Machine Learning
Medidas de distância de variáveis qualitativas

Prof. Hugo de Paula

Medidas de similaridade: variáveis binomiais ou binárias

Atributos de tipo binário ou booleano só têm dois valores : 1 ou 0, sim ou não, alto ou baixo.

Tratar como valores numéricos pode levar a análises errôneas.

Amostra		Objeto j	
Objeto i	Valor	1	0
	1	a	b
	0	c	d

Medidas de similaridade: variáveis binomiais ou binárias

Valores casados: **a + d**

Valores distintos: **b + c**

Numero de atributos: **a + b + c + d**

Medidas de similaridade: variáveis binomiais ou binárias

Medida de distância (atributos simétricos)

$$d(i,j) = \frac{b + c}{a + b + c + d}$$

Distância de Jaccard (atributos assimétricos)

$$d(i,j) = \frac{b + c}{a + b + c}$$

Medidas de similaridade: variáveis binomiais ou binárias

Casamento simples - *matching* (atributos simétricos)

$$matching(i,j) = \frac{a+d}{a+b+c+d}$$

Similaridade de Jaccard (atributos assimétricos)

$$sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$$

Medidas de similaridade: variáveis binomiais ou binárias

Nome	Gênero	Febre	Tosse	Teste 1	Teste 2	Teste 3	Teste 4
João	M	S	N	S	N	N	N
Maria	F	S	N	S	N	S	N
José	M	S	S	N	N	N	N

- Gênero é um atributo simétrico
- Os outros atributos são assimétricos
- Seja S = 1, e N = 0

$$d(joão, maria) = \frac{0+1}{2+0+1} = 0.33$$

$$d(maria, josé) = \frac{1+2}{1+1+2} = 0.75$$

Medidas de distância: variáveis binomiais ou binárias

Distância	Fórmula	Propriedade
Hamming (Manhattan)	b+c	não normalizada
Euclidiana	sqrt(b+c)	não normalizada
Chebyshev discreto	max(b; c)	não normalizada
Soergel	(b+c)/(b+c+d)	normalizada
Hamming média	(b+c)/(a+b+c+d)	normalizada
Euclidiana média	sqrt((b+c)/(a+b+c+d))	normalizada

Medidas de similaridade: variáveis binomiais ou binárias

Similaridade	Fórmula	Propriedade
Russel & Rao	a/(a+b+c+d)	normalizada
Jaccard	a/(a+b+c)	normalizada
Rogers & Tanimoto	(a+d)/(a+2*(b+c)+d)	normalizada
Hamann	((a+d) - (b+c))/(a + b + c + d)	normalizada
Dice	2*a/(2*a+b+c)	normalizada
Match simples	(a+d)/(a+b+c+d)	normalizada
McConnoughy	(a*a - b*c) / sqrt((a+b)*(a+c))	normalizada

Medidas de distância: variáveis nominais ou categóricas

Casamento simples - matching

– m: num de matches, p: num total de variáveis

$$matching(i,j) = \frac{p-m}{p}$$

Medidas de distância: variáveis nominais ou categóricas

Converter para o formato de planilha binomial

- Para cada atributo A, criar P atributos binários para os P estados nominais (categorias) de A
- Exemplo: A1: Temp = alta; A2: Temp = média; A3: Temp = baixa

Medidas de distância: variáveis categóricas ordinais

Rank

Trocar x_{if} pelo seu rank $r_{if} \in \{1, \cdots, M_f\}$

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- mapeia a faixa de cada variável em um intervalo [0, 1].
- computa dissimilaridade usando método para variáveis contínuas comuns.

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