Machine Learning
Medidas de distância de variáveis quantitativas

Prof. Hugo de Paula

Medidas de distância: variáveis contínuas

Qualquer distância métrica pode ser utilizada.

Mais importantes são classes de distâncias de Minkowski:

$$d(i,j) = \sqrt[n]{\left(\left|x_{i1} - x_{j1}\right|^n + \left|x_{i2} - x_{j2}\right|^n + \cdots + \left|x_{ip} - x_{jp}\right|^n\right)}$$

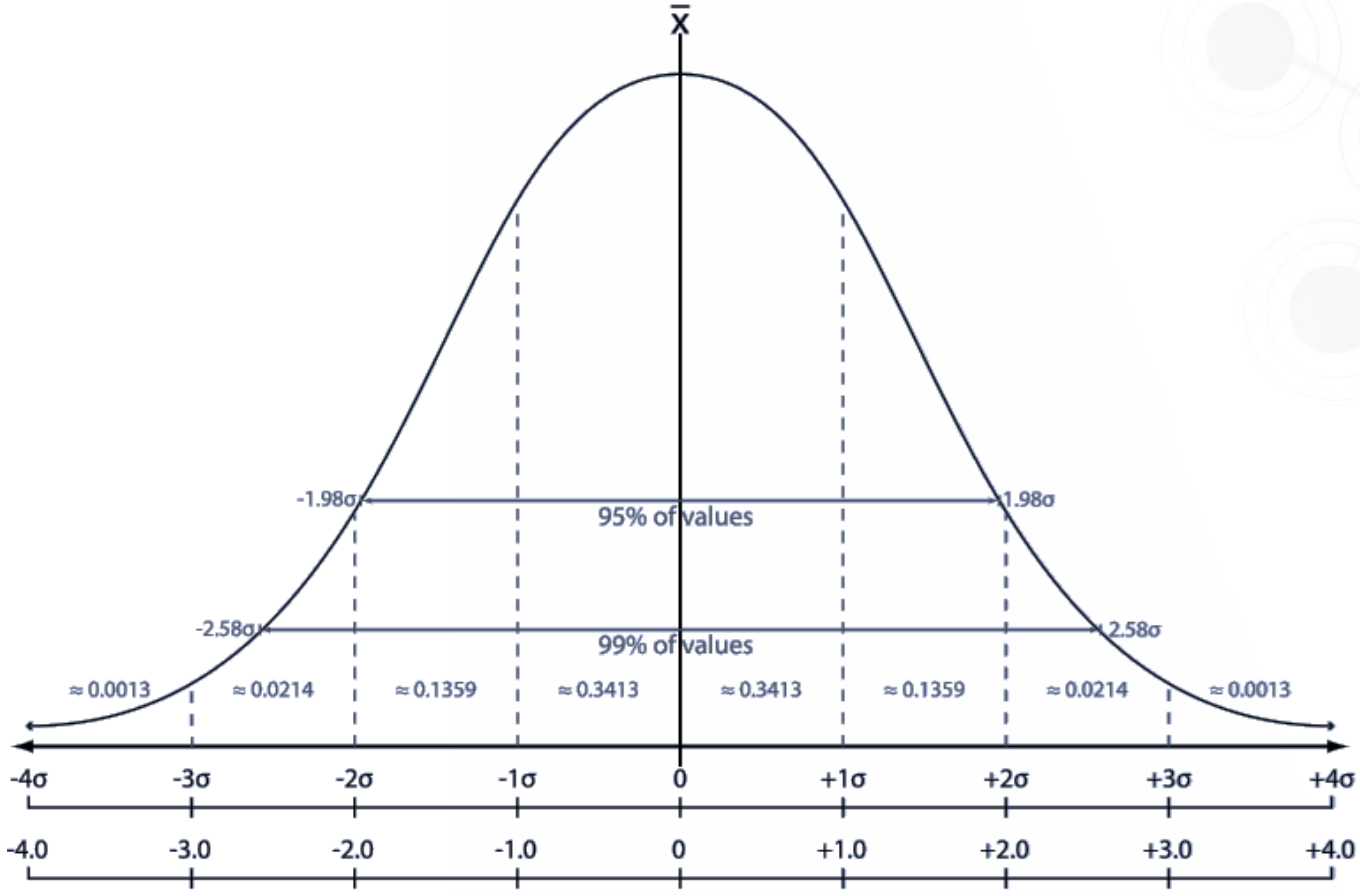
- Se q = 1, d é a distância de Manhattan
- Se q = 2, d é a distância Euclidiana

Normalização e padronização de dados numéricos

Z-score: $z = \frac{x-\mu}{\sigma}$

- x: valor, μ: média, σ: desvio padrão
- Distância entre o dado e a população em termos do desvio padrão
- Negativo quando abaixo da média, e positivo caso acima

Z-score



desvio padrão da média

Z-score

Normalização e padronização de dados numéricos

Normalização Min-Max:

$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}(max_{novo} - min_{novo}) + min_{novo}$$

Normalização Min-Max

ID	Gênero	Idade	Salário
1	F	27	19.000
2	M	51	64.000
3	M	52	100.000
4	F	33	55.000
5	M	45	45.000

ID	Gênero	Idade	Salário
1	1	0.00	0.00
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32

Medidas de similaridade baseadas em vetor

Em alguns casos, medidas de distância provêm visão distorcida

- Ex. Quando o dado é muito esparso e 0´s no vetor não são significativos
- Nesses casos, melhor utilizar medidas de distância baseada em vetor

$X = \langle x_1, x_2, \cdots, x_n \rangle$ $\quad$ $Y = \langle y_1, y_2, \cdots, y_n \rangle$

Similaridade de cosseno

- Produto escalar: $sim(X, Y) = X \cdot Y = \sum x_i \times y_i$
- A norma do vetor X é: $\|X\| = \sqrt{\sum x_i^2}$
- A similaridade de cosseno é:

$$sim(X, Y) = \frac{X \cdot Y}{\|X\| \times \|Y\|} = \frac{\sum x_i \times y_i}{\sqrt{\sum x_i^2} \times \sqrt{\sum y_i^2}}$$

Similaridade de cosseno

Exemplo: X = <2, 0, 3, 2, 1, 4>

$$\|X\| = \sqrt{(4 + 0 + 9 + 4 + 1 + 16)} = 5{,}83$$

$$X^* = \frac{X}{\|X\|} = < 0.343, 0, 0.514, 0.343, 0.171, 0.686 >$$

- Note que $\|X\| = 1$
- Torna o vetor de comprimento unitário

Exemplo: Similaridade entre documentos

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	0	4	0	0	0	2	1	3
Doc2	3	1	4	3	1	2	0	1
Doc3	3	0	0	0	3	0	3	0
Doc4	0	1	0	3	0	0	2	0
Doc5	2	2	2	3	1	4	0	2

Doc2 · Doc4 = <3,1,4,3,1,2,0,1> * <0,1,0,3,0,0,2,0>

0 + 1 + 0 + 9 + 0 + 0 + 0 + 0 = 10

Norma(Doc2) = SQRT(9+1+16+9+1+4+0+1) = 6.4

Norma(Doc4) = SQRT(0+1+0+9+0+0+4+0) = 3.74

Cosseno(Doc2, Doc4) = 10 / (6.4 * 3.74) = 0.42

Correlação

Em casos onde há uma variância média alta entre os dados (ex. avaliação de filmes), o coeficiente de correlação de Pearson é a melhor opção.

Correlação de Pearson

$$coor(x, y) = \frac{cov(x, y)}{stdev(x) \cdot stdev(y)}$$

Aviso legal

O material presente nesta apresentação foi produzido a partir de informações próprias e coletadas de documentos obtidos publicamente a partir da Internet. Este material contém ilustrações adquiridas de bancos de imagens de origem privada ou pública, não possuindo a intenção de violar qualquer direito pertencente à terceiros e sendo voltado para fins acadêmicos ou meramente ilustrativos. Portanto, os textos, fotografias, imagens, logomarcas e sons presentes nesta apresentação se encontram protegidos por direitos autorais ou outros direitos de propriedade intelectual.

Ao usar este material, o usuário deverá respeitar todos os direitos de propriedade intelectual e industrial, os decorrentes da proteção de marcas registradas da mesma, bem como todos os direitos referentes a terceiros que por ventura estejam, ou estiveram, de alguma forma disponíveis nos slides. O simples acesso a este conteúdo não confere ao usuário qualquer direito de uso dos nomes, títulos, palavras, frases, marcas, dentre outras, que nele estejam, ou estiveram, disponíveis.

É vedada sua utilização para finalidades comerciais, publicitárias ou qualquer outra que contrarie a realidade para o qual foi concebido. Sendo que é proibida sua reprodução, distribuição, transmissão, exibição, publicação ou divulgação, total ou parcial, dos textos, figuras, gráficos e demais conteúdos descritos anteriormente, que compõem o presente material, sem prévia e expressa autorização de seu titular, sendo permitida somente a impressão de cópias para uso acadêmico e arquivo pessoal, sem que sejam separadas as partes, permitindo dar o fiel e real entendimento de seu conteúdo e objetivo. Em hipótese alguma o usuário adquirirá quaisquer direitos sobre os mesmos.

O usuário assume toda e qualquer responsabilidade, de caráter civil e/ou criminal, pela utilização indevida das informações, textos, gráficos, marcas, enfim, todo e qualquer direito de propriedade intelectual ou industrial deste material.

PUC Minas

Virtual