MAC 0459 / 5865

Topics in Data Science and Engineering

R. Hirata Jr. (hirata@ime.usp.br)

Class 11 (2020)

Distance

- a positive number that represents the separations between two objects
  - S a set of objects
  - $D : S \times S \rightarrow \mathbb{R}^+$
- D must satisfy four properties:
  - non-negative
  - identity
  - symmetry
  - triangle inequality

Distance properties

$D : S \times S \rightarrow \mathbb{R}^+$ **is a distance function (or metric) if it satisfies the properties bellow** $a, b, c \in S$**:**

- **positive** - $D(a, b) \geq 0$ **for any** a **and** b
- **identity** - $D(a, b) = 0$ **if and only if** $a = b$
- **simmetry** - $D(a, b) = D(b, a)$ **for any** a **and** b
- **triangle inequality** - **if** $D(a, c) \leq D(a, b) + D(b, c)$ **for any** a**,** b **and** c

Distance vs Topology

- Distance is usualy defined from R^n to R^+
- What if the objects “live” in a discrete space?

Distance vs Topology

Distance categories

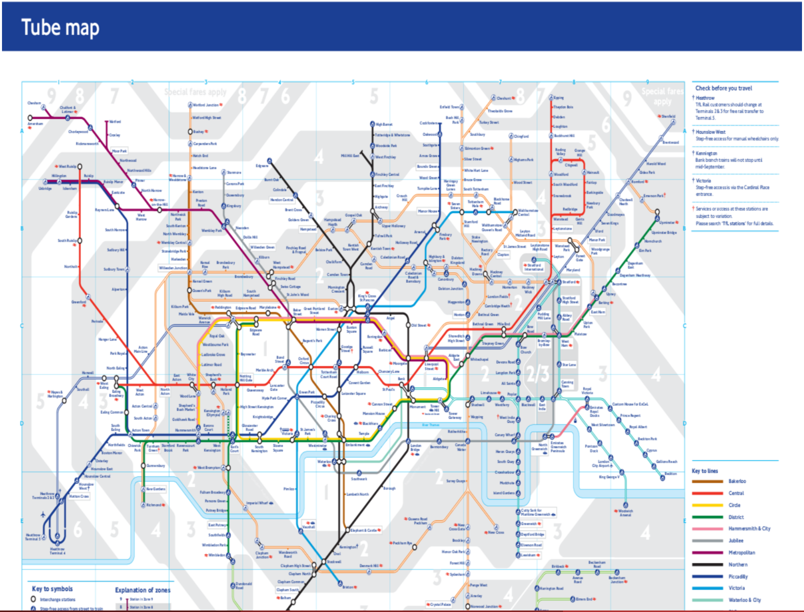
- **Intrinsic - obeys the topology and curvature of the space of objects**
- **Extrinsic - obeys the topology and curvature of the representation space**

Distance vs Topology

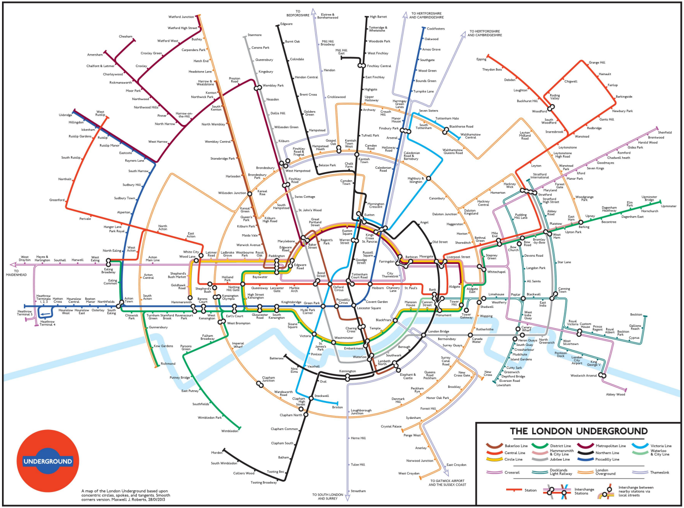
Distance categories

- **Intrinsic - obeys the topology and curvature of the space of objects d: connected digital path between two objects $\rightarrow Z^+$**
- **Extrinsic - obeys the topology and curvature of the representation space d: point coordinates that represent the objects $\rightarrow R^+$**

Distance vs Topology



Distance vs Topology



Distance vs Topology

Tube map vs real map

https://www.reddit.com/r/dataisbeautiful/comments/b8ihhr/comparison_between_the_london_tube_map_and_its/

Distance properties

$D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ **is a distance function (or metric) if it satisfies the properties bellow** $\mathrm{x}_i, \mathrm{x}_j, \mathrm{x}_k \in \mathbb{R}^d$**:**

- **positive, i.e.,** $D(\mathrm{x}_i, \mathrm{x}_j) \geq 0$ **for any** x_i **and** x_j**.**
- **identity, i.e.,** $D(\mathrm{x}_i, \mathrm{x}_j) = 0$ **if and only if** $\mathrm{x}_i = \mathrm{x}_j$**,**
- **simmetry, i.e.,** $D(\mathrm{x}_i, \mathrm{x}_j) = D(\mathrm{x}_j, \mathrm{x}_i)$ **for any** x_i **e** x_j**,**
- **triangle inequality, i.e., if** $D(\mathrm{x}_i, \mathrm{x}_j) \leq D(\mathrm{x}_i, \mathrm{x}_k) + D(\mathrm{x}_k, \mathrm{x}_j)$ **for any** $\mathrm{x}_i, \mathrm{x}_j$ **and** x_k**,**

Some distance functions

Minkowski or norm L_p :

$$D(\mathrm{x}_i, \mathrm{x}_j) = \Big[\sum_{i=k}^{d} (\mathrm{x}_{ik} - \mathrm{x}_{jk})^p \Big]^{1/p}$$

Some distance functions

Minkowski or norm L_p :

$$D(\mathrm{x}_i, \mathrm{x}_j) = \Big[\sum_{i=k}^{d} (\mathrm{x}_{ik} - \mathrm{x}_{jk})^p \Big]^{1/p}$$

$p = 1$ **city-block, Manhattan**

Some distance functions

Minkowski or norm L_p :

$$D(\mathrm{x}_i, \mathrm{x}_j) = \Big[\sum_{i=k}^{d} (\mathrm{x}_{ik} - \mathrm{x}_{jk})^p\Big]^{1/p}$$

$p = 1$ **city-block, Manhattan**

$p = 2$ **Euclidean**

Some distance functions

Minkowski or norm L_p :

$$D(\mathrm{x}_i, \mathrm{x}_j) = \Big[\sum_{i=k}^{d} (\mathrm{x}_{ik} - \mathrm{x}_{jk})^p\Big]^{1/p}$$

$p = 1$ **city-block, Manhattan**

$p = 2$ **Euclidean**

$p \to \infty$ **sup-distance** *(Chebychev distance)*

Some distance functions

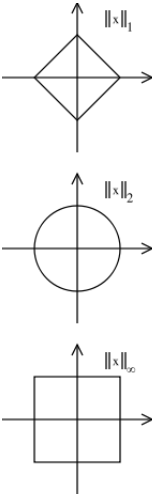
Minkowski or norm L_p :

$$D(x_i, x_j) = \left[\sum_{i=k}^{d} (x_{ik} - x_{jk})^p \right]^{1/p}$$

$p = 1$ **city-block, Manhattan**

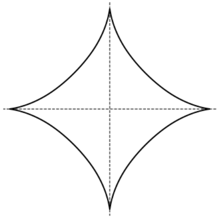
$p = 2$ **Euclidean**

$p \rightarrow \infty$ **sup-distance** *(Chebychev distance)*



Norm L_p

$p = 2/3$ *(subellipse)*

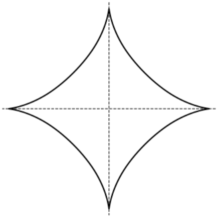


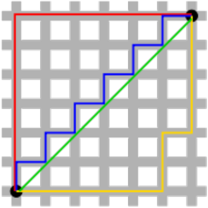
Some distance functions

Norm L_p

$p = 2/3$ *(subellipse)*

$p = 1$ **city-block, Manhattan**





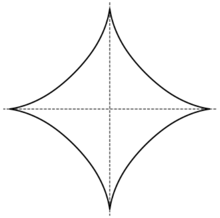
Some distance functions

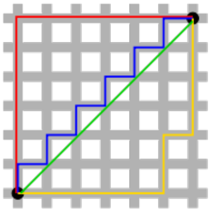
Norm L_p

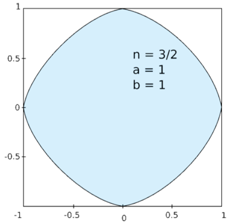
$p = 2/3$ *(subellipse)*

$p = 1$ **city-block, Manhattan**

$p = 3/2$ *(superellipse)*







Some distance functions

Camberra (weighted version of Manhattan):

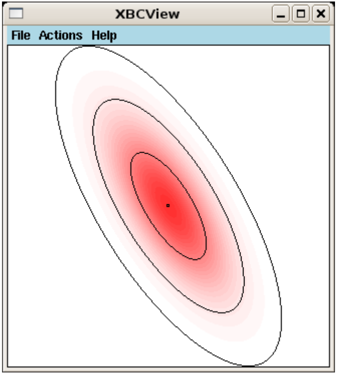
$$D(\mathrm{x}_i, \mathrm{x}_j) = \sum_{i=k}^{d} \frac{|\mathrm{x}_{ik} - \mathrm{x}_{jk}|}{|\mathrm{x}_{ik}| + |\mathrm{x}_{jk}|}$$

Some distance functions

Mahalanobis

$$D(x_i, x_j) = (x_i - x_j)^t \, \Sigma^{-1} \, (x_i - x_j)$$

Σ symmetric positive definite.



Similarity functions

A **similarity function** S **satisfies:**

- $0 \leq S(\mathrm{x}_i, \mathrm{x}_j) \leq 1$ **for any** x_i **and** x_j**,**
- $S(\mathrm{x}_i, \mathrm{x}_j) = S(\mathrm{x}_j, \mathrm{x}_i)$ **for any** x_i **and** x_j**,**
- $S(\mathrm{x}_i, \mathrm{x}_j) = 1$ **if and only if** $\mathrm{x}_i = \mathrm{x}_j$**,**
- $S(\mathrm{x}_i, \mathrm{x}_j)S(\mathrm{x}_j, \mathrm{x}_k) \leq \left[S(\mathrm{x}_i, \mathrm{x}_j) + S(\mathrm{x}_j, \mathrm{x}_k)\right] S(\mathrm{x}_i, \mathrm{x}_k)$

Some similarity functions

Cosine similarity

$$S(\mathrm{x}_i, \mathrm{x}_j) = cos\alpha = \frac{\mathrm{x}_i\,\mathrm{x}_j}{||\mathrm{x}_i||\,||\mathrm{x}_j||}$$

Similarities from distances:

$$S(\mathrm{x}_i, \mathrm{x}_j) = \frac{1}{1 + D(\mathrm{x}_i, \mathrm{x}_j)}$$

if $D(\mathrm{x}_i, \mathrm{x}_j) = 0$ then $S(\mathrm{x}_i, \mathrm{x}_j) = 1$.

Distances between objects with categorical characteristics

Binary data (absent/present characteristic)

$\mathrm{x} = (x_1, x_2, \ldots, x_d)$: object with d characteristcs

$x_k = 1$: *characteristic k present in the object*

$x_k = 0$: *characteristic k absent in the object*

x_i and x_j: any two objects

n_{11}: *characteristics with value* 1 *in both objects*

n_{00}: *characteristics with value* 0 *in both objects*

n_{10}: *characteristics with value* 1 *in* x_i *and value 0 in* x_j

n_{01}: *characteristics with value* 0 *in* x_i *and value 1 in* x_j

$d = n_{11} + n_{00} + n_{10} + n_{01}$

Distances between objects with binary characteristics

Measures that consider some ratio between the matches:

$$\frac{n_{11} + n_{00}}{n_{11} + n_{00} + \omega(n_{10} + n_{01})}$$

$\omega = 1$	**simple**
$\omega = 2$	**Roger & Tanimoto**
$\omega = 1/2$	**Gower & Legendre**

Distances between objects with binary characteristics

Measures that consider some ratio only between the $1-1$ matches ($0-0$ are not considered):

$$\frac{n_{11}}{n_{11} + \omega(n_{10} + n_{01})}$$

$\omega = 1$	**Jaccard index**
$\omega = 2$	**Sokal & Sneath**
$\omega = 1/2$	**Gower & Legendre**

Distances between objects with categorical characteristics

A possible similarity measure when the characteristics are not binary is:

$$S(\mathrm{x}_i, \mathrm{x}_j) = \frac{1}{d} \sum_{l=1}^{d} S_{ijl}$$

which

$$S_{ijl} = \begin{cases} 0 & \textbf{if } \mathrm{x}_i \textbf{ and } \mathrm{x}_j \textbf{ do not match in } l, \\ 1 & \textbf{if } \mathrm{x}_i \textbf{ and } \mathrm{x}_j \textbf{ match in } l. \end{cases}$$

Value “Normalization”

- Distint characteristics can have different scales of representation
- This can introduce some bias to the distance/similarity measure
- At the end, to all the analysis

"Normalization"

Value "Normalization"

$$\hat{x} = \frac{x - \overline{x}}{\sigma}$$

$\overline{x}$ sample mean
σ sample standard deviation

Linear "normalization" to $[0, 1]$ or $[-1, 1]$

Non-linear "normalization"

$$\hat{x} = \frac{x - \overline{x}}{r\sigma} \qquad \hat{x} = \frac{1}{1 + e^{-x}}$$