L3 Similarity

Monday, January 13, 2020 3:02 PM

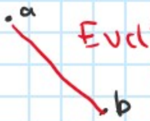
- Goal

Raw Text $\xrightarrow{1}$ Abstract Representation {sets} $\xrightarrow{2}$ Vectors $\xrightarrow{3}$ LSH

· Min Hashings

* Fast Comparison

Distance:



Euclidean Distance

$a = (a_1, a_2, \ldots, a_d) \in \mathbb{R}^d$

$b = (b_1, b_2, \ldots, b_d) \in \mathbb{R}^d$

$d_E(a,b) = \|a-b\| = \sqrt{\sum_{j=1}^{d} (a_j - b_j)^2}$

* The inverse of distance is known as similarity

Distance	Similarity
$d(a,b)$	$s(a,b)$
· small if a, b are close	· large if a, b close
· if large → a, b far.	· if small → a, b far.
· 0, if the same.	· 1, if the same.
· $d(a,b) \in [0, \infty)$	· $s(a,b) \in [0,1]$

↳ $d(a,b) = 1 - s(a,b)$

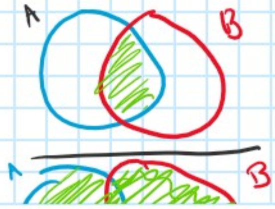
↳ $d(a,b) = \sqrt{s(a,a) + s(b,b) - 2s(a,b)}$

Jaccard Similarity:

$JS(A,B)$ where $A = \{0,1,2,5,6\}$
$B = \{0,2,3,5,7,9\}$

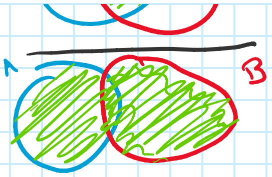
$$= \frac{|A \cap B|}{|A \cup B|} = \frac{|\{0,2,5\}|}{|\{0,1,2,3,5,6,7,9\}|} = \frac{3}{8}$$

$= 0.375$



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

"Jacard Distance"



"Jacard Distance"

Similarities Between sets:

$$S_{x,y,z,z'}(A,B) = \frac{x|A \cap B| + y|\overline{A \cup B}| + z|A \triangle B|}{x|A \cap B| + y|\overline{A \cup B}| + z'|A \triangle B|}$$

for $x, y, z, z' \geq 0$, $z' > z$

Ex.

- $JS(A,B) = S_{1,0,0,1}(A,B) = \frac{|A \cap B|}{|A \cap B| + |A \triangle B|}$
- $Ham(A,B) = S_{1,1,0,1}(A,B) = 1 - \frac{|A \triangle B|}{|[n]|}$
- $Andb(A,B) = S_{1,0,0,2}(A,B) = \frac{|A \cap B|}{|A \cup B| + |A \triangle B|}$
- $RT(A,B) = S_{1,1,0,2}(A,B) = \frac{|[n]| - |A \triangle B|}{|[n]| + |A \triangle B|}$
- $Pice(A,B) = S_{2,0,0,1}(A,B) = \frac{2|A \cap B|}{|A| + |B|}$

Modeling Text:

I am Sam.
Sam I am.
I do not like green eggs and ham.
I do not like them, Sam I am.

Text → vectors $\mathbb{R}^d$
$(d=11)$

Bag-of-words:
(am, and, do, eggs, green, ham, I, like, not, Sam, them, zebra)

Bag-of-words is a count of each word at ith coordinate

$v_1 = (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0)$
$v_2 = (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0)$
$v_3 = (0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$
$v_4 = (1, 0, 1, 0, 0, 0, 2, 1, 1, 1, 1, 0)$

K-Grams with Words:

Words $K=1$:
{ [I], [am], [Sam], . . . [them] }

words $K=2$ (I am Sam. Sam I am...)
{ [I am], [am Sam], [Sam Sam] }

words k=2 (I am Sam. Sam I am...)
{[I am], [am Sam], [Sam Sam]

characters k=3
{[iam], [ams], [msa], [sam] .....}

↳ Jaccard needs sets (even though [iam] occurs twice)

Modeling Choices:

- words vs characters
  ↳ more interpretable
- new lines
- value of K
- capitalization
- punctuation
  ↳ highlight '#'

More complex representation
(larger K, words vs characters, punctuation, etc)
↕
more data.

- K-Grams and Jaccard Example:

D_1: [I am], [am Sam]

D_2: [sam I], [I am]

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} = 1/3 = 0.333$$

- Continuous bag of Words:

- each word $\xrightarrow[\text{maps}]{}$ vector $v_{word} \in \mathbb{R}^d$

↳ bow (0, 0, 0, 1, 0, 0, ..., 0)