

DATOS MASIVOS II

PROYECCIONES ALEATORIAS

Blanca Vázquez-Gómez y Gibran Fuentes-Pineda

29 de agosto de 2022

- 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.

EL LEMA DE JOHNSON-LINDENSTRAUSS

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

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

- 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

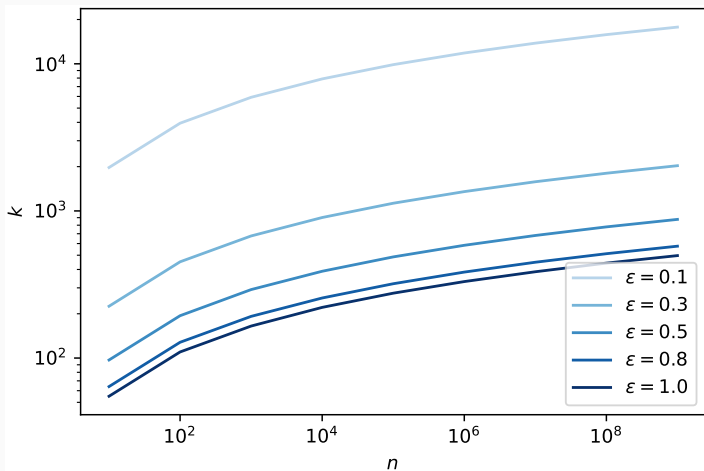


Imagen generada con ejemplo de sklearn

COTAS: ϵ VS k

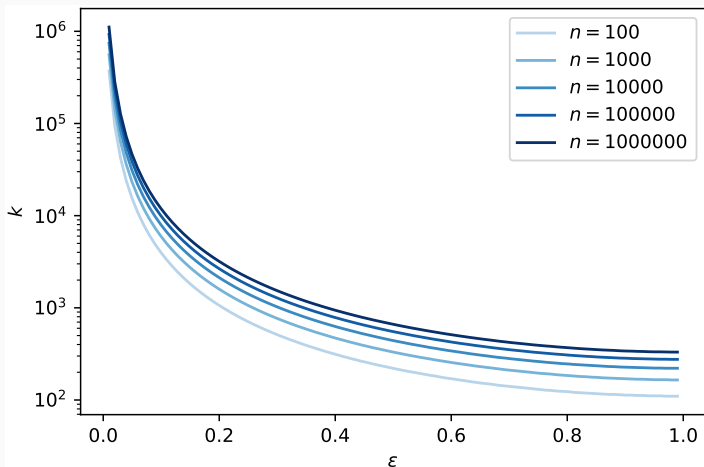
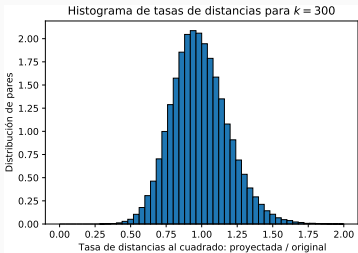
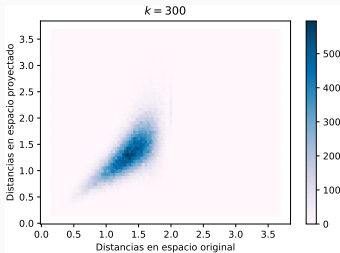


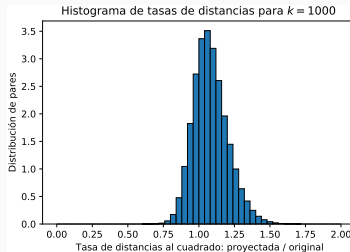
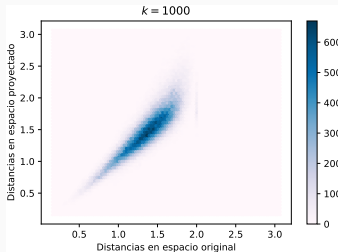
Imagen generada con ejemplo de sklearn

DISTORSIÓN: $k = 300$



Imágenes generadas con ejemplo de sklearn

DISTORCIÓN: $k = 1000$



Imágenes generadas con ejemplo de sklearn

- Dado $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \in \mathbb{R}^d$, una proyección aleatoria se define por una matriz aleatoria \mathbf{A} de $k \times d$
- La matriz \mathbf{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 $\mathbf{R} \in O(d)$, \mathbf{AR} y \mathbf{A} tienen la misma distribución
 - Ortogonalidad: Las filas de \mathbf{A} son ortogonales
 - Normalidad: Las filas de \mathbf{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

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

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