text{para } 1 \leq j \leq k$$

Ilustración de asignación

Considere como ilustración un conjunto de 7 vectores los cuales tienen el siguiente de vector de etiquetas

$$c = [3, 1, 1, 2, 1, 2, 3]^\top$$

Entonces, el conjunto de las etiquetas son, aquí representado por vectores

$$G_1 = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}, \quad G_3 = \begin{bmatrix} 1 \\ 7 \end{bmatrix}$$

Conceptos de clustering

Representantes. En cada uno de los grupos definimos un vector representante. Los representantes de cada uno de los clusters los denotamos por

$$\{z_1, z_2, ..., z_k\}$$

Idealmente, queremos que cada representante este cerca de los vectores de su grupo, es decir, matemáticamente queremos que las siguientes normas sean pequeñas

$$\|x_i - z_{c_i}\|, \quad 1 \leq i \leq k$$

Objetivo matemático del cluster

Queremos escoger un vector de asignaciones o etiquetas $c = [c_1, ..., c_N]^\top$ y una elección de vectores representantes $\{z_1, ..., z_k\}$ de los grupos que minimicen la función objetivo del cluster

$$J^{\text{cluster}} = \frac{1}{N}\left(\|x_1 - z_{c_1}\|^2 + \|x_2 - z_{c_2}\|^2 + ... + \|x_N - z_{c_N}\|^2\right)$$

Al cluster o asignación que minimiza esta función lo llamamos cluster optimo. Encontrar la solución de este problema en la práctica debido a la alta dimensionalidad del problema. En la práctica se usan algoritmos heurístico como por ejemplo el algoritmo de k-means, el cual requiere considerablemente menos cómputos y encuentra una solución cercana al optimo. Llamamos a estos clusters, solución encontrada por los algoritmos, sub-óptimos.

Derivación del algoritmo.

Dividimos el análisis en dos tareas:

1. El problema de encontrar el vector de asignación dado representantes fijos. Suponga que ya contamos con los k representantes $\{z_1, z_2, ..., z_k\}$ de los k-clusters y buscamos el vector de asignaciones o etiquetas $c = [c_1, c_2, ..., c_N]^\top$ que dan el menor valor posible de la función de objetivo. Este problema si podemos resolverlo de forma exacta, debemos simplemente asignar a cada vector x su representante más cercano, esto calculando las distancias de este a los representantes y eligiendo la mínima.
2. El problema de encontrar los representantes dado un vector de etiquetas fijas. Si asumimos que el vector de asignación $c = [c_1, c_2, ..., c_N]^\top$ es dado, podemos encontrar los k-representantes $\{z_1, z_2, ..., z_k\}$ simplemente calculando un vector promedio de los vectores con la misma etiqueta.

Algoritmo K-means

Algorithm k-means (1957, S. Lloyd)

Input: Vectores $\{x_1, x_2, ..., x_N\}$, $x_i \in \mathbb{R}^n$, y k-representantes iniciales $\{z_1, ..., z_k\}$.
Output: Lista de representantes $\{z_1, ..., z_k\}$ y vector de etiquetas $c = [c_1, ..., c_N]^\top$

Repita los siguientes pasos hasta que la asignación del vector c se repita:

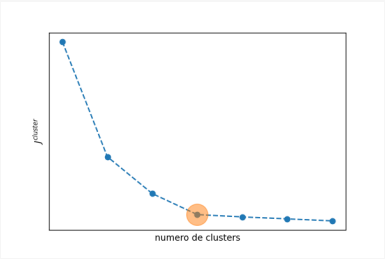
1. Calcule el vector de etiquetas, asignando cada vector x a su representante mas cercano.

2. Actualice los representantes calculando el vector promedio dentro de los vectores con la misma etiqueta.

Número óptimo de clusters

Como elegir el número de clusters o grupos óptimo k para un conjunto de datos?

Lo estándar es calcular el algoritmo de k-means para distintos valores de k y comparar los resultados. En general podemos decir que si el valor de la función objetivo J^{cluster} para un número determinado de clusters k es mucho menor que el valor para números de clusters menor que k y no significativamente mas grande que el valor para un numero mayor de clusters entonces podemos concluir que los datos se particionan adecuadamente en k grupos.



UC | Chile