Johnson-Lindenstrauss Lemma[1]

Vázquez Choreño Luis Ernesto

Abril 21, 2020

Reducción de dimensiones

Dados $N >= 2$ distintos vectores, cada uno perteneciente a $\mathbb{R}^d$. Si la dimensión d es larga puede ser costoso guardar o manipular los datos. La idea de reducción de dimensiones es mediante un mapeo $F\colon \mathbb{R}^d \to \mathbb{R}^m$ con dimensión de proyección m sustancialmente menor que d, preservando características esenciales del conjunto de datos.

Para este ejemplo se va a considerar preservar las distancias a pares, especificamente se desea tener una función de mapeo F tal que se garantice que dado un $\delta \in (0, 1)$

$$(1-\delta) <= \frac{\|F(u^i) - F(u^j)\|_2^2}{\|u^i - u^j\|_2^2} <= (1+\delta) \tag{1}$$

Requerimientos

Variable aleatoria sub-exponencial

Una variable aleatoria X con $\mu = \mathbb{E}[X]$ es sub-exponencial si existe dos parámetros no negativos (v, α) tal que

$$\mathbb{E}[e^{\lambda(X-\mu)}] <= e^{\frac{v^2\lambda^2}{2}} \qquad \forall |\lambda| < \frac{1}{\alpha} \tag{2}$$

Ejemplo

Sea $Z \sim \mathcal{N}(0, 1)$, considerar una variable aleatoria $X = Z^2$ para $\lambda < \frac{1}{2}$ tenemos

$$\mathbb{E}[e^{\lambda(X-1)}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} e^{\frac{-z^2}{2}} dz$$

$$= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}$$

Por lo tanto $X = Z^2$ es sub-exponencial asignando los parámetros

$$(v, \alpha) = (2, 4)$$

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} <= e^{2\lambda^2} = e^{\frac{4\lambda^2}{2}} \tag{3}$$

Y sea $Y = \sum_{k=1}^{n}(Z_k)^2$ cada Z_i independiente, entonces Y es también una variable sub-exponencial con parámetros $(v, \alpha) = (2\sqrt{n}, 4)$ y se tiene el siguiente tail bound:

$$\mathbb{P}[|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1| >= t] <= 2e^{\frac{-nt^2}{8}}, \quad \forall t \in (0,1) \tag{4}$$

Construcción

Para probar (1) se usará de un procedimiento aleatorio, se construirá una matriz $X \in \mathbb{R}^{mxd}$ construidos con entradas independientes $\mathcal{N}(0,1)$

Se define una función de mapeo $F\colon \mathbb{R}^d \to \mathbb{R}^m$ de la forma $u \to Xu/\sqrt{m}$

Para un vector específico $u \neq 0$ se define la variable Y donde cada uno de sus sumandos $\langle x_i, \frac{u}{\|u\|_2}\rangle$ es una distribución $\mathcal{N}(0,1)$ al cuadrado, la cual se demostro en (3) ser una variable sub-exponencial

$$Y = \frac{\|Xu\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m}\langle x_i, \frac{u}{\|u\|_2}\rangle^2$$

Por (4) tenemos que la distribución de Y cumple con

$$\mathbb{P}\left[\Big|\frac{\|Xu\|_2^2}{m\|u\|_2^2} - 1\Big| >= \delta\right] <= 2e^{\frac{-m\delta^2}{8}}, \quad \forall \quad \delta \in (0,1)$$

$$\mathbb{P}\left[\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin [(1-\delta),(1+\delta)]\right] <= 2e^{\frac{-m\delta^2}{8}}, \quad para \quad cualquier \quad u \neq 0 \quad \in \mathbb{R}^d$$

Aplicamos union bound para trabajar con todos los N pares de vectores

$$\mathbb{P}\left[\frac{\|F(u^i - u^j)\|_2^2}{\|u^i - u^j\|_2^2} \notin [(1-\delta),(1+\delta)] para \quad algun \quad u^i \neq u^j\right] <= 2\binom{N}{2}e^{\frac{-m\delta^2}{8}},$$

References

[1] Martin J. Wainwright. High-dimensional statistics.