Datos Masivos II

Proyecciones aleatorias

Blanca Vázquez-Gómez y Gibran Fuentes-Pineda

Octubre 2020

- Es un método sencillo y eficiente de reducir dimensiones.
- Las direcciones de las proyecciones son independientes de los datos
- Preservan las distancias entre cualquier par de datos de forma aproximada.

- Lema de Johnson-Lindenstrauss: un conjunto de datos en un espacio euclidiano de altas dimensiones puede proyectarse a un espacio de menores dimensiones con una distorsión controlada de sus distancias

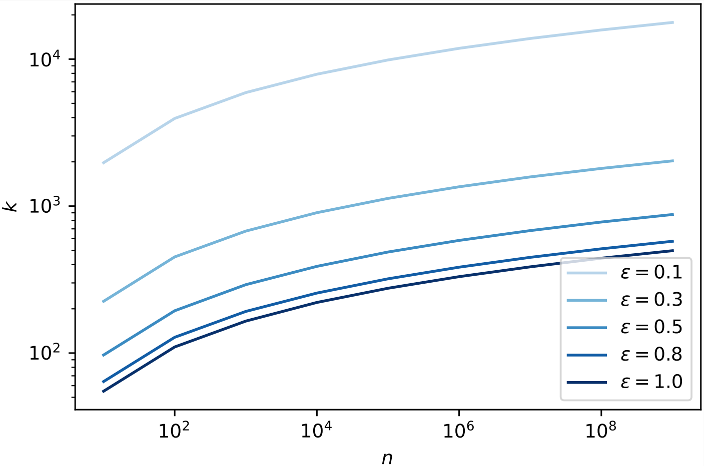
- *Dado* $\epsilon \in (0,1)$, $X = \{x^{(1)}, \ldots, x^{(n)}\} \in \mathbb{R}^d$ *y* $k > 9\epsilon^{-2}\log(n)$, *hay una proyección* $\mathcal{PA} : \mathbb{R}^d \to \mathbb{R}^k$ *tal que*

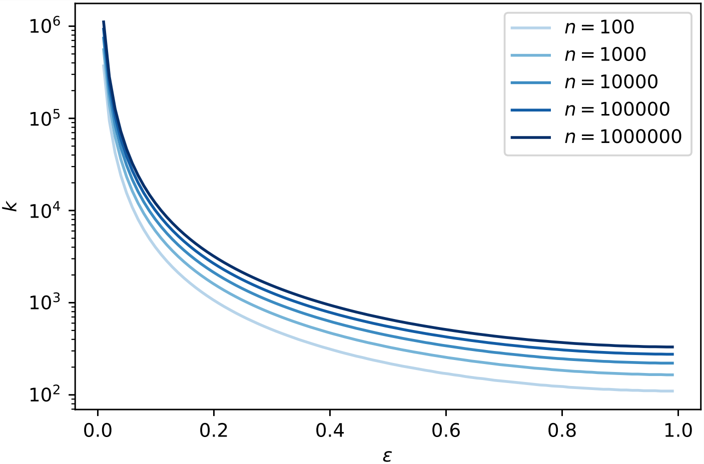
$$1 - \epsilon \leq \frac{\|\mathcal{PA}(x^{(i)}) - \mathcal{PA}(x^{(j)})}{\|x^{(i)} - x^{(j)}\|^2}\| \leq 1 + \epsilon$$

- El número mínimo k para garantizar esto es

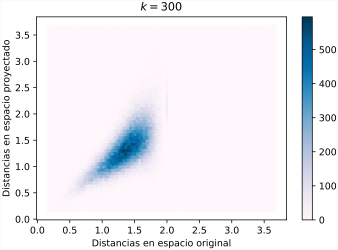
$$k = \geq \left(\frac{4 \cdot \log(n)}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \right)$$

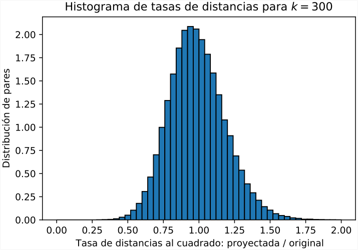
Cotas: n vs k


10^4
10^3
10^2
k
10^2
10^4
10^6
10^8
n
$\varepsilon = 0.1$
$\varepsilon = 0.3$
$\varepsilon = 0.5$
$\varepsilon = 0.8$
$\varepsilon = 1.0$

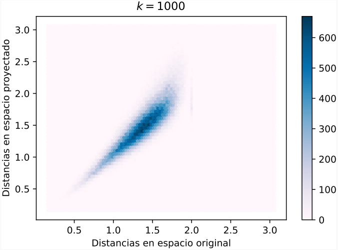

$n = 100$
$n = 1000$
$n = 10000$
$n = 100000$
$n = 1000000$
10^6
10^5
10^4
10^3
10^2
k
0.0
0.2
0.4
0.6
0.8
1.0
ε

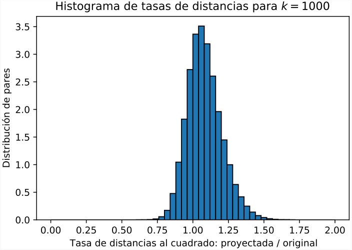
Distorción: $k = 300$





Distorción: $k = 1000$





- Dado $\mathsf{X} = \{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(n)}\} \in \mathbb{R}^d$, una proyección aleatoria se define por una matriz aleatoria A de $k \times d$
- La matriz A se puede generar con una distribución gaussiana $\mathcal{N}(0, \frac{1}{k})$ de tal forma que satisfaga las siguientes propiedades
  - Simetría esférica: para cualquier matriz ortogonal $\mathsf{R} \in O(d)$, AR y A tienen la misma distribución
  - Ortogonalidad: Las filas de A son ortogonales
  - Normalidad: Las filas de A son vectores unitarios

Proyección aleatoria dispersa

- Reducción de dimensiones usando una matriz aleatoria dispersa
- Más rápido y eficiente en memoria
- La matriz aleatoria se construye sacando muestras de

$$\begin{cases} -\sqrt{\frac{s}{k}} \text{ con probabilidad } \frac{1}{2s} \\ 0 \text{ con probabilidad } 1 - \frac{1}{s} \\ +\sqrt{\frac{s}{k}} \text{ con probabilidad } \frac{1}{2s} \end{cases}$$

donde $s = \frac{1}{densidad}$