Licap

Formação Cientista de Dados

PUC Minas

Licap

Formação do Cientista de Dados
Análise de Entropia – Módulo Básico

Luis Enrique Zárate

Conteúdo do Curso

Licap

Definições básicas para Ciência de Dados

1. O que é dado
2. O que é idéia acerca do dado
3. O que é informação
4. O que é conhecimento
5. O que é aprendizado
6. O que é descoberta de conhecimento
7. Exemplo ilustrativo

Análise de Entropia - Teoria da Informação

Outro método para ranking de atributos, objetivando a diferenciação entre classes é baseado na *Teoria da Informação,* através do conceito de *entropia*:

Seja X uma variável discreta

$$X = \{x_k \mid k = 0, \pm 1, \ldots, \pm K\}$$

Onde xk é um número discreto e (2K+1) é o número total de níveis discretos.

O evento xk ocorre com uma probabilidade de:

$$(p_k = P(X = x_k)$$

Para o qual se cumpre:

$$0 \leq p_k \leq 1 \qquad \sum_{k=-K}^{K} p_k = 1$$

Suponha que o evento $(X = x_k)$

ocorra com probabilidade $(p_k = 1)$

pelo qual $(p_i = 1)$ i ≠ k, nesta situação não há “surpresa” nem

“informação” transmitida pelo evento.

A quantidade de informação após observar o evento **xk**, com probabilidade **pk** é uma função logarítmica.

$$I(x_k) = Log\left(\frac{1}{p_k}\right) = -Log(p_k)$$

Quando a base 10 é utilizada as unidades serão *nats* e quando a base 2 é utilizada as unidades serão *bits*.

Propriedades:

1. Se estivermos absolutamente certos de um evento, nenhuma informação é ganha.

$$I(x_k) = 0 \quad para \quad p_k = 1$$

2. Qualquer evento xk, fornece alguma ou nenhuma informação, mas nunca é perda de informação

$$I(x_k) \geq 0 \quad para \quad 0 \leq p_k \leq 1$$

Propriedades:

3. Quanto menos provável seja um evento, mas informação é ganha através de sua ocorrência

$$I(x_k) > I(x_i) \quad \quad para \quad p_k < p_i$$

A Entropia é o valor médio de I(xk) sobre o intervalo completo de 2K+1 valores discretos. É dada por:

$$H(X) = E[I(x_k)]$$

$$H(X) = \sum_{k=-K}^{K} p_k I(x_k)$$

$$H(X) = -\sum_{k=-K}^{K} p_k Log(p_k)$$

A Entropia é limitada pelo intervalo

$$0 \leq H(X) \leq Log(2K+1)$$

Eliminação de atributos pela entropia

Por exemplo, considere a seguinte variável discreta:

$$X = \{2,2,2,2,2,2,2,2,2,2\}$$

A probabilidade do evento ser X=2 é dado por: $(p_2 = 1)$

A quantidade de informação contida no evento X=2 é:

$$I(x_k) = Log\left(\frac{1}{p_k}\right) = -Log(1) = 0$$

A Entropia é:

$$H(X) = -\sum_{k=-K}^{K} p_k Log(p_k) = 0$$

Por exemplo, considere a seguinte variável discreta:

$$X = \{2,2,2,2,2,3,3,3,3,3\}$$

A probabilidade do evento ser X=2 é dado por: $(p_2 = 0{,}5)$

A probabilidade do evento ser X=3 é dado por: $(p_3 = 0{,}5)$

A quantidade de informação contida no evento X=2 é:

$$I(x_k = 2 = 3) = Log\left(\frac{1}{p_k}\right) = -Log(0{,}5) = 0{,}30$$

A Entropia é:

$$H(X) = -\sum_{k=-K}^{K} p_k Log(p_k) = 0{,}5 * 0{,}3 + 0{,}5 * 0{,}3$$

$$H(X) = 0{,}3 \quad nats$$

Por exemplo, considere a seguinte variável discreta:

$$X = \{2,2,3,3,3,3,4,4,4,4\}$$

A probabilidade do evento ser X=2 é dado por: $(p_2 = 0{,}2)$

A probabilidade do evento ser X=3 é dado por: $(p_3 = 0{,}4)$

A probabilidade do evento ser X=4 é dado por: $(p_4 = 0{,}4)$

A quantidade de informação contida no evento X=2, X=3 e X=4 é:

$$I(x_k = 2) = -Log(0,2) = 0,699$$

$$I(x_k = 3 = 4) = -Log(0,4) = 0,398$$

A Entropia é:

$$H(X) = \sum_{k=-K}^{K} p_k I(x_k) = 0,2 * 0,699 + 0,4 * 0,398 + 0,4 * 0,398$$

$$H(X) = 0,3623 \quad nats$$

Eliminação de atributos pelo poder classificatório

Outro método para ranking de atributos, objetivando a diferenciação entre classes é baseado na *Teoria da Informação,* através do conceito de *entropia*:

X1	x2	X3	Classe
P	70	V	A
P	90	V	B
P	85	F	B
P	95	F	B
P	70	F	A
Q	90	V	A
Q	78	F	A
Q	65	V	A
Q	75	F	A
R	80	V	B
R	70	V	B
R	80	F	A
R	80	F	A
R	96	F	A

$$H(X) = -\sum_{k=-K}^{K} p_k Log(p_k)$$

X1	x2	X3	Classe
P	70	V	A
P	90	V	B
P	85	F	B
P	95	F	B
P	70	F	A
Q	90	V	A
Q	78	F	A
Q	65	V	A
Q	75	F	A
R	80	V	B
R	70	V	B
R	80	F	A
R	80	F	A
R	96	F	A

$$H(X1) = -\frac{5}{14}\left[\frac{2}{5}\log_2(2/5) + \frac{3}{5}\log_2(3/5)\right]$$
$$-\frac{4}{14}\left[\frac{4}{4}\log_2(4/4) + \frac{0}{4}\log_2(0/4)\right]$$
$$-\frac{5}{14}\left[\frac{3}{5}\log_2(3/5) + \frac{2}{5}\log_2(2/5)\right] = 0{,}694 \quad bits$$

$$H(X2) = -\frac{9}{14}\left[\frac{7}{9}\log_2(7/9) + \frac{2}{9}\log_2(2/9)\right]$$

$$-\frac{5}{14}\left[\frac{2}{5}\log_2(2/5) + \frac{3}{5}\log_2(3/5)\right] = 0{,}837 \quad bits$$

O atributo 2 foi discretizado considerando:

X2 ≤ 80 e X2 > 80

$$H(X3) = -\frac{6}{14}\left[\frac{3}{6}\log_2(3/6) + \frac{3}{6}\log_2(3/6)\right]$$

$$-\frac{8}{14}\left[\frac{6}{8}\log_2(6/8) + \frac{2}{8}\log_2(2/8)\right] = 0{,}892 \quad bits$$

Eliminação de conjuntos de atributos pela similaridade

É uma técnica que consiste em analisar o grau da Entropia das medidas de similaridade entre conjuntos com diferentes números de atributos.

A aproximação consiste na remoção de irrelevantes e redundantes características da base de dados.

Quando os atributos são numéricos, a medida de similaridade entre duas instâncias pode ser medida por:

$$S_{ij} = e^{-\alpha D_{ij}}$$

Onde Dij é a distância entre as instâncias xi e xj e α é um parâmetro expresso como:

$$\alpha = -(\ln 0{,}5) / D$$

D é a distância média entre as instâncias da base de dados e é calculada como:

$$D = \frac{\sum D_{ij}}{M} \qquad M = C_r^n = \frac{n!}{r!(n-r)!}$$

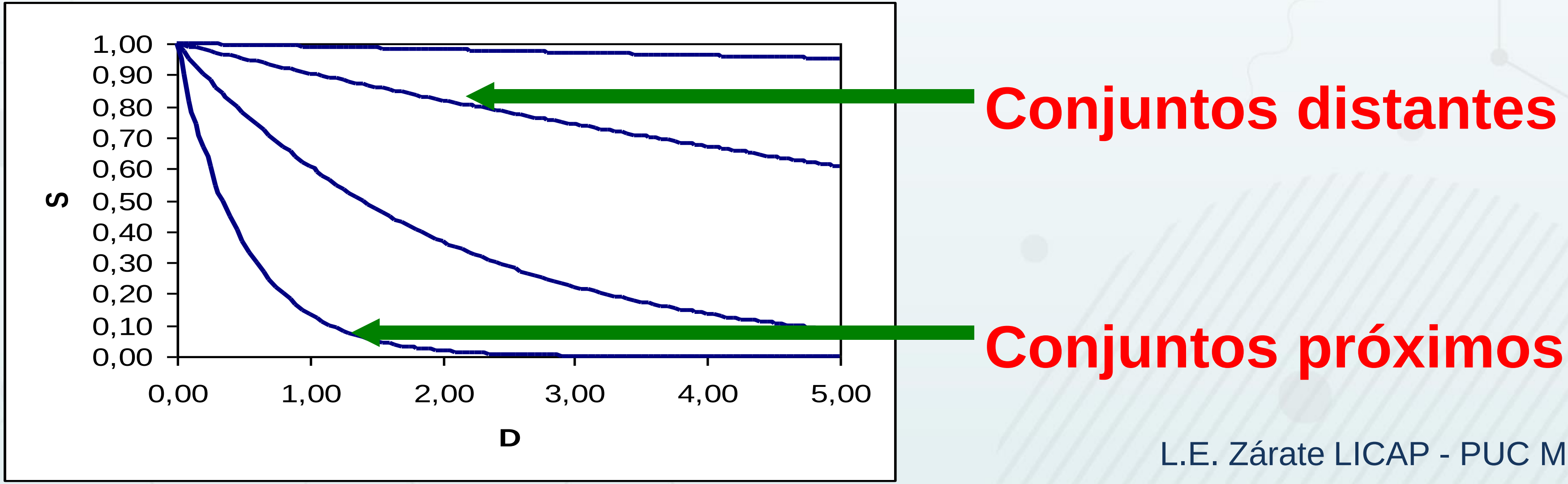
Na prática α tem sido ajustado para o valor 0,5

A medida da Distância Euclideana Normalizada é utilizada para calcular as distâncias Dij.

$$D_{ij} = \left[\sum_{k=1}^{n} ((x_{ik} - x_{jk}) / (\max_k - \min_k))^2 \right]^{1/2}$$

A menor distância média é aproximadamente: D ≈ 0 e
a maior distância média é aproximadamente: D ≈ n^0,5

O menor valor de α é aproximadamente: $\alpha \approx$ 0,69/(n^0,5)
O maior valor de α é aproximadamente: $\alpha \approx$ 0,69/0 ≈ ∝



Exemplo:

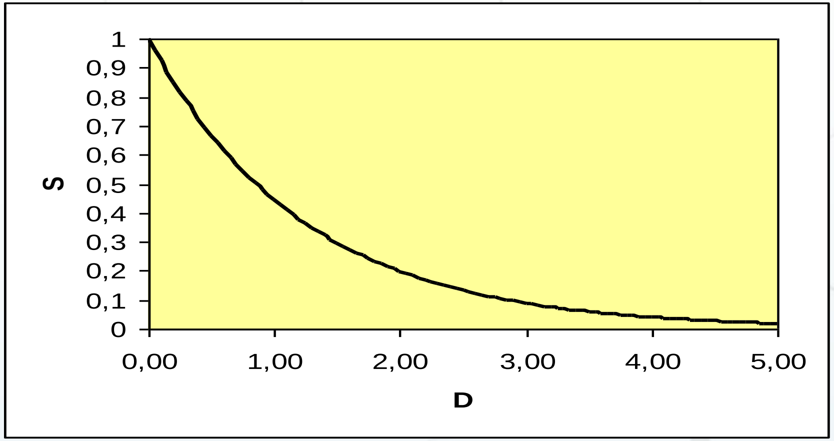
Registro	Idade	Peso(Kg)	Altura (cm)	Norm_Idade	Norm_Peso	Norm_Altura
1	22,00	86,63	183,52	0,00	0,44	0,64
2	23,00	77,13	172,09	0,33	0,16	0,00
3	24,00	92,13	180,98	0,67	0,61	0,50
4	24,00	104,25	184,79	0,67	0,97	0,71
5	24,00	105,13	189,87	0,67	1,00	1,00
6	25,00	71,88	184,15	1,00	0,00	0,68
Mínimo=	22,00	71,88	172,09			
Máximo=	25,00	105,13	189,87			

Dij	1	2	3	4	5	6
1	0,00	0,78	0,70	0,85	0,94	1,09
2		0,00	0,75	1,13	1,35	0,96
3			0,00	0,42	0,63	0,72
4				0,00	0,29	1,03
5					0,00	1,10
6						0,00

$$D = \frac{\sum D_{ij}}{M} = 0,85$$

$$\alpha = -(\ln 0,5)\,/\,D = 0,81$$

$$S_{ij} = e^{-0,81 D_{ij}}$$



Sij	1	2	3	4	5	6
1	1,00	0,53	0,56	0,50	0,47	0,41
2		1,00	0,54	0,40	0,33	0,46
3			1,00	0,71	0,60	0,56
4				1,00	0,79	0,43
5					1,00	0,41
6						1,00

Eliminação de conjuntos de atributos pela similaridade

Quando todos os **atributos são não numéricos**, a medida de similaridade entre duas instâncias pode ser medida pela distância de Hamming:

$$S_{ij} = \left(\sum_{k=1}^{n} \left| x_{ik} = x_{jk} \right| \right) / \, n$$

Onde |Xik=Xjk| = 1 se Xik = Xjk e 0 nos outros casos

Exemplo:

Registro	Atrib 1	Atrib 2	Atrib 3
1	A	X	1
2	B	Y	2
3	C	Y	2
4	B	X	1
5	C	Z	3

Registro	1	2	3	4	5
1	1	0/3	0/3	2/3	0/3
2		1	2/3	1/3	0/3
3			1	0/3	1/3
4				1	0/3
5					1

Sij	1	2	3	4	5
1	1,00	0,00	0,00	0,67	0,00
2		1,00	0,67	0,33	0,00
3			1,00	0,00	0,33
4				1,00	0,00
5					1,00

Para dados misturados: atributos numéricos e não numéricos, é recomendável discretizar os valores numéricos e transformar características numéricas em características nominais.

A técnica consiste em comparar as entropias, das medidas de similaridade, antes e depois de remover um atributo.

Se as duas medidas são próximas, então o conjunto reduzido é satisfatório e é aproximado ao original.

A expressão para cálculo da entropia é dado por:

$$E = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} [S_{ij} . Log S_{ij} + (1 - S_{ij}) . Log(1 - S_{ij})]$$

Algoritmo:

i) Iniciar com o conjunto completo de características F e calcule sua entropia EFf

ii) Para cada característica f $\in$ F, remover um atributo f de F e obtenha um subconjunto Ff. Encontre a diferença entre a entropia de F e as entropias de todos os Ff. Exemplo: (EF-EF-F1), (EF-EF-F2), (EF-EF-F3).

iii) Considere fk como a característica tal que a diferença entre (EF-EF-Ffk) é mínima

iV) Atualizar o conjunto de características F=F- {fk}. Exemplo se (EF-EF-F1) é mínimo o novo conjunto é {F2,F3}

V) Repetir os passos 2-4 até existir somente uma característica em F.