Distances

paul

Going the distance

$$d(A, B)$$

is the quantification of

- the quantity of 1D-space between A and B , a length, in meters
- similarity between A and B, a metric $\in \mathbb{R}$
- quantity of separation , as in the social distance between two people in termes of classes

A, B may be sets ...

My own distances taxonomy

- Physical Length : Geometry, Physics, Mecanics
- Error function : Statistical model fitting
- Fitness function : Genetic Algorithm
- Loss function : Machine Learning
- Edit distance : Natural Language Processing
- Paths Lengths: Graph theory, Optics, Acoustics
- Likelyhood : Probabilities

Disclaimer

Due to curse of dimensionality, there is no good distance in a high dimensional data

Distance as Physical Length

Euclidean

$$d(A, B) = \sqrt{\sum_i (A_i - B_i)^2} \in \mathbb{R}$$

Requires cartesian coordinates

Pros

- well known /widely used
- simple/intuitive
- perfect for 2D and 3D

Cons

- subject to scale/units
- subject to curse of dimensionality
- Earth is not flat
- high-dimensional data may include correlations beetween dimensions

toy example : k-means on 2D points

Manhattan Distance

$$d(A, B) = \sum_i |A_i - B_i| \in \mathbb{R}$$

Pros

- ok with high-dimensional data
- perfectily understandable if 1D ;-)

Cons

- "not the shortest"
- hard to interpret

Chebyshev Distance

$$d(A, B) = max_i |A_i - B_i| \in \mathbb{R}$$

"King distance" on a chessboard

Pros

- ?

Cons

-
- hard to interpret

Minkowski Distance

$$d(A,B) = \Big(\sum_i |A_i - B_i|^p\Big)^{\frac{1}{p}} \in \mathbb{R}$$

the *"paramterizable norm"*

$p = 1$: Manhatan

$p = 2$: Euclidean

$p = \infty$: Chebyshev

Pros

- tunable with p

Cons

- shipped with others cons depending on values of p
- hard to interpret (what if $p = 0.3$?)

Chebyshev Distance

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Pros

- correlation taken into account
- variance taken into account

Cons

- distance between an element and a set of others
- outliers sensitive (because variance and mean are)

Haversine Distance

$$d(A,B) = 2r \arcsin \sqrt{\sin^2\left(\frac{\varphi_B - \varphi_A}{2}\right) + cos(\varphi_A)cos(\varphi_B)\sin^2\left(\frac{\lambda_B - \lambda_A}{2}\right)}$$

φ is latitude , λ is longitude, r is the sphere radius.

Pros

- adapted for earth surface points

Cons

- distortions if not on a regular sphere
- scary looking

Distance as Similarity

Cosine similarity

$$d(A, B) = cos(\theta) = \frac{A.B}{||A||.||B||} \in [-1; 1]$$

Requires scalar product and a norm.

Pros

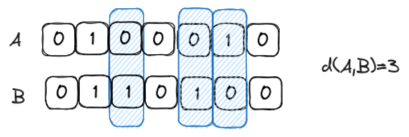
- still simple
- normalised values
- used for high-dimensional data

Cons

- captures "orientation" only
- magnitudes meaningless
- degraded by sparse data

Hamming distance

$$d(A, B) = Card\{i : A_i \neq B_i\}$$



The number of values that differ from A to B.

Requires same length objects

Pros

- intuitive (regarding objects size)
- simple

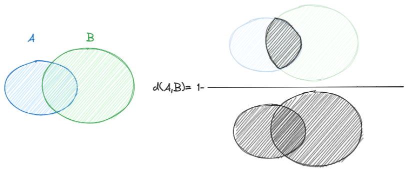
Cons

- same length constraint
- count differences occurences, not the *gap*

use case : similarity using qualitative variables only

Jaccard index

$$d(A,B) = 1 - \frac{A \cap B}{A \cup B} \in [0;1]$$



also called IOU

Distance is 1- Jaccard index

Pros

- intuitive : similarity of sets
- simple with cardinality

Cons

- tend to be low for huge sets ($\cup$ is always big)

use case : similarity between documents as common words count

Kullback-Liebler divergence

$$d(A, B) = \sum_{x \in X} A(x) log \frac{A(x)}{B(x)}$$

A and B are probability distributions on X

Pros

- well known
- feat. entropy

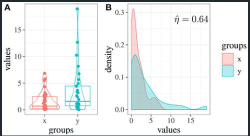
Cons

- how to handle zeros in probabilities ?
- $\Longrightarrow$ additional smoothing required
- not a distance! (no symmetry + no triangle inequality)
- strange if multimodalities

Overlapping index

$$d(A,B) = \int_{\mathbb{R}^n} min[f_A(x), f_B(x)]\ dx \in [0;1]$$

f_A and f_B are probability distribution functions



dug by Kirana, thx!

Pros

- intuitive
- no distributions assumptions (unimodality, symmetry)
- works with different sizes samples

Cons

- ?

Pastore, M., and Calcagnì, A. (2019).Measuring distribution similarities between samples: a distribution-free overlapping index. Front. Psychol. 10:1089. doi: 10.3389/fpsyg.2019.01089

Références

- Initial blog post on Towards Data Science