Unidad 3: Búsqueda de pares similares

Funciones *hash* sensibles a la localidad

Gibran Fuentes Pineda

Abril 2021

Hashing sensible a la localidad (LSH)

- Método para realizar búsqueda del vecino más cercano aproximado en espacios de alta dimensionalidad.
- La idea es proyectar el espacio original a otro de mucho menores dimensiones que preserve las distancias entre los objetos de forma aproximada con alta probabilidad.
- Para ello se define una familia de funciones $\mathcal{H}$ sensibles a la localidad para una distancia $dist(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

Familias de funciones sensibles a la localidad

- Una familia de funciones $\mathcal{H} = \{h : \mathbb{R}^d \to \mathbb{Z}\}$ se llama *sensible a la localidad* para una distancia *dist* si para cualquier par $\{\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\} \in \mathbb{R}^d$, existen números reales r_1, r_2, p_1, p_2 tal que:

$$dist(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \leq r_1 \Rightarrow P[h(\mathbf{x}^{(i)}) = h(\mathbf{x}^{(j)})] \geq p_1$$
$$dist(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \geq r_2 \Rightarrow P[h(\mathbf{x}^{(i)}) = h(\mathbf{x}^{(j)})] \leq p_2$$

  donde $r_1 < r_2$.

- En general, es deseable que $p_1 \gg p_2$.

Tuplas de funciones LSH

- Para ampliar el margen entre p_1 y p_2, se generan l tuplas $g_1, \dots g_l$ de r funciones $hash$[1]:

$$g_1 = (h_{11}, \dots, h_{1r})$$
$$\vdots \qquad\qquad \vdots$$
$$g_l = (h_{l1}, \dots, h_{lr})$$

- Se pueden ver como una familia de funciones con $d_1, d_2, (p_1)^r, (p_2)^r$.
- Para buscar se construyen l tablas (una por tupla) y se almacena cada punto en la cubeta correspondiente.[2]

[1] Sacadas de forma independiente y uniforme de $\mathcal{H}$

[2] Esto se logra mediante una función *hash* universal que toma la tupla y la mapea a un índice de la tabla.

LSH para distancia de Hamming

- Para vectores binarios $\mathbf{x}^{(i)} \in \{0,1\}^d$ (o cadenas de bits de longitud d) y la distancia de Hamming, una familia LSH se construye obteniendo el valor de una posición j

$$h_j(\mathbf{x}^{(i)}) = x_j^{(i)}$$

- Más generalmente, esta familia de funciones se puede aplicar a vectores M-arios $\mathbf{x}^{(i)} \in \{0,1,\ldots,M\}^d$.

Extensión a distancia ℓ_1

- Sean $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ puntos en un espacio de d dimensiones y C el valor máximo de cualquier coordenada, cada punto se transforma a un vector de Cd bits:

$$f(\mathbf{x}^{(i)}) = [t(x_1); t(x_2); \cdots ; t(x_d)]$$

  donde $t(x_k)$ es una cadena de bits con x_k unos seguidos de $C - x_k$ ceros.

- La distancia de Hamming sobre $f(\mathbf{x}^{(i)})$ y $f(\mathbf{x}^{(j)})$ es igual a la distancia ℓ_1 sobre $\mathbf{x}^{(i)}$ y $\mathbf{x}^{(j)}$

LSH para distancias ℓ_s

- Se elige aleatoriamente una proyección de $\mathbb{R}^d$ sobre una línea, se desplaza por b y se córta en segmentos de tamaño w, esto es,

$$h_{\mathbf{a},b} = \left\lfloor \frac{\mathbf{a} \cdot \mathbf{x} + b}{w} \right\rfloor$$

  donde $b \sim \mathit{Unif}(0, w)$

- Si $\mathbf{a}$ se muestrea de una distribución normal se obtiene una familia LSH para la distancia ℓ_2.

- Si $\mathbf{a}$ se muestrea de una distribución de Cauchy se obtiene una familia LSH para la distancia ℓ_1

LSH para distancia angular

- Dado un par de puntos $\{\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\} \in \mathbb{R}^d$, el ángulo entre ellos se obtiene de la siguiente manera:

$$\theta(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \arccos\left(\frac{\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}}{\|\mathbf{x}^{(i)}\| \cdot \|\mathbf{x}^{(j)}\|}\right)$$

- Una familia LSH para la distancia angular $1 - \theta(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ es:

$$h_{\mathbf{v}}(\mathbf{x}^{(i)}) = signo(\mathbf{v} \cdot \mathbf{x}^{(i)})$$

  donde $\mathbf{v} \in \mathbb{R}^d$ es un vector aleatorio de tamaño unitario.

- Esta función se puede ver como dividir el espacio en 2 por un hiperplano elegido aleatoriamente

$$Pr[h_{\mathbf{v}}(\mathbf{x}^{(i)}) = h_{\mathbf{v}}(\mathbf{x}^{(j)}] = 1 - \frac{\theta(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{\pi}$$