CADERNO DE EXERCÍCIOS – PARTE 02

5 CAPÍTULO 05: CLASSIFICAÇÃO DE DADOS

5.1 EXERCÍCIOS CONCEITUAIS

5.1.1 Durante o desenvolvimento de um modelo de classificação a base de dados é dividida em dois conjuntos. Quais são estes conjuntos e qual é o propósito desta separação

Resposta: Os dois conjuntos são: treinamento e testes. Seu propósito é gerar modelos preditivos capazes de identificar classes ou valores de registros não rotulados com um resultado satisfatório.

5.1.2 Discuta por quê para bases de dados reais na maioria das vezes treinar um sistema preditivo até que o erro para os dados de treinamento seja zero não é aconselhável.

Resposta: Durante o processo de treinamento de modelos, mesmo com o desempenho cada vez melhor e a quantidade de erros possa estar decaindo, o desempenho de generalização pode começar a se deteriorar após determinado número de iterações.
O ideal é utilizar uma validação cruzada para interromper o treinamento para conseguir obter um melhor resultado.

5.1.3 O que é a validação cruzada em k-pastas? Qual é a sua finalidade?

Resposta: A validação cruzada em k-pastas consiste em dividir a base de dados em k subconjuntos, sendo k –1 pastas para treinamento e 1 pasta para teste. Esse processo de treinamento e teste é repetido com todos os k subconjuntos, e a média dos desempenhos para as bases de treinamento e as bases de teste é adotada como indicador de qualidade do modelo.
A validação cruzada em k-pastas é bastante usual para se estimar o erro de generalização de preditores.

5.1.4 O que é a matriz de confusão de um problema de classificação binária? Explique os quatro valores apresentados pela matriz.

Resposta: Matriz de confusão é uma forma de apresentar integralmente o desempenho de um algoritmo de classificação binária utilizando uma matriz que relaciona as classes desejadas com as classes preditas.
Essa matriz também é conhecida como *matriz de contingência* ou *matriz de erro* e na sua composição ela tem nas linhas os objetos nas classes originais e nas colunas os objetos nas classes preditas.

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	VP	FN
	Negativa	FP	VN

Abaixo estão as explicações dos quatro valores apresentados na tabela.

- **VP (verdadeiro positivo):** você previu positivo e é verdadeiro
  Ex.: Você previu que uma mulher está grávida e ela realmente está.
- **VN (verdadeiro negativo):** você previu negativo e é verdadeiro
  Ex.: Você previu que um homem não está grávido e ele realmente não está.
- **FP (falso positivo):** você previu positivo e é falso
  Ex.: Você previu que um homem está grávido, mas na verdade ele não está.
- **FN (falso negativo):** você previu negativo e é falso
  Ex.: Você previu que uma mulher não está grávida, mas na verdade ela está.

5.2 EXERCÍCIOS NUMÉRICOS

5.2.1 Utilizando a árvore de decisão para o conjunto de treinamento da base de dados Cogumelos, apresentada na Figura 5.19, extraia todas as regras de classificação para cogumelos venenosos.

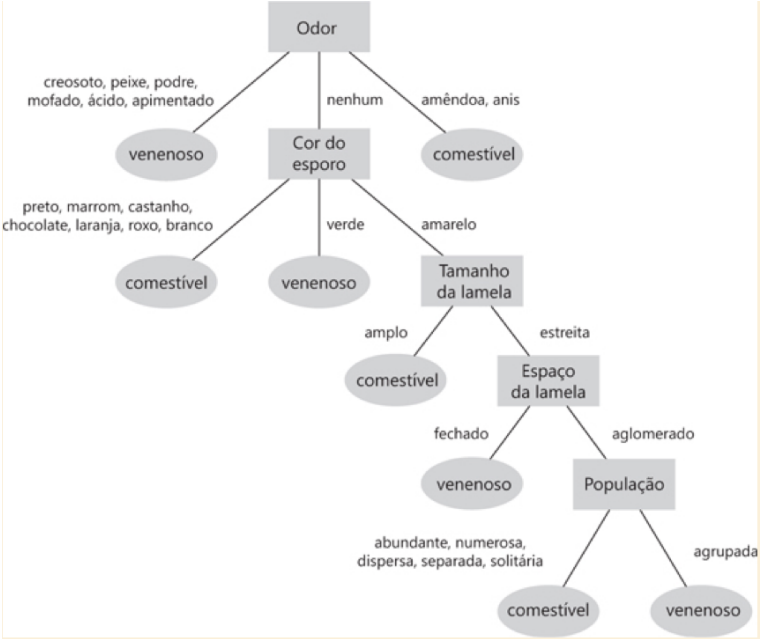


Figura 5.19

Resposta:

Atributo	Valor	Classe
odor	creosoto	venenoso
odor	peixe	venenoso
odor	podre	venenoso
odor	mofado	venenoso
odor	ácido	venenoso
odor	apimentado	venenoso
cor do esporo	verde	venenoso
espaço da lamela	fechado	venenoso
população	agrupada	venenoso

5.2.2 Considere o seguinte problema: Um determinado algoritmo é responsável por indicar se um paciente está doente ou não. Após a realização de um experimento com 150 pacientes, onde 30 estavam doentes, o algoritmo informou a existência de 28 doentes sendo que 8 destes estavam saudáveis. Calcule a taxa de verdadeiro positivo *(TPR)*, a taxa de falso positivo *(FPR)*, a acurácia *(ACC)*, e a precisão *(Pr)*.

Resposta:

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	20	10
	Negativa	8	112

Taxa de verdadeiro positivo (TVP):

$$TVP = \frac{VP}{VP + FN} = \frac{20}{30} = 0.6666 \cdot 100 = 66.66\%$$

Taxa de falso positivo (TFP):

$$TFP = \frac{FP}{FP + VN} = \frac{8}{120} = 0.0666 \cdot 100 = 6,67\%$$

Acurácia (ACC):

$$ACC = \frac{VP + VN}{VP + FP + VN + FN} = \frac{132}{150} = 0.88 \cdot 100 = 88\%$$

Precisão (PR):

$$Pr = \frac{VP}{FP + VP} = \frac{20}{28} = 0.7143 \cdot 100 = 71.43\%$$

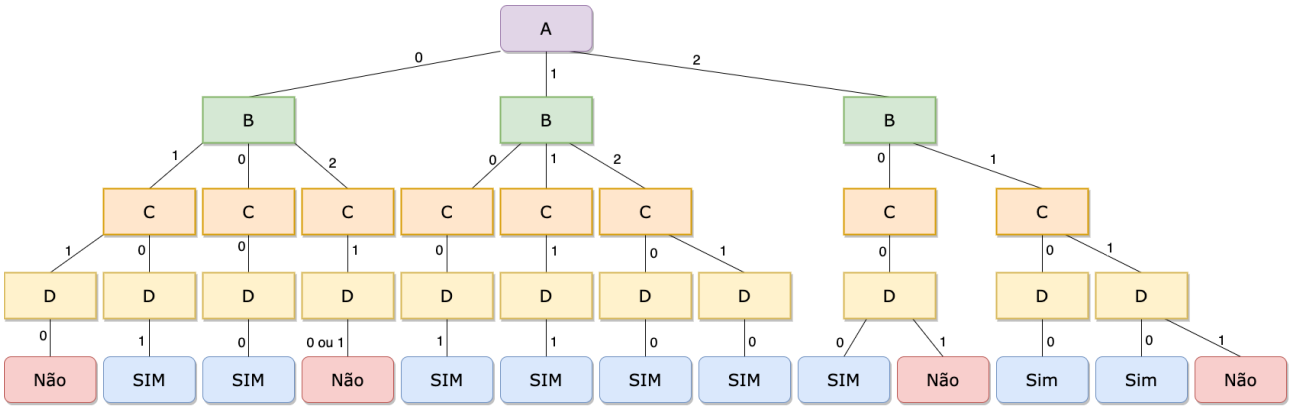
Conclusão:

$$TVP = 66.66\%$$
$$TFP = 6.67\%$$
$$ACC = 88\%$$
$$Pr = 71.43\%$$

5.2.3 Construa uma árvore de decisão para a base de dados abaixo.

A	B	C	D	Classe
2	1	1	0	Sim
2	0	0	0	Sim
2	0	0	1	Não
2	1	0	0	Sim
2	1	1	1	Não
0	2	1	0	Não
0	2	1	1	Não
0	1	1	0	Não
0	0	0	0	Sim
0	1	0	1	Sim
1	2	1	0	Sim
1	0	0	1	Sim
1	1	1	1	Sim
1	2	0	0	Sim

Resposta:



5.2.4 Aplique o classificador *Naïve Bayes* na base de dados abaixo e determine o valor para classe Jogar dos seguintes objetos:

a X = (Ensolarado, Branda, Alta, Não)

b Y = (Chuvoso, Fria, Alta, Sim)

c Z = (Ensolarado, Branda, Normal, Não)

d W = (Fechado , Fria, Normal, Sim)

Tempo	Tempera-tura	Umidade	Vento	Jogar
Ensolarado	Quente	Alta	Não	**Não**
Ensolarado	Quente	Alta	Sim	**Não**
Fechado	Quente	Alta	Não	**Sim**
Chuvoso	Branda	Alta	Não	**Sim**
Chuvoso	Fria	Normal	Não	**Sim**
Chuvoso	Fria	Normal	Sim	**Não**
Fechado	Fria	Normal	Sim	**Sim**
Ensolarado	Branda	Alta	Não	**Não**
Ensolarado	Fria	Normal	Não	**Sim**
Chuvoso	Branda	Normal	Não	**Sim**
Ensolarado	Branda	Normal	Sim	**Sim**
Fechado	Branda	Alta	Sim	**Sim**
Fechado	Quente	Normal	Não	**Sim**
Chuvoso	Branda	Alta	Sim	**Não**

Resposta: **Teorema de Bayes:** $P(C_i \mid x) = \dfrac{P(x \mid C_i)P(C_i)}{P(x)}$

Probabilidade por atributo:

Tempo	Jogar = Sim	Jogar = Não	Total
Ensolarado	$2/9 = 0.22$	$3/5 = 0.6$	$5/14 = 0.36$
Fechado	$4/9 = 0.44$	$0/5 = 0$	$4/14 = 0.28$
Chuvoso	$3/9 = 0.33$	$2/5 = 0.55$	$5/14 = 0.36$

Temperatura	Jogar = Sim	Jogar = Não	Total
Quente	$2/9 = 0.22$	$2/5 = 0.4$	$4/14 = 0.28$
Branda	$4/9 = 0.44$	$2/5 = 0.4$	$6/14 = 0.43$
Fria	$3/9 = 0.33$	$1/5 = 0.2$	$4/14 = 0.28$

Umidade	Jogar = Sim	Jogar = Não	Total
Alta	$3/9 = 0.33$	$4/5 = 0.8$	$7/14 = 0.5$
Normal	$6/9 = 0.67$	$1/5 = 0.2$	$7/15 = 0.5$

Vento	Joga = Sim	Jogar = Não	Total
Sim	$3/9 = 0.33$	$3/5 = 0.6$	$6/14 = 0.43$
Não	$6/9 = 0.67$	$2/5 = 0.4$	$8/14 = 0.57$

Probabilidade de Jogar (Sim/Não) $P(C)$**:** $P(Jogar = SIM) = 9/14 = 0.64$
$P(Jogar = NAO) = 5/14 = 0.36$

Problema: a) - X = (Ensolarado, Branda, Alta, Não):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Ensolarado \mid Jogar = Sim) = 0.22$
$P(Temperatura = Branda \mid Jogar = Sim) = 0.44$
$P(Umidade = Alta \mid Jogar = Sim) = 0.33$
$P(Vento = Nao \mid Jogar = Sim) = 0.67$
$P(Jogar = Sim) = 0.64$
$P(X \mid Jogar = Sim)P(Jogar = Sim) = 0.22 \cdot 0.44 \cdot 0.33 \cdot 0.67 \cdot 0.64 = 0.0137$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Ensolarado \mid Jogar = Nao) = 0.6$
$P(Temperatura = Branda \mid Jogar = Nao) = 0.4$
$P(Umidade = Alta \mid Jogar = Nao) = 0.8$
$P(Vento = Nao \mid Jogar = Nao) = 0.4$
$P(Jogar = Nao) = 0.36$
$P(X \mid Jogar = Nao)P(Jogar = Nao) = 0.6 \cdot 0.4 \cdot 0.8 \cdot 0.4 \cdot 0.36 = 0.0276$

Probabilidade do total dos atributos:
$P(X) = P(Tempo = Ensolarado) \cdot P(Temperatura = Branda) \cdot P(Umidade = Alta) \cdot P(Vento = Nao)$
$P(X) = 0.36 \cdot 0.43 \cdot 0.5 \cdot 0.57$
$P(X) = 0.0441$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid X) = \frac{0.0137}{0.0441} = 0.3106$
$P(Jogar = Nao) \mid X) = \frac{0.0276}{0.0441} = \mathbf{0.6258}$

Conclusão de X: Probabilidade de Jogar = **NÃO**

Problema: b) - Y = (Chuvoso, Fria, Alta, Sim):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Chuvoso \mid Jogar = Sim) = 0.33$
$P(Temperatura = Fria \mid Jogar = Sim) = 0.33$
$P(Umidade = Alta \mid Jogar = Sim) = 0.33$
$P(Vento = Sim \mid Jogar = Sim) = 0.33$
$P(Jogar = Sim) = 0.64$
$P(Y \mid Jogar = Sim)P(Jogar = Sim) = 0.33 \cdot 0.33 \cdot 0.33 \cdot 0.33 \cdot 0.64 = 0.0076$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Chuvoso \mid Jogar = Nao) = 0.6$
$P(Temperatura = Fria \mid Jogar = Nao) = 0.2$
$P(Umidade = Alta \mid Jogar = Nao) = 0.8$
$P(Vento = Sim \mid Jogar = Nao) = 0.6$
$P(Jogar = Nao) = 0.36$
$P(Y \mid Jogar = Nao)P(Jogar = Nao) = 0.6 \cdot 0.2 \cdot 0.8 \cdot 0.6 \cdot 0.36 = 0.0207$

Probabilidade do total dos atributos:
$P(Y) = P(Tempo = Chuvoso) \cdot P(Temperatura = Fria) \cdot P(Umidade = Alta) \cdot P(Vento = Sim)$
$P(Y) = 0.36 \cdot 0.28 \cdot 0.5 \cdot 0.43$
$P(Y) = 0.0217$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid Y) = \frac{0.0076}{0.0217} = 0.3502$
$P(Jogar = Nao) \mid Y) = \frac{0.0207}{0.0217} = \mathbf{0.9539}$

Conclusão de Y: Probabilidade de Jogar = **NÃO**

Problema: c) - Z = (Ensolarado, Branda, Normal, Não):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Ensolarado \mid Jogar = Sim) = 0.22$
$P(Temperatura = Branda \mid Jogar = Sim) = 0.44$
$P(Umidade = Normal \mid Jogar = Sim) = 0.67$
$P(Vento = Nao \mid Jogar = Sim) = 0.67$
$P(Jogar = Sim) = 0.64$
$P(Z \mid Jogar = Sim)P(Jogar = Sim) = 0.22 \cdot 0.44 \cdot 0.67 \cdot 0.67 \cdot 0.64 = 0.0278$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Ensolarado \mid Jogar = Nao) = 0.6$
$P(Temperatura = Branda \mid Jogar = Nao) = 0.4$
$P(Umidade = Normal \mid Jogar = Nao) = 0.2$
$P(Vento = Nao \mid Jogar = Nao) = 0.4$
$P(Jogar = Nao) = 0.36$
$P(Z \mid Jogar = Nao)P(Jogar = Nao) = 0.6 \cdot 0.4 \cdot 0.2 \cdot 0.4 \cdot 0.36 = 0.0069$

Probabilidade do total dos atributos:
$P(Z) = P(Tempo = Ensolarado) \cdot P(Temperatura = Branda) \cdot P(Umidade = Normal) \cdot P(Vento = Nao)$
$P(Z) = 0.36 \cdot 0.43 \cdot 0.5 \cdot 0.57$
$P(Z) = 0.0441$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid Z) = \dfrac{0.0278}{0.0441} = \mathbf{0.6303}$
$P(Jogar = Nao) \mid Z) = \dfrac{0.0069}{0.0441} = 0.1564$

Conclusão de Z: Probabilidade de Jogar = **SIM**

Problema: d) - W = (Fechado, Fria, Normal, Sim):

Probabilidade de Jogar=Sim por atributo:
$P(Tempo = Fechado \mid Jogar = Sim) = 0.44$
$P(Temperatura = Fria \mid Jogar = Sim) = 0.33$
$P(Umidade = Normal \mid Jogar = Sim) = 0.67$
$P(Vento = Sim \mid Jogar = Sim) = 0.33$
$P(Jogar = Sim) = 0.64$
$P(W \mid Jogar = Sim)P(Jogar = Sim) = 0.44 \cdot 0.33 \cdot 0.67 \cdot 0.33 \cdot 0.64 = 0.0109$

Probabilidade de Jogar=Nao por atributo:
$P(Tempo = Fechado \mid Jogar = Nao) = 0$
$P(Temperatura = Fria \mid Jogar = Nao) = 0.33$
$P(Umidade = Normal \mid Jogar = Nao) = 0.2$
$P(Vento = Sim \mid Jogar = Nao) = 0.6$
$P(Jogar = Nao) = 0.36$
$P(W \mid Jogar = Nao)P(Jogar = Nao) = 0 \cdot 0.33 \cdot 0.2 \cdot 0.6 \cdot 0.36 = 0$

Probabilidade do total dos atributos:
$P(W) = P(Tempo = Fechado) \cdot P(Temperatura = Fria) \cdot P(Umidade = Normal) \cdot P(Vento = Sim)$
$P(W) = 0.28 \cdot 0.28 \cdot 0.5 \cdot 0.43$
$P(W) = 0.0168$

Probabilidade de Jogar:
$P(Jogar = Sim) \mid W) = \dfrac{0.0109}{0.0168} = \mathbf{0.6488}$
$P(Jogar = Nao) \mid W) = \dfrac{0}{0.0168} = 0$

Conclusão de W: Probabilidade de Jogar = **SIM**