Universidade do Minho (DMA)
janeiro 2019

Machine Learning:: Fundamental and Application
Teste Modelo

Nome ______________________________

Número ______________________ Curso ______________________

Existe dois tipos de dados com atributos binarios: atributos exclusivos ou atributos independentes.

- Atributos `red`, `blue`, `green`, com valores 0 para Não e 1 para Sim, são exclusivos, porque não podemos ter ao mesmo tempo a cor vermelha e azul. Apenas um dos três atributos tem o valor 1.
- Atributos binarios independentes tais como `sol`, `vento`, `humidade` não apresentam qualquer tipo de exclusão entre eles.

O objetivo do estudo é de identificar as metricas binarias adequadas para cada tipo de situações. Recordamos que $\mathcal{A} = A_1 \times \cdots \times A_I$ é o espaço dos atributos, com $A_i = \{0,1\}, i = 1, \cdots, I$ para os atributos binarios e se $x = (x_1, \cdots, x_I) \in \mathcal{A}$, temos $x_i \in \{0,1\}$. Sejam $x, x' \in \mathcal{A}$, definimos

$$a(x,x') = \sum_{i=1}^{I} x_i x'_i, \quad b(x,x') = \sum_{i=1}^{I} x_i(1-x'_i), \quad c(x,x') = \sum_{i=1}^{I} (1-x_i)x'_i, \quad d(x,x') = \sum_{i=1}^{I} (1-x_i)(1-x'_i).$$

1) Mostrar que para quaisquer eventos x e x' temos $0 \leq a,b,c,d \leq I$ e $a+b+c+d = I$.
2) Mostrar que $a(x,x) + d(x,x) = I$ e $b(x,x) = c(x,x) = 0$.
3) Mostrar que, no caso dos atributos binarios exclusivos, temos $d(x,x') < I$.
4) Notamos por índice de Jaccard $J(x,x')$ e índice de Sokal $Sok(x,x')$ as quantidades seguintes

$$Jac(x,x') = \frac{a}{a+b+c}, \qquad Sok(x,x') = \frac{a+d}{a+b+c+d}.$$

Explicar porque, no caso de atributos binarios independentes, Jac não é bem definido. Justificar porque o índice Jac torna se bem definido no caso de atributos exclusivos.
5) Mostrar que se os atributos são exclusivos, $Sok(x,x')$ pode tomar apenas dois valores: $\frac{I-2}{I}$ ou 1, e $Jac(x,x')$ pode tomar apenas os valores 0 ou 1.
6) Justificar porque Jac é adequado para os atributos exclusivos e Sok para os atributos independentes.
7) Mostrar que Jac e Sok são simetricas. Mostrar que Jac e Sok são *definiteness*, a saber que $Jac(x,x') = 1 \Rightarrow x = x'$ e a mesma coisa com Sok.
8) Consideramos a tabela seguinte

	Feature	Dove	Hen	Duck	Goose	Owl	Hawk	Eagle	Fox	Dog	Wolf	Cat	Tiger	Lion	Horse	Zebra	Cow
Is	small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
	medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
likes to	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
	run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Calcular Jac(Dove,Tiger) e Sok(Dove,Tiger). Qual valor parece mais apropriado? Mesma pergunta com Jac(Lion,Tiger) e Sok(Lion,Tiger)?

Correção

1) Como $x_i, x'_i \in \{0,1\}$, logo $(1-x_i), (1-x'_i) \in \{0,1\}$ e qualquer produto $x_i x'_i$, $x_i(1-x'_i)$, $(1-x_i)x'_i$, $(1-x_i)(1-x'_i)$ retorna um valor 0 ou 1. Consequência as somas ficam sempre entre 0 e I.
Por outro lado, temos com o desenvolvimento dos produtos e simplificações

$$a+b+c+d = \sum_{i=1}^{I} \Big[x_i x'_i + x_i(1-x'_i) + (1-x_i)x'_i + (1-x_i)(1-x'_i)\Big] = \sum_{i=1}^{I} 1 = I$$

2) Como $x_i \in \{0,1\}$, verificamos que $x_i = x_i x_i$ logo $x_i x_i + (1-x_i)(1-x_i) = 1 - 2x_i + 2x_i x_i = 1$ e deduzimos

$$a(x,x) + d(x,x) = \sum_{i=1}^{I} \Big[x_i x_i + (1-x_i)(1-x_i)\Big] = \sum_{i=1}^{I} 1 = I$$

Por outro lado, temos sempre $x_i(1-x_i) = 0$ (experimentar os dois valores 0 e 1) logo

$$b(x,x) = c(x,x) = \sum_{i=1}^{I} x_i(1-x_i) = \sum_{i=1}^{I} 0 = 0.$$

3) $d(x,x') = I$ acontece se e somente se temos ambos $x_i = 0$ e $x'_i = 0$, $i = 1, \cdots, I$. No caso de atributos binarios exclusivos, o valor 1 deve aparecer só uma vez. Logo, temos pelo menos um $x_i = 1$ e não podemos ter todos os valores nulos. A contradição implique que $d(x,x') \neq I$.
4) No caso de atributos independentes podemos ter $d(x,x') = I$, logo $a = 0$ e $a+b+c = 0$. Por consequência, o quociente de $Jac(x,x')$ não faz sentido. No caso de atributos exclusivos, temos $d(x,x') < I$, logo $a+b+c = I-d > 0$. O quociente de $Jac(x,x')$ agora faz sempre sentido porque o denominador é sempre não nulo.
5) Como os atributos são exclusivos, $x = (x_1, \cdots, x_I)$ tem uma unica componente que vale 1 e as outras são zeros. Seja x' um outro evento. Se $x = x'$, deduzimos que $a(x,x') = 1$, $b(x,x') = c(x,x') = 0$ e $d(x,x') = I-1$, logo $Sok(x,x') = 1$ e $Jac(x,x') = 1$. Caso contrario, significa que a posição do valor 1 em x e different da posição em x'. Por consequência, temos $a(x,x') = 0$, $b(x,x') = c(x,x') = 1$, $d(x,x') = I-2$, logo $Sok(x,x') = \frac{I-2}{I}$ e $Jac(x,x') = 0$.
6) A diferença entre $Jac(x,x')$ e $Sok(x,x')$ é a introdução da quantidade d.

- Indice de Jaccard. Em primeiro lugar, o índice de Jaccard funciona apenas com atributos exclusivos. Neste contexto, a quantidade d não é util porque os valores zeros não são informativos e correspondem á consequência da exclusão. Finalmente, da questão 5), podemos observar que $Jac(x,x') = 1$ se $x = x'$ e $Jac(x,x') = 0$ se $x \neq x'$ o que corresponde à agrupar os atributos num unico atributo com a semelhança do *matching*.

- Indice de Sokal. Com dados binarios independentes, os valores zeros dos atributo são realmente novas informações logo é fundamental contabiliza-las quando dois eventos têm zeros em comum. No contexto de dados exclusivos, o índice de Sokal pode tomar apenas dois valores: $Sok(x,x') = 1$ quando $x = x'$ e $Sok(x,x') = \frac{I-2}{I}$ quando $x \neq x'$. Logo se I é grande, os dois valores são muito próximos quando $x = x'$ ou quando $x \neq x'$.

7) Temos

$$b(x,x') = \sum_{i=1}^{I} x_i(1-x'_i) = \sum_{i=1}^{I}(1-x'_i)x_i = c(x',x)$$

Logo $b(x,x') + c(x,x') = b(x',x) + c(x',x)$. Por outro lado, temos $a(x,x') = a(x',x)$ e $d(x,x') = d(x',x)$. Deduzimos que

$$Jac(x,x') = \frac{a(x,x')}{a(x,x') + b(x,x') + c(x,x')} = \frac{a(x',x)}{a(x',x) + c(x',x) + b(x',x)} = Jac(x',x)$$

e

$$Sok(x,x') = \frac{a(x,x') + d(x,x')}{I} = \frac{a(x',x) + d(x',x)}{I} = Sok(x',x),$$

e concluimos que as duas semelhanças são simetricas.
Se $Jac(x,x') = 1$, então temos $a(x,x') = a(x,x') + b(x,x') + c(x,x')$, logo $b(x,x') = c(x,x') = 0$. Isto implique que se $x_i = 0$, temos $x'_i = 0$ e se $x_i = 1$ temos $x'_i = 1$. Logo $x = x'$. Do mesmo modo, se $Sok(x,x') = 1$, temos $a(x,x') + d(x,x') = a(x,x') + b(x,x') + c(x,x') + d(x,x')$, logo $b(x,x') = c(x,x') = 0$. O raciocinho conduz a mesma conclusão que $x = x'$.
8) O calculo não envolve qualquer dificuldade. O índice de Jaccard é mais adequado para este tipo de dados que são subconjuntos de atributos exclusivos.