UFM UNIVERSIDAD FRANCISCO MARROQUÍN
VERITAS • LIBERTAS • JVSTITIA

FACULTAD de CIENCIAS ECONÓMICAS

Escalamiento Multidimensional

Alan Reyes-Figueroa

Elements of Machine Learning (Aula 09) 16.febrero.2023

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 $\mathbf{x}_i^* \in \mathbb{R}^r$,(estamos proyectando a un espacio de dimensión menor) de modo que en el nuevo espacio, se preserven las distancias lo mejor posible.

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

Consideremos las matrices de distancias al cuadrado

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

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

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

Además, consideramos las matrices de Gram

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

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

Escalamiento multidimensional

Recordemos que la matriz de proyección para un conjunto de n datos, $\mathbb{J}$, está dada por

$$\mathbb{J} = I_n - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T.$$

Esto es

$$\mathbb{J} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix}.$$

Esta matriz sirve para

- centrar la matriz de datos $\mathbb{X}$, mediante el cálculo $\mathbb{X}_c = \mathbb{J}\,\mathbb{X}$.
- centrar la matriz de Gram $\mathbb{G}$, mediante el cálculo $\mathbb{G}_c = \mathbb{J}\,\mathbb{G}\,\mathbb{J}$.

Escalamiento multidimensional

Si queremos proyectar a un espacio de dimensión r, $0 < r \leq d$, 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:

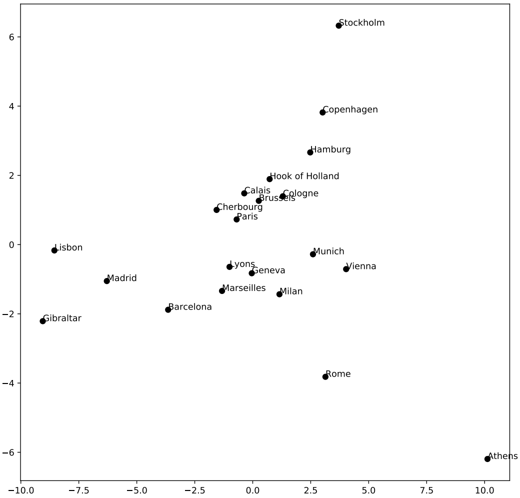
1. Dada una matriz de distancias $\mathbb{D} \in \mathbb{R}^{n \times n}$, simétrica, entre n datos.
2. Calcular la matriz de productos internos $\mathbb{G}_c = -\frac{1}{2}\mathbb{J}\mathbb{D}\mathbb{J}$, con $\mathbb{J} = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.
3. Hallar la descomposición SVD de $\mathbb{G}_c$
$$\mathbb{G}_c = U\Sigma V^T.$$
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	$\dots$
Athens	0	3313	2963	3175	2762	3276	$\dots$
Barcelona	3313	0	1318	1326	1498	2218	$\dots$
Brussels	2963	1318	0	204	206	966	$\dots$
Calais	3175	1326	204	0	409	1136	$\dots$
Cologne	2762	1498	206	409	0	760	$\dots$
Copenhagen	3276	2218	966	1136	760	0	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

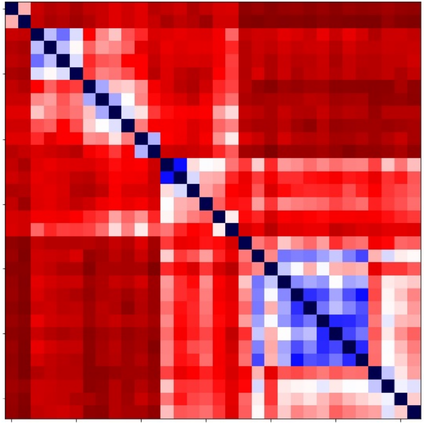
Ejemplo



UFM UNIVERSIDAD FRANCISCO MARROQUÍN

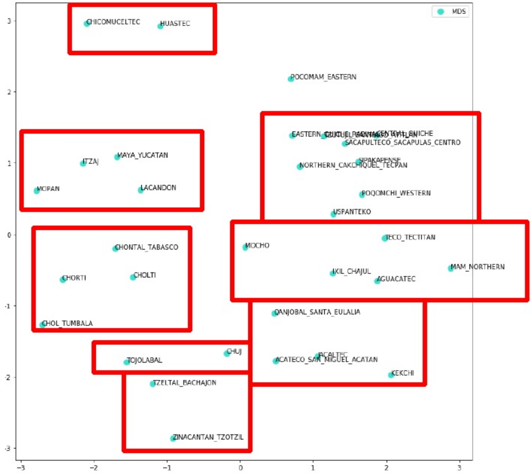
Ejemplo

Ejemplo: Idiomas mayas



Matriz de distancias entre idiomas mayas.

Ejemplo



Escalamiento multidimensional a 2 dimensiones.

UFM UNIVERSIDAD FRANCISCO MARROQUÍN