A.Z
Data Mining
Dec, 2023

lec 1 Distance

Abstract

overview: a brief introduction for methods of distance(similarity), i.e., how measure the degree of difference between two objects

1 Metric

Metric: diatance d that satisfies 4 properties

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ (non-negativity or separation)

2. $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ (coincidence axiom)

3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symmetry)

4. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (triangle inequality)

Metric space $(\mathcal{S}, d)$: data space equipped with a metric

Distance with metric can perform more efficiently.

2 Lp norm

$$L_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{k} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

L_2 called Euclidean distance

L_1 called Manhattan distance

L_{inf} called Chebyshev distance

If $p >= 1$, L-p norm is metric

2.1 Curse of dimensionality

$L_p - norm$ does not work well in high dimensions.

Contrasts between largest and smallest distances $\frac{D_{\max}-D_{\min}}{D_{avg}}$ disappear in high dimensions.

Imporvement:

- Generalized Minkowski distance give weights a_i reflecting importance

$L_p(\mathbf{x}, \mathbf{y}) = (\sum_i a_i \, |x_i - y_i|^p)^{\frac{1}{p}}$

- Fractional L_p quasinorms set $p \in (0, 1)$ (not metrics)

- Match-based similarity with proximity thresholding

3 Match-based similarity with proximity thresholding

We assume that:

1. Features may be only locally relevant (e.g., blood glucose for diabetic patients but not for epileptic)

2. In large dimensions, two objects are unlikely to have similar values, unless the feature is relevant

So the match-based similarity emphasizes dimensions where objects are close/similar, ignores dimensions where x and y not in proximity

Method:

1. Discretize all dimentions to m equi-depth bin_{ij} , i=dimension j=bin number

2. Two points, x and y, are considered to be in proximity on dimension i if both x_i and y_i (the i-th components of x and y) fall into the same bin on that dimension.

3. The proximity set $S(\mathbf{x}, \mathbf{y}, m)$ is defined as the list of dimensions where x_i and y_i are in the same bin.

Similarity measure:

$$\text{PSelect}(\mathbf{x}, \mathbf{y}, m) = \left[\sum_{i \in S(\mathbf{x}, \mathbf{y}, m); x_i \in \text{bin}_{i,j}} \left(1 - \frac{|x_i - y_i|}{width_{i,j}} \right)^p \right]^{1/p}$$

width $_{i,j}$ refers to the width of the bin in the i-th dimension and the j-th bin.

If $\mathbf{x} = \mathbf{y}$:

Then for all dimensions i, $x_i = y_i$

Every dimension would be in $S(\mathbf{x}, \mathbf{y}, m)$ because for every dimension x_i and y_i would fall in the same bin

So, $\mathrm{PSelect}(\mathbf{x}, \mathbf{y}, m) = \left[\sum_{i \in S(\mathbf{x}, \mathbf{y}, m)} 1^p\right]^{1/p}$

Given $S(\mathbf{x}, \mathbf{y}, m)$ contains all dimensions, so $\mathrm{PSelect}(\mathbf{x}, \mathbf{y}, m) = d^{1/p}$, d is the number of dimension

If $S(\mathbf{x}, \mathbf{y}, m) = \emptyset$:

It means that there are no dimensions in which the components of x and y fall into the same bin. In other words, for every dimension, the components of x and y are in different bins, indicates that x and y are not considered similar in any of the dimensions based on the binning criteria used. The similarity measure is 0

paper: The IGrid index: reversing the dimensionality curse for similarity indexing in high dimensional space

4 Cosine similarity

Cosine similarity:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

- suitable for numerical (continuous or integers) and binary data

- in [-1, 1], most similar if $\cos(\mathbf{x}, \mathbf{y}) = 1$

- polular for text documents (their numerical persentation)

Relationship to Euclidean distance L_2 :if vectors are normalized (length 1)

$L_2^2(\mathbf{x}, \mathbf{y}) = 2(1 - \cos(\mathbf{x}, \mathbf{y}))$

5 Mahalanobis distance

Idea: Should distance reflect **data distribution**? High variance direction is more likely to be distant.

Mahalanobis distance

$$\mathrm{Maha}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})\Sigma^{-1}(\mathbf{x} - \mathbf{y})^T} \quad \Sigma = \textit{convariance matrix}$$

6 IOSMAP method

Idea: Measure distances along shortest paths in a nearest neighbour graph

1. Create a nearest neighbour graph $G = (V, E)$ where each $v \in V$ in connected to K nearest neighbours and edge weights represent distances.

2. For any points $v_1, v_2 \in V$

$$\mathrm{Dist}\left(v_1, v_2\right) = \left|\text{shortest path}\left(v_1, v_2\right)\right|$$

This means "intrinsic" or geodesic distances. This distance is the shortest path between two vertices in the graph.

3. Optional step: embed the data into multidimensional space with multidimensional scaling results in lower dimensional representation. Then use either $\mathrm{Dist}\left(v_1, v_2\right)$ or L_p distances in the new space

7 Similarity in categorical data

Generic function:

$$\operatorname{sim}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{k} w_i s\left(x_i, y_i\right)$$

Typically weight $w_i = \frac{1}{k}$ (k = number of features) and many choices for s e.g.

7.1 Overlap similarity

$$s\left(x_i, y_i\right) = \begin{cases} 1 & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{cases}$$

Overlap similarity = fraction of dimensions where $\mathbf{x}$ and $\mathbf{y}$ have equal value

7.2 Goodall measure (its one variant)

$$s\left(x_i, y_i\right) = \begin{cases} 1 - p_i^2\left(x_i\right) & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{cases}$$

where $p_i\left(x_i\right) = \frac{fr\left(A_i = x_i\right)}{n}$, means fraction of records having $A_i = x_i$

paper: Similarity measures forcategorical data: A comparative evaluation.

8 Similarity in mixed data

Give weight to numerical and categorial components

$$\operatorname{sim}(\mathbf{x}, \mathbf{y}) = \lambda \cdot \text{NumSim} + (1-\lambda) \cdot \text{CatSim}$$

but NumSim and CatSim often in different scales, so we need to calculate standard deviations

$$\operatorname{sim}(\mathbf{x}, \mathbf{y}) = \lambda \cdot \text{NumSim} / \sigma_N + (1-\lambda) \cdot \text{CatSim} / \sigma_C$$

9 Similarity in binary data

9.1 Hamming distance

Hamming distance $= L_1 - norm$ for binary data

$$L_1(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|$$

9.2 Jaccard coefficient

$$J(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}$$

What if we consider the meaning of the string?

9.3 Levenshtein distance

Using edit distance (insert, delete, substitute) , special for string

Levenshtein distance=minimum number of unit operations

Levenshtein distance is **metric**

10 Similarity for documents

$\mathbf{x}$ and $\mathbf{y}$ are m-dimensional vectors ($m =$ lexicon size)

$x_i =$ frequency of term i in the document $\mathbf{x}$

Then we can use cosine similarity

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$