MLSEC
Similarity

Terminology

We recall some terminology:

Let S a finite set of samples, F a finite set of features and

$$V : S \times F \ni (s, f) \mapsto v_f(s) \in \mathbb{R}_0^+$$

a function; we will call $v_f(s)$ the **value** of the feature f for the sample s, $v(s) = [v_f(s)]_{f \in F}$ its **feature vector** and

$$F_s = \{f \in F \mid v_f(s) > 0\}$$

its **feature set**.

Example

Let S be some set of texts, F a set of words and $v_f(s)$ the number of occurences of f in the text s.

Example

- $s =$ John likes to watch movies. Mary likes movies too.
- $F = \{Anna, cinema, John, likes, Mary, movies, to, too, watch\}$.
- $v_{Anna}(s) = 0, v_{movies}(s) = 2, \ldots$
- $V(s, \cdot) = \{"Anna" : 0, "cinema" : 0, "John" : 1, "likes" : 2, "to" : 1, "watch" : 1, "movies" : 2, "Mary" : 1, "too" : 1\}$
- $v(s) = [0, 0, 1, 2, 1, 2, 1, 1, 1]$
- $F_s = \{John, likes, Mary, movies, to, too, watch\}$.

Options for measuring similarity/closeness

s, s' are close/similar if

- The vectors $[v_f(s)]_f$ and $[v_f(s')]_f$ are close, e.g. if

$$\sum_{f \in F} |v_f(s) - v_f(s')|$$

  is below a certain threshold.
- F_s and $F_{s'}$ are close.

The Jaccard index

Definition

Let $\mathcal{X}$ be non-empty. Let $A, B \subseteq \mathcal{X}$, not both empty. The **Jaccard index** $J(A, B)$ is defined by

$$J(A, B) = \frac{\#(A \cap B)}{\#(A \cup B)}$$

it measures to what extent the sets have elements in common.

Remarks

Jaccard metric

$$d_J(A,B) = 1 - J(A,B)$$

defines a metric on

$$\{Y \mid Y \subseteq \mathcal{X}, 0 < \#Y < \infty\}$$

Practical implementation

Two samples s, s' are similar, if

$$J(F_s, F_{s'}) \geq \alpha$$

for a threshold $\alpha \in (0,1)$, e.g. $\alpha = 0.8$.

MinHash 1

Let $H : \mathcal{X} \to \mathbb{Z}$ be a one-to-one function and let π be a permutation on $\mathcal{X}$.

Theorem

We have

$$J(A, B) = \frac{1}{m!} \#\{\pi \mid \min(H \circ \pi)(A) = \min(H \circ \pi)(B)\},$$

where $m = \#\mathcal{X}$.

MinHash 2

Example

- $\mathcal{X} = \{1, \ldots, 21\}$
- $A = \{1, 2, 4, \ldots, 10\}, B = \{2, 10\}$
- $J(A, B) = 2/9$
- $H = identity$
- $\pi = 1 \rightarrow 2 \rightarrow ... \rightarrow 21 \rightarrow 1, \min \pi(A) = 2 \neq 3 = \min \pi(B)$
- $\pi : x \mapsto 22 - x, \min \pi(A) = 12 = \min \pi(B)$
- $21! = 51090942171709440000 > 2^{65}$

MinHash 3

MinHash estimator

Let $\pi_1, \ldots, \pi_k$ independent and uniformly distributed random permutations and put

$$\hat{J}(A, B) = \frac{1}{k} \#\{j \mid 1 \leq j \leq k, \min H_j(A) = \min H_j(B)\},$$

where $H_j = H \circ \pi_j$.

MinHash 4

Theorem

$\hat{J}(A,B)$ is an unbiased estimator of $J(A,B)$:

$$\mathbb{E}(\hat{J}(A,B)) = J(A,B)$$

and

$$\mathbb{P}(|J(A,B) - \hat{J}(A,B)| < \epsilon) > 1 - \delta,$$

whenever $k > \frac{2}{\epsilon^2}\ln(\frac{2}{\delta})$.

Remark

This estimate is independent of $m = \#\mathcal{X}$.

Example

$k = 512$, $\epsilon = 0.1$, $\delta = 0.2$

MinHash 5

MinHash estimator

For independent and identically distributed random hash functions $H_1, \dots, H_k$ with a common range $\mathcal{Y} \subset \mathbb{Z}$ put

$$\hat{J}(A, B) = \frac{1}{k} \#\{j \mid 1 \leq j \leq k, \min H_j(A) = \min H_j(B)\}.$$

MinHash paradigm

- $\mathbb{E}(\hat{J}(A, B)) \approx J(A, B)$
- $\mathbb{P}(|J(A, B) - \hat{J}(A, B)| < \epsilon) > 1 - \delta$, if $\delta \in (0, 1)$, $\epsilon > 0$ and $k > \frac{2}{\epsilon^2} \ln(\frac{2}{\delta})$.