Unidad 3: Búsqueda de pares similares

Resúmenes de conjuntos con preservación de similitud

Gibran Fuentes Pineda

Abril 2021

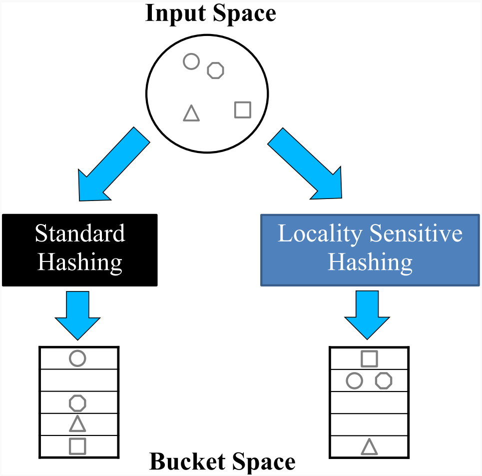
El problema del vecino más cercano aproximado

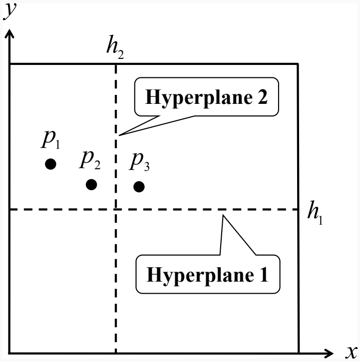
- Dado un conjunto de puntos $\mathcal{X} = \{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(n)}\}$ y un punto de consulta q, los cuales residen en un espacio de d dimensiones $\mathsf{x}^{(k)} \in \mathbb{R}^d, i = 1, \ldots, n$ bajo una función de distancia *dist*, encontrar los puntos en $\mathcal{X}$ que:

$$dist(\mathsf{q} - \mathsf{x}^{(k)}) \leq (1 + \epsilon) \cdot \min_{\mathsf{x}^{(j)} \in \mathcal{X}} dist(\mathsf{q} - \mathsf{x}^{(j)})$$

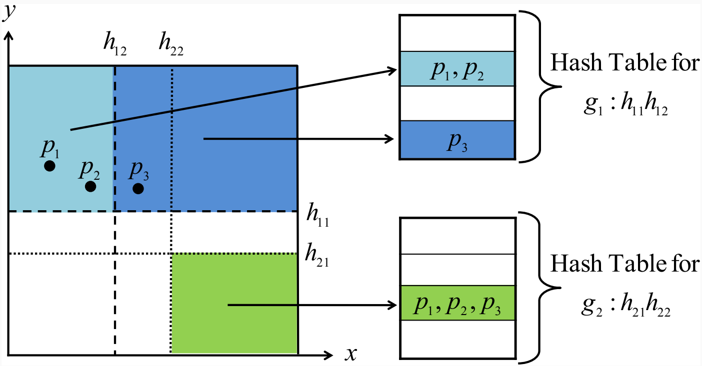
  donde $\epsilon > 0$ y $\mathsf{x}^{(j)}$ es el verdadero vecino más cercano de q.

Funciones de hash sensibles a la localidad (LSH)




y
h2
Hyperplane 2
p1
p2
p3
h1
Hyperplane 1
x
y
h12
h22
Second Hyperplane 2
p1
p2
p3
h11
h21
Second Hyperplane 1
x

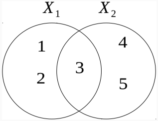
Búsqueda de pares similares



- Genera permutación aleatoria del conjunto universo $\mathbb{U}$
- Asigna a cada conjunto su 1er elemento bajo la permutación

$$h(\mathcal{C}^{(1)}) = min(\pi(\mathcal{C}^{(1)}))$$

- Ejemplo:



$$\pi_1 = \{2,4,5,3,1\} \longrightarrow (h_1(\mathcal{C}^{(1)}) = 2, h_1(\mathcal{C}^{(2)}) = 4)$$
$$\pi_2 = \{4,3,1,5,2\} \longrightarrow (h_2(\mathcal{C}^{(1)}) = 3, h_2(\mathcal{C}^{(2)}) = 4)$$
$$\pi_3 = \{3,1,4,2,5\} \longrightarrow (h_3(\mathcal{C}^{(1)}) = 3, h_3(\mathcal{C}^{(2)}) = 3)$$
$$\pi_4 = \{3,4,1,5,2\} \longrightarrow (h_4(\mathcal{C}^{(1)}) = 3, h_4(\mathcal{C}^{(2)}) = 3)$$

- Probabilidad de colisión de 2 conjuntos es igual a su **similitud de Jaccard**:

$$P[h(\mathcal{C}^{(1)}) = h(\mathcal{C}^{(2)})] = \frac{|\mathcal{C}^{(1)} \cap \mathcal{C}^{(2)}|}{|\mathcal{C}^{(1)} \cup \mathcal{C}^{(2)}|} \in [0,1]$$

- Considera los conjuntos

$$\mathcal{C}^{(1)} = \{1, 2, 5, 7, 9\}$$
$$\mathcal{C}^{(2)} = \{3, 4, 5, 8, 9\}$$
$$\mathcal{C}^{(3)} = \{2, 5, 7, 8\}$$

- Y las permutationes

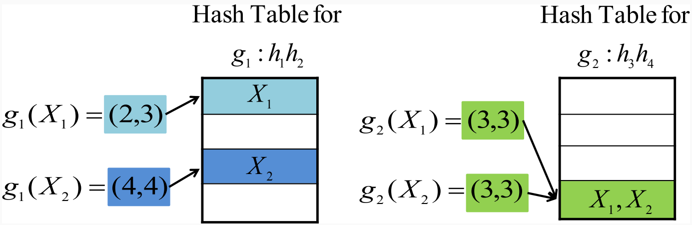
$$\pi_1 = \{5, 6, 9, 2, 3, 4, 8, 0, 7, 1\}$$
$$\pi_2 = \{3, 6, 0, 1, 8, 2, 7, 5, 4, 9\}$$
$$\pi_3 = \{9, 2, 6, 7, 4, 3, 1, 8, 5, 0\}$$
$$\pi_4 = \{2, 4, 0, 7, 9, 3, 8, 1, 5, 6\}$$

- Encuentre los valores MinHash para $\mathcal{C}^{(1)}$, $\mathcal{C}^{(2)}$ y $\mathcal{C}^{(3)}$

- Tuplas de valores MinHash

$$g_1(\mathcal{C}^{(1)}) = (h_1(\mathcal{C}^{(1)}), h_2(\mathcal{C}^{(1)}), \ldots, h_r(\mathcal{C}^{(1)}))$$
$$g_2(\mathcal{C}^{(1)}) = (h_{r+1}(\mathcal{C}^{(1)}), h_{r+2}(\mathcal{C}^{(1)}), \ldots, h_{2 \cdot r}(\mathcal{C}^{(1)}))$$
$$\ldots$$
$$g_l(\mathcal{C}^{(1)}) = (h_{(l-1)\cdot r+1}(\mathcal{C}^{(1)}), h_{(l-1)\cdot r+2}(\mathcal{C}^{(1)}), \ldots, h_{l \cdot r}(\mathcal{C}^{(1)}))$$

- Conjuntos con tupla idéntica se almacenan en el mismo registro



Probabilidad de colisión

- La probabilidad de que los valores MinHash de 2 conjuntos sean idénticos es

$$P[g_k(\mathcal{C}^{(1)}) = g_k(\mathcal{C}^{(2)})] = sim(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})^r$$

- La probabilidad de que no tengan ninguna tupla idéntica de l posibles es

$$P[g_k(\mathcal{C}^{(1)}) \neq g_k(\mathcal{C}^{(2)})] = (1 - sim(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})^r)^l, \forall k$$

- Por lo tanto la probabilidad de que 2 conjuntos tengan al menos una tupla idéntica es

$$P_{colisin}[\mathcal{C}^{(1)}, \mathcal{C}^{(2)}] = 1 - (1 - sim(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})^r)^l$$

Extensión a bolsas con multiplicidades enteras

- Cada bolsa $\mathcal{B}^{(i)}$ se convierte a un conjunto $\hat{\mathcal{C}}^{(i)}$, reemplazando cada multiplicidad con un elemento distinto
- El conjunto universal extentido sería

$$U_{ext} = \{1, \ldots, F_1, \ldots, F_1 + \cdots + F_{D-1} + 1, \ldots, F_1 + \cdots + F_D\}$$

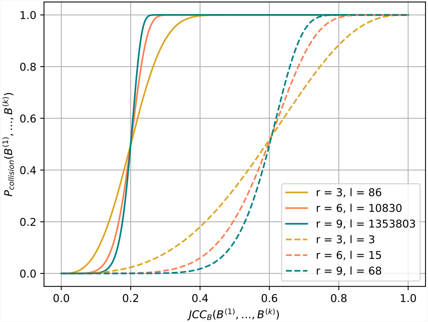
donde $F_1, \ldots, F_D$ son las multiplicidades máximas de los elementos $1, \ldots, D$

- Si aplicamos el esquema de MinHash a los conjuntos $\hat{\mathcal{C}}^{(i)}, \hat{\mathcal{C}}^{(j)} \subseteq U_{ext}$ se cumple

$$P[h(\hat{\mathcal{C}}^{(i)}) = h(\hat{\mathcal{C}}^{(j)})] = \frac{\sum_{w=1}^{D} \text{mín}(\mathcal{B}_w^{(i)}, \mathcal{B}_w^{(j)})}{\sum_{w=1}^{D} \text{máx}(\mathcal{B}_w^{(i)}, \mathcal{B}_w^{(j)})} = J_{\mathcal{B}}(\mathcal{B}_w^{(i)}, \mathcal{B}_w^{(j)})$$

- Dado r y un umbral de similitud η, el número de tuplas para aproximar un escalón unitario es

$$l = \frac{log(0.5)}{log(1-\eta^r)}$$



1. *Uniformidad*: Cada muestra (w, z_w) debe ser sacada aleatoriamente de forma uniforme de $\bigcup_{w=1}^{D} \{\{w\} \times [0, \mathcal{B}_w^{(i)}]\}$, es decir, la probabilidad de sacar w de $\mathcal{B}^{(i)}$ es proporcional a $\mathcal{B}_w^{(i)}$ y z_w está distribuido uniformemente.

2. *Consistencia*: Si $\mathcal{B}_w^{(j)} \leq \mathcal{B}_w^{(i)}, \forall w$, entonces cualquier muestra (w, z_w) sacada de $\mathcal{B}^{(i)}$ que satisface $z_w \leq \mathcal{B}_w^{(j)}$ también será una muestra de $\mathcal{B}^{(j)}$.

- Para cada elemento (no cero) de la bolsa $\mathcal{C}^{(i)}$
    1. $r_k \sim Gamma(2, 1)$
    2. $c_k \sim Gamma(2, 1)$
    3. $\beta_k \sim Unif(0, 1)$
    4. Calcula

$$t_k = \left\lfloor \frac{\ln \mathcal{C}_k^{(i)}}{r_k} + \beta_k \right\rfloor$$

$$y_k = e^{r_k \cdot (t_k - \beta_k)}$$

$$z_k = y_k \cdot e^{r_k}$$

$$a_k \frac{c_k}{z_k}$$

    5. $k^* = \arg\min_k \mathrm{a}$
- $\mathcal{C}_k^{(i)}$ es la multiplicidad del elemento k y puede ser entera o real.

- Particiones inducidas por Min-Hashing preservan relaciones de orden mayor dadas por el coeficiente de co-ocurrencia de Jaccard

$$JCC(\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(k)}) = \frac{|\mathcal{C}^{(1)} \cap \mathcal{C}^{(2)} \cap \cdots \cap \mathcal{C}^{(k)}|}{|\mathcal{C}^{(1)} \cup \mathcal{C}^{(2)} \cup \cdots \cup \mathcal{C}^{(k)}|}$$

- Particiones inducidas por Min-Hashing preservan relaciones de orden mayor dadas por el coeficiente de co-ocurrencia de Jaccard

$$JCC(\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(k)}) = \frac{|\mathcal{C}^{(1)} \cap \mathcal{C}^{(2)} \cap \dots \cap \mathcal{C}^{(k)}|}{|\mathcal{C}^{(1)} \cup \mathcal{C}^{(2)} \cup \dots \cup \mathcal{C}^{(k)}|}$$

- Una tupla de *hash* se puede ver como una partición del conjunto universo basada en la co-occurrencia de sus elementos