Unidad 3: Búsqueda de pares similares

Funciones *hash* sensibles a la localidad

Gibran Fuentes Pineda

Marzo 2020

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 $\mathcal{H} = \{h : \mathbf{x}^d \to \mathcal{U}\}$ se llama *sensible a la localidad* para d 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 las siguientes dos propiedades se mantienen:

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

- Es deseable que $p_1 > p_2$ y $r_1 < r_2$.

Tuplas de funciones LSH

- Se amplía margen entre p_1 y p_2 generando 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 lograusando 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 $\{0,1\}^d$, una familia LSH se obtiene sacando un bit de forma aleatoria (independiente y uniforme)

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

- Esta familia de funciones se mantiene para vectores M-arios.

Extensión a distancia ℓ_1

- Sean $\{\mathbf{x}^1 \ldots, \mathbf{x}^1\}$ 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

- Elige aleatoriamente una proyección de $\mathbb{R}^d$ sobre una línea, desplázala por b y córtala en segmentos de tamaño w, esto es,

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

  donde $b \in [0, w)$
- Si $\mathbf{a}$ se muestrea de una distribución normal se obtiene una familia LSH para distancia ℓ_2.

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

LSH para distancia angular

- Para cualquier par de puntos $\{\mathbf{x}^{(i)}\mathbf{x}^{(j)} \in \mathbb{R}^d\}$

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

- Una familia LSH se forma eligiendo aleatoriamente un vector de tamaño unitario $u \in \mathbb{R}^d$ y $h_u(\mathbf{x}^{(i)}) = signo(u \cdot \mathbf{x}^{(i)})$
- Se puede ver como dividir el espacio en 2 por un hiperplano elegido aleatoriamente

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