Escalamiento Multidimensional

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 13) 21.febrero.2022

Escalamiento multidimensional

Dada una matriz de datos $\mathbf{x} \in \mathbb{R}^{n\times d}$, $n > d$, asociamos a cada vector $\mathbf{x}_i \in \mathbb{R}^d$ de la matriz, un representante $\mathbf{x}_i^* \in \mathbb{R}^r$, de modo que

$$\min_{\mathbf{x}_i^*,\mathbf{x}_j^*} \sum_{i=1}^{n}\sum_{j=1}^{n} \Big(d(\mathbf{x}_i,\mathbf{x}_j)^2 - d(\mathbf{x}_i^*,\mathbf{x}_j^*)^2\Big)^2. \tag{1}$$

Consideremos las matrices de distancias al cuadrado

$$\mathbb{D}^2 = \big(d(\mathbf{x}_i,\mathbf{x}_j)^2\big), \;\; \mathbb{D}^{*2} = \big(d(\mathbf{x}_i^*,\mathbf{x}_j^*)^2\big) \in \mathbb{R}^{n\times n}.$$

Con esta notación, la ecuación (1) se escribe como

$$\min_{\mathbf{x}_i^*,\mathbf{x}_j^*} ||\mathbb{D}^2 - \mathbb{D}^{*2}||_F^2.$$

Escalamiento multidimensional

Además, consideramos las matrices de Gram

$$\mathbb{G} = \mathbb{X}\mathbb{X}^T = \left(\mathbf{x}_i^T\mathbf{x}_j\right) = \left(\langle \mathbf{x}_i, \mathbf{x}_j\rangle\right), \;\; \mathbb{G}^* = \mathbb{X}^*(\mathbb{X}^*)^T = \left((\mathbf{x}_i^*)^T\mathbf{x}_j^*\right) = \left(\langle \mathbf{x}_i^*, \mathbf{x}_j^*\rangle\right) \in \mathbb{R}^{n\times n}.$$

Tenemos una relación entre distancias y productos internos:
Denotamos $g_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j\rangle = \mathbf{x}_i^T\mathbf{x}_j$. Entonces,

$$\begin{aligned} d_{ij}^2 &= ||\mathbf{x}_i - \mathbf{x}_j||^2 = \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j\rangle = \langle \mathbf{x}_i, \mathbf{x}_i\rangle - 2\langle \mathbf{x}_i, \mathbf{x}_j\rangle + \langle \mathbf{x}_j, \mathbf{x}_j\rangle \\ &= g_{ii} - 2g_{ij} + g_{jj}. \end{aligned}$$

Recordemos que si $\mathbb{J} = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, entonces $\mathbb{J}$ es una matriz de proyección, y $\mathbb{X}_c = \mathbb{J}\mathbb{X}$ es la matriz de datos centrados.

Luego, $\mathbb{G} = \mathbb{X}\mathbb{X}^T \;\Rightarrow\; \mathbb{G}_c = \mathbb{X}_c\mathbb{X}_c^T = (\mathbb{J}\mathbb{X})(\mathbb{J}\mathbb{X})^T = \mathbb{J}\mathbb{X}\mathbb{X}^T\mathbb{J}^T = \mathbb{J}\mathbb{X}\mathbb{X}^T\mathbb{J} = \mathbb{J}\mathbb{G}\mathbb{J}$.

Similarmente, $d_{ij}^{*2} = g_{ii}^* - 2g_{ij}^* + g_{jj}^*$.

Observe que centrar los datos es una transformación rígida, esto es, preserva distancias. Luego,

$$d_{ij}^2 = g_{ii}^c - 2g_{ij}^c + g_{jj}^c.$$

Luego, sumando sobre i, y sumando sobre j, respectivamente, se tiene

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n d_{ij}^2 &= \frac{1}{n}\sum_{i=1}^n g_{ii}^c - \underbrace{\frac{2}{n}\sum_{i=1}^n g_{ij}^c}_{=0} + \frac{1}{n}\sum_{i=1}^n g_{jj}^c = \frac{1}{n}\sum_{i=1}^n g_{ii}^c + g_{jj}^c,\\
\frac{1}{n}\sum_{j=1}^n d_{ij}^2 &= \frac{1}{n}\sum_{j=1}^n g_{ii}^c - \underbrace{\frac{2}{n}\sum_{j=1}^n g_{ij}^c}_{=0} + \frac{1}{n}\sum_{j=1}^n g_{jj}^c = \frac{1}{n}\sum_{j=1}^n g_{jj}^c + g_{ii}^c.
\end{aligned}$$

Escalamiento multidimensional

Juntando ambas ecuaciones, resulta

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n} d_{ij}^2\Big) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n} g_{jj}^c + g_{ii}^c\Big)\\ &= \frac{1}{n}\sum_{j=1}^{n} g_{jj}^c + \frac{1}{n}\sum_{i=1}^{n} g_{ii}^c = \frac{2}{n}\sum_{i=1}^{n} g_{ii}^c.\end{aligned}$$

Denotando $a_{ij} = \frac{1}{2}d_{ij}^2$, $a_{i.} = \frac{1}{n}\sum_j a_{ij}$, $a_{.j} = \frac{1}{n}\sum_i a_{ij}$ y $a_{..} = \frac{1}{n^2}\sum_{i,j} a_{ij}$, se muestra que

$$-2g_{ij}^c = a_{ij} - a_{i.} + a_{.j} + a_{..},$$

$g_{ij} = -\frac{1}{2}(a_{ij} - a_{i.} + a_{.j} + a_{..})$.

Escalamiento multidimensional

En notación matricial, esto es $\mathbb{G}^c = -\frac{1}{2}\mathbb{J}\mathbb{D}^2\mathbb{J}$.

Así, en lugar de resolver el problema de optimización (1)

$$\min_{\mathbf{x}_i^*} \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(d(\mathbf{x}_i,\mathbf{x}_j)^2 - d(\mathbf{x}_i^*,\mathbf{x}_j^*)^2\Big)^2 = \min_{\mathbf{x}_i^*} ||\mathbb{D}^2 - \mathbb{D}^{*2}||_F^2.$$

podemos resolver el problema equivalente

$$\min_{\mathbf{x}_i^*} ||\tfrac{1}{2}\mathbb{J}(\mathbb{D}^2 - \mathbb{D}^{*2})\mathbb{J}||_F^2 = \min_{\mathbf{x}_i^*} ||\mathbb{G}^c - \mathbb{G}^{*c}||_F^2.$$

Esta última ecuación corresponde a encontrar la matriz $\mathbb{G}^{*c}$ de rango $1 \leq r \leq d$ que mejor aproxima $\mathbb{G}^c$:

$$\min_{\mathbb{G}^*\succeq 0,\ rank(\mathbb{G}^*)=r} ||\mathbb{G}^c - \mathbb{G}^{*c}||_F^2.$$

Escalamiento multidimensional

Por el Teorema de Eckart-Young, la solución a este problema está dada de la siguiente forma: Si

$$\mathbb{G}^c = USV^T = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{u}_i^T,$$

es la descomposición SVD de $\mathbb{G}^c$, entonces $\mathbb{G}^{*c}$ es

$$\mathbb{G}^{*c} = U_r S_r V_r^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{u}_i^T.$$

¿Para qué se hace esto?

- No siempre es posible representar datos como vectores.
- Más adelante vamos a hacer el análisis sin referirnos explícitamente a los $\mathbf{x}_i$. En lugar de ello, usaremos distancias o algún otro tipo de métrica.

Escalamiento multidimensional

Objetivo: Crear coordenadas (sintéticas) en los datos, a partir una matriz de distancias.

Receta para hacer escalamiento multidimensional:

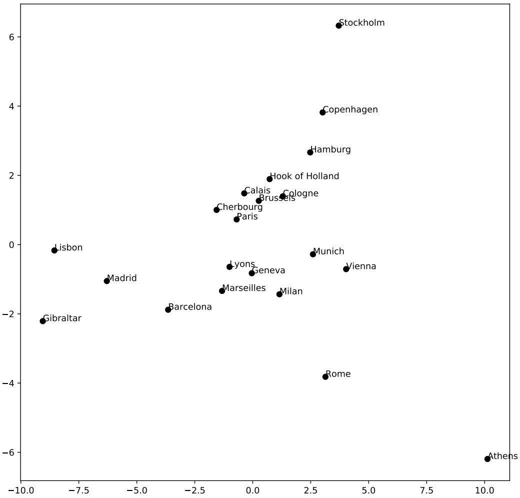
1. Dada una matriz de distancias $\mathbb{D} \in \mathbb{R}^{n \times n}$, simétrica, entre n datos.
2. Calcular la matriz de productos internos $\mathbb{G}^c = -\frac{1}{2}\mathbb{J}\mathbb{D}\mathbb{J}$, con $\mathbb{J} = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.
3. Hallar la descomposición SVD de $\mathbb{G}^c$
$$\mathbb{G}^c = U\Sigma V^T.$$
4. Si queremos representar los datos como vectores en $\mathbb{R}^k$, con $1 \leq k \leq n$, tomamos la proyección de $\mathbb{D}$ generada por las primeras k columnas de V:
$$\mathbb{X} = \mathbb{D}\, V[:, : k].$$

Ejemplo

Ejemplo: Distancias entre 21 ciudades europeas (en Km).

	Athens	Barcelona	Brussels	Calais	Cologne	Copenhagen	...
Athens	0	3313	2963	3175	2762	3276	...
Barcelona	3313	0	1318	1326	1498	2218	...
Brussels	2963	1318	0	204	206	966	...
Calais	3175	1326	204	0	409	1136	...
Cologne	2762	1498	206	409	0	760	...
Copenhagen	3276	2218	966	1136	760	0	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

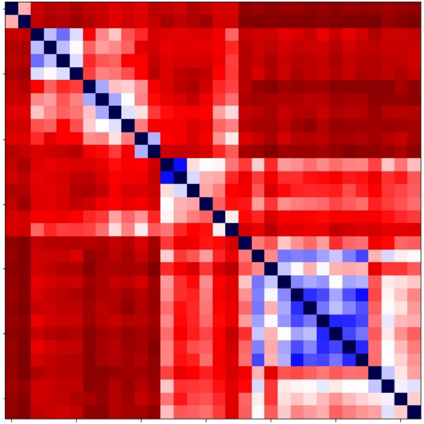
Ejemplo



Stockholm
Copenhagen
Hamburg
Hook of Holland
Calais
Cologne
Brussels
Cherbourg
Paris
Lisbon
Munich
Lyons
Geneva
Vienna
Madrid
Marseilles
Milan
Barcelona
Gibraltar
Rome
Athens
6
4
2
0
−2
−4
−6
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

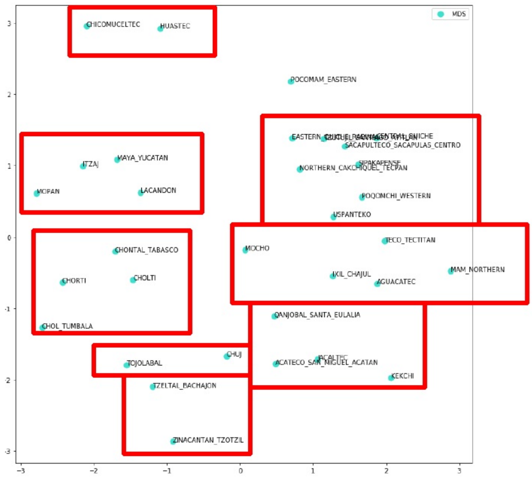
Ejemplo

Ejemplo: Idiomas mayas



Matriz de distancias entre idiomas mayas.

Ejemplo



Escalamiento multidimensional a 2 dimensiones.

UNIVERSIDAD DEL VALLE DE GUATEMALA