

Medidas de Distância

Prof. Jefferson T. Oliva

Aprendizado de Máquina e Reconhecimento de Padrões (AM28CP)
Engenharia de Computação
Departamento Acadêmico de Informática (Dainf)
Universidade Tecnológica Federal do Paraná (UTFPR)
Campus Pato Branco

- Distância Euclidiana
- Distância de Minkowski
- Distância de Mahalanobis
- Similaridade entre Vetores Binários
- Similaridade do Cosseno
- Propriedades Comuns de Distância e de Similaridade

- Medidas de distância são utilizadas para quantificar a similaridade ou dissimilaridade em objetos ou conjuntos de exemplos em um conjunto de dados
- Diversas abordagens de aprendizado de máquina fazem o uso de medidas de distância:
 - Classificação
 - Agrupamento
 - Detecção de outliers
 - Redução de dimensionalidade

- Similaridade
 - Determina o quão semelhantes são dois objetos
 - Quanto maior o valor da medida de distância, maior é a similaridade entre os objetos
 - Geralmente resultam em valores entre 0 e 1
- Dissimilaridade
 - Estabelece o quão diferentes são dois objetos
 - Quanto menor for o valor, maior é a similaridade entre os objetos
 - O valor mínimo é 0, mas o limite superior varia (∞)
- Proximidade se refere à similaridade ou dissimilaridade

- Diversas medidas podem ser utilizadas para o cálculo da similaridade/dissimilaridade
 - Distância Euclidiana
 - Distância de Manhattan
 - Distância de Minkowski
 - Distância de Mahalanobis
 - Coeficiente de casamento simples
 - Similaridade do cosseno
 - ...

Distância Euclidiana

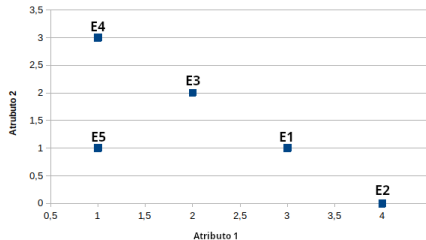
$$d(p, q) = \sqrt{\sum_{i=1}^m (p_i - q_i)^2}$$

onde m é o número de características, p_i e q_i são as i -ésimas características dos exemplos p e q , respectivamente

- Caso as escalas entre os atributos forem diferentes, é necessária a padronização dos seus respectivos valores

Distância Euclidiana

Exemplo	Atributo 1	Atributo 2
E1	1	3
E2	4	0
E3	2	2
E4	3	1
E5	1	1



	E1	E2	E3	E4	E5
E1	0	5,00	2,24	3,61	2,24
E2	5,00	0	2,83	1,41	3,16
E3	2,24	2,83	0	1,41	1,41
E4	3,61	1,41	1,41	0	2,00
E5	2,24	3,16	1,41	2,00	0

Matriz de distâncias

Distância de Minkowski

- Generalização da distância Euclidiana

$$d(p, q) = \left[\sum_{i=1}^m (p_i - q_i)^r \right]^{\frac{1}{r}}$$

- $r = 1$: distância de Manhattan (Norma L_1)
- $r = 2$: distância Euclidiana (Norma L_2)
- $r \rightarrow \infty$: distância de Chebyshev (Norma L_∞)

Distância de Minkowski

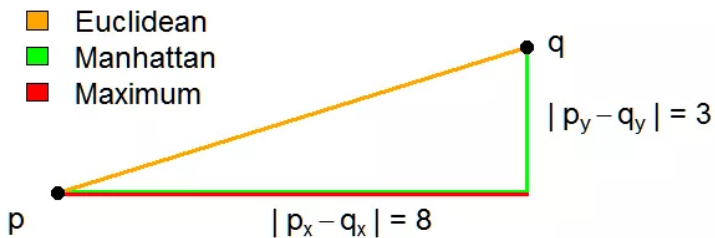
Exemplo	Atributo 1	Atributo 2
E1	1	3
E2	4	0
E3	2	2
E4	3	1
E5	1	1

L1	E1	E2	E3	E4	E5
E1	0	7,00	3,00	5,00	3,00
E2	7,00	0	4,00	2,00	4,00
E3	3,00	4,00	0	2,00	2,00
E4	5,00	2,00	2,00	0	2,00
E5	3,00	4,00	2,00	2,00	0

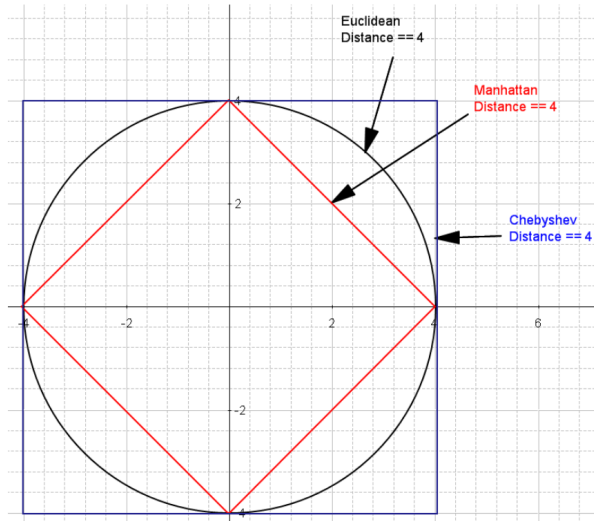
L2	E1	E2	E3	E4	E5
E1	0	5,00	2,24	3,61	2,24
E2	5,00	0	2,83	1,41	3,16
E3	2,24	2,83	0	1,41	1,41
E4	3,61	1,41	1,41	0	2,00
E5	2,24	3,16	1,41	2,00	0

L ∞	E1	E2	E3	E4	E5
E1	0	5	2	3	2
E2	5	0	3	1	3
E3	2	2	0	1	1
E4	3	1	1	0	2
E5	2	3	1	2	0

Distância de Minkowski



Distância de Minkowski



Distância de Mahalanobis

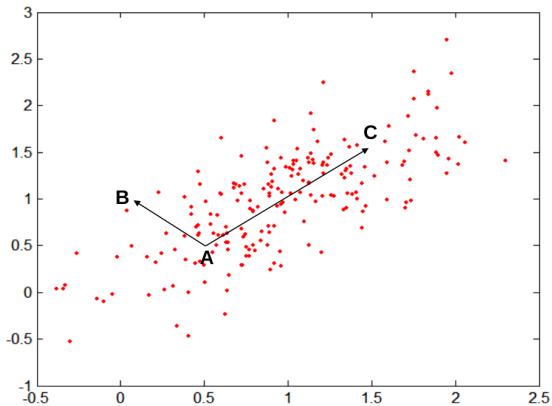
$$d(p, q) = \sqrt{(p - q)^T \sum^{-1} (p - q)}$$

- Onde:
 - \sum é a matriz de co-variância da base de dados X

$$\sum_{i,j} = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

- Caso \sum seja uma matriz identidade, a distância de Mahalanobis é a mesma que a Euclideana
- Essa medida de distância é útil para a detecção de *outliers*

Distância de Mahalanobis



$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Similaridade entre Vetores Binários

Similaridade entre Vetores Binários

- Os exemplos comparados contêm apenas atributos binários
- Para mensurar a similaridade entre exemplos binários, são computados os seguintes parâmetros:
 - M_{00} número de atributos em que $p = 0$ e $q = 0$
 - M_{01} número de atributos em que $p = 0$ e $q = 1$
 - M_{10} número de atributos em que $p = 1$ e $q = 0$
 - M_{11} número de atributos em que $p = 1$ e $q = 1$
- Exemplos de medidas de distância: coeficiente de casamento simples e distância de Jaccard

- Coeficiente de casamento simples

$$d(p, q) = \frac{M_{00} + M_{11}}{M_{00} + M_{10} + M_{01} + M_{11}}$$

- Conta igualmente 0's e 1's
- Adequada para atributos simétricos
- Distância (coeficiente) de Jaccard

ou

$$d(p, q) = \frac{M_{11}}{M_{10} + M_{01} + M_{11}}$$

- Ignora as ocorrências de pares de 0's (e.g. $p_i = p_j = 0$) para lidar adequadamente com atributos assimétricos
 - Afinal, 0 indica apenas a ausência de uma característica

- Exemplo

$$p = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$$

$$q = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1]$$

- $M_{00} = 7$

- $M_{01} = 2$

- $M_{10} = 1$

- $M_{11} = 0$

- Coeficiente de casamento simples = $\frac{7+0}{7+2+1+0} = 0,7$

- Coeficiente de Jaccard = $\frac{0}{2+1+0} = 0$

- Exercício: qual é a medida de distância que resulta no menor valor para os vetores abaixo? Manhattan, Euclideana, coeficiente de casamento simples ou coeficiente de Jaccard?

$$p = 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0$$

$$q = 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1$$

Similaridade do Cosseno

Similaridade do Cosseno

- Comumente utilizada no processamento de linguagem natural
 - Os documentos são representados por vetores, onde cada atributo representa a frequência de ocorrência de uma palavra no texto

$$\cos(p, q) = \frac{p \bullet q}{||p|| \cdot ||q||}$$

onde

- é o produto interno entre os vetores

$||p||$ é a norma do vetor p

- Exemplo:

$$p = [3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0]$$

$$q = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2]$$

- $p \bullet q =$

$$3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

- $\|p\| = \sqrt{3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = \sqrt{42} = 6,48074069840786$

- $\|q\| = \sqrt{1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2} = \sqrt{6} = 2,449489742783178$

- $\cos(p, q) = \frac{5}{6,48074069840786 * 2,449489742783178} = 0,314970394174356$

- Distância = $1 - \cos(p, q) = 0,685029605825644$

Propriedades Comuns de Distância e de Similaridade

- Medidas de distância, como a Euclidiana, possuem algumas propriedades bem conhecidas
 - $d(p, q) \geq 0$ para todo p e q
 - $d(p, q) = 0$ apenas se $p = q$
 - $d(p, q) = d(q, p)$ para todo p e q (simetria)
 - $d(p, r) \leq d(p, q) + d(q, r)$ para todo p, q e r (desigualdade triangular)

Onde $d(p, q)$ é a distância (dissimilaridade) entre os exemplos p e q

- Uma distância que satisfaz essas propriedades é considerada uma medida de distância

Propriedades Comuns de Distância e de Similaridade

- Medidas de similaridade também possuem algumas propriedades bem conhecidas:
 - $s(p, q) = 1$ (similaridade máxima) apenas se $p = q$
 - $s(p, q) = s(q, p)$ para todo p e q (simetria)

Onde $s(p, q)$ é a similaridade entre os exemplos p e q



CASANOVA, D.

Distance Measures. Aprendizado de Máquina.

Slides. Engenharia de Computação. Dainf/UTFPR, 2020.



RASCHKA, S.; MIRJALILI, V.

Python Machine Learning.

Packt, 2017.



TAN P.; STEINBACK M.; KUMAR V.

Introduction to Data Mining.

Pearson, 2006.