Universidade Federal de São Carlos
Aprendizado de Máquina 1 – 01/2022
Prof. Murilo Naldi

Lista de exercícios 3 – Classificadores

Adaptados do livro Introduction data mining / c2006 - (Livros) PANG-NING, Tan; STEINBACH, Michael; KUMAR, Vipin. Introduction data mining. Boston: Pearson Education, c2006. 769 p. ISBN 0-321-32136-7

1. Desenhe a árvore de decisão inteira para a função de paridade de quatro atributos boleanos, A, B, C e D. É possível simplificar a árvore?

2. Analise os exemplos de treinamento mostrados na Tabela 4.7 para um problema de classificação binária.

(a) Calcule o índice Gini para o conjunto geral de exemplos de treinamento.
(b) Calcule o índice Gini para o atributo ID do Cliente.
(c) Calcule o índice Gini para o atributo Sexo.
(d) Calcule o indice Gini para o atributo Tipo de Carro usando uma divisão múltipla.
(e) Calcule o índice Gini para o atributo Tamanho da Camisa usando uma divisão múltipla.
(f) Qual atributo é melhor, Sexo, Tipo do Carro ou Tamanho da Camisa?
(g) Explique por que a ID do Cliente não deve ser usada como a condição de teste de atributo embora tenha o Gini mais baixo.

Tabela 4.7. Conjunto de dados para o Exercício 2.

ID do Cliente	Sexo	Tipo de Carro	Tamanho de Camisa	Classe
1	M	Familiar	Pequeno	C0
2	M	Esportivo	Médio	C0
3	M	Esportivo	Médio	C0
4	M	Esportivo	Grande	C0
5	M	Esportivo	Extra Grande	C0
6	M	Esportivo	Extra Grande	C0
7	F	Esportivo	Pequeno	C0
8	F	Esportivo	Pequeno	C0
9	F	Esportivo	Médio	C0
10	F	de Luxo	Grande	C0
11	M	Familiar	Grande	C1
12	M	Familiar	Extra Grande	C1
13	M	Familiar	Médio	C1
14	M	de Luxo	Extra Grande	C1
15	F	de Luxo	Pequeno	C1
16	F	de Luxo	Pequeno	C1
17	F	de Luxo	Médio	C1
18	F	de Luxo	Médio	C1
19	F	de Luxo	Médio	C1
20	F	de Luxo	Grande	C1

3. Analise os exemplos de treinamento mostrados na Tabela 4.8 para um problema de classificação binária.

(a) Qual a entropia deste conjunto de exemplos de treinamento com respeito à classe positiva?
(b) Quais os ganhos de informação de a_1 e a_2 relativos a estes exemplos de treinamento?
(c) Para a_3, que é um atributo contínuo, calcule o ganho de informação para cada divisão possível.
(d) Qual a melhor divisão (entre a_1, a_2 e a_3) de acordo com o ganho de informação?
(e) Qual a melhor divisão (entre a_1, a_2) de acordo com a taxa de erros de classificação?
(f) Qual a melhor divisão (entre a_1, a_2) de acordo com o índice Gini?

Tabela 4.8. Conjunto de dados para o Exercício 3.

Instância	a_1	a_2	a_3	Classe Alvo
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

4. Mostre que a entropia de um nodo nunca aumenta após a divisão em um número de nodos sucessores menor.

5. Analise o seguinte conjunto de dados para um problema de classe binária.

A	B	Rótulo da Classe
V	F	+
V	V	+
V	V	+
V	F	−
V	V	+
F	F	−
F	F	−
F	F	−
V	V	−
V	F	−

(a) Calcule o ganho de informação na divisão sobre A e B. Que atributo o algoritmo

de indução de árvore de decisão escolheria?

(b) Calcule o ganho no índice Gini na divisão sobre A e E. Que atributo o algoritmo de indução de árvore de decisão escolheria?

(c) A Figura 4.13 mostra que entropia e índice de Gini estão ambos aumentando monotonamente na faixa de [O, 0.5] e que estão decrescendo monotonamente na faixa [0.5, 1]. É possível que o ganho de informação no índice Gini favoreça atributos diferentes? Explique.

6. Analise o seguinte conjunto de exemplos de treinamento.

X	Y	Z	Nº de Exemplos da Classe C1	Nº de Exemplos da Classe C2
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

(a) Calcule uma árvore de decisão de dois níveis usando a abordagem ávida. Use a taxa de erros de classificação como critério de divisão. Qual a taxa de erro geral da árvore induzida?

(b) Repita a parte (a) usando X como o primeiro atributo de divisão e a seguir escolha o melhor atributo restante para divisão em cada um dos nodos sucessores. Qual a taxa de erro da árvore induzida?

(e) Compare os resultados das partes (a) e (b). Comente a respeito da conveniência da heurística usada para seleção de atributos de divisão.

7. A tabela a seguir resume um conjunto de dados com três atributos A, B e C e dois rótulos de classe +, -. Construa uma árvore de decisão de dois níveis.

A	B	C	Número de Instâncias	
			+	−
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

(a) De acordo com a taxa

de erros de classificação, qual atributo seria escolhido como o primeiro atributo de divisão? Para cada atributo, mostre a tabela de contingência e os ganhos na taxa de erros de classificação.

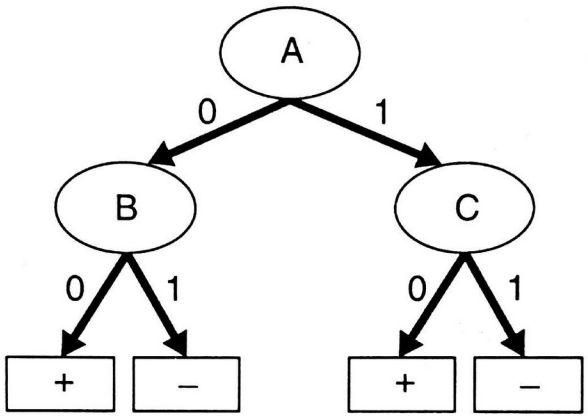
(b) Repita para os dois filhos do nodo raiz.

(e) Quantas instâncias estão mal classificadas pela árvore de decisão resultante?

(d) Repita as partes (a), (b) e (c) usando C como atributo de divisão.

(e) Use os resultados nas partes (e) e (d) para concluir sobre a natureza ávida do algoritmo de indução de árvore de decisão.

8. Analise a árvore de decisão mostrada a seguir.



Treinamento:

Instância	A	B	C	Classe
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	–
5	1	0	0	+
6	1	0	0	+
7	1	1	0	–
8	1	0	1	+
9	1	1	0	–
10	1	1	0	–

Validação:

Instância	A	B	C	Classe
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	–
15	1	0	0	+

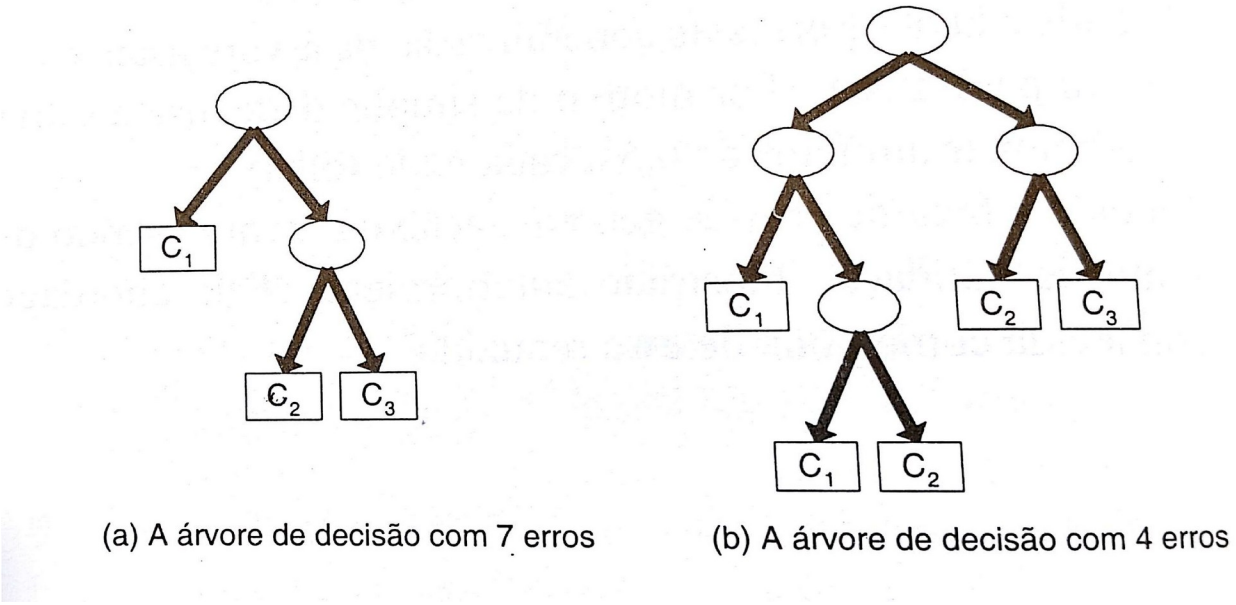
(a) Calcule a taxa de erros de generalização da árvore usando a abordagem pessimista. (Por motivo de simplicidade, use a estratégia de adicionar um fator de 0,5 a cada nodo folha).

(b) Calcule a taxa de erro de generalização da árvore usando o conjunto de validação mostrado anteriormente. Esta abordagem é conhecida como poda de erro reduzido.

9. Analise as árvores de decisão mostradas. Suponha que elas sejam geradas a partir de um conjunto de dados que contenha 16 atributos binários e 3 classes: C1, C2 e C3.

(a) Árvore de decisão com 7 erros.

(b) Árvore de decisão com 4 erros.



(a) A árvore de decisão com 7 erros

(b) A árvore de decisão com 4 erros

Calcule o comprimento total da descrição de cada árvore de decisão de acordo com o princípio do comprimento mínimo de descrição.

- O comprimento mínimo de uma árvore é dado por:

$$\text{Custo(árvore, dados)} = \text{Custo (árvore)} + \text{Custo(dados|árvore)}$$

- Cada nodo interno da árvore é codificado pelo ID do atributo de divisão. Se houve m atributos, o custo de codificar cada atributo é log2m bits.
- Cada folha é codificada usando o ID da classe com a qual está associada. Se houver k classes, o custo de codificar uma classe é log2k bits.
- O Custo(árvore) é o custo de codificar todos os nodos da árvore. Para simplificar o cálculo, você pode supor que o custo total da árvore seja obtido somando-se os custos de codificar cada nodo interno e cada nodo folha.
- O Custo(dados|árvore) é codificado usando os erros de classificação que a árvore comete no conjunto de treinamento. Cada erro é codificado por log2n bits, onde n é o número total de instâncias de treinamento.

Qual árvore de decisão é melhor, de acordo com o princípio do comprimento mínimo de descrição?

10. Embora a abordagem bootstrap seja útil para se obter uma avaliação confiável da precisão do modelo, ela possui uma limitação. Analise um problema de duas classes, onde há um número igual de exemplos positivos e negativos nos dados. Suponha que os rótulos de classe para os exemplos sejam gerados aleatoriamente. O classificador usado é uma árvore de decisão não podada (i.e., um memorizador perfeito).

Determine a precisão do classificador usando cada um dos seguintes métodos.

(a) O método holdout, onde dois terços dos dados são usados para treinamento e o terço restante é usado para teste.

(b) A validação cruzada de 10 vezes.

(c) O método bootstrap.

(d) A partir dos resultados das partes (a), (b) e (c), qual método fornece uma avaliação mais confiável da precisão do classificador?

11. Analise a seguinte abordagem para testar se um Classificador A é melhor que outro Classificador B. Seja N o tamanho de um determinado conjunto de dados, pA a precisão do Classificador A, pB a precisão do Classificador B e p = (pA + pB)/2 a precisão média de ambos os classificadores. Para testar se o Classificador A é significativamente melhor do que o B, a seguinte estatística Z é usada:

$$Z = \frac{p_A - p_B}{\sqrt{\frac{2p(1-p)}{N}}}.$$

Supõe-se que o Classificador A seja melhor do que o Classificador B se Z > 1,96.

A Tabela 4.9 compara as precisões de três diferentes classificadores, classificadores de árvore de decisão, classificadores Bayes simples e máquinas de suporte de vetores, sobre vários conjuntos de dados.

Resuma o desempenho dos classificadores dados na Tabela 4.9 usando a seguinte tabela 3 x 3:

Vitórias, derrotas e empates	Árvore de decisão	Bayes simples	Máquina de suporte de vetores
Árvore de decisão	0 - 0 - 23		
Bayes simples		0 - 0 - 23	
Máquina de suporte de vetores			0 - 0 - 23

Cada célula contém o número de Vitórias, derrotas e empates quando comparando o classificador de uma determinada linha com o de uma determinada coluna.

Tabela 4.9. Comparando a precisão de diversos métodos de classificação.

Conjunto de Dados	Tam. (N)	Árvore de Decisão(%)	Bayes Simples (%)	Máquina de suporte de vetores (%)
Anneal	898	92.09	79.62	87.19
Australia	690	85.51	76.81	84.78
Auto	205	81.95	58.05	70.73
Tórax	699	95.14	95.99	96.42
Cleve	303	76.24	83.50	84.49
Crédito	690	85.80	77.54	85.07
Diabetes	768	72.40	75.91	76.82
Alemão	1000	70.90	74.70	74.40
Vidro	214	67.29	48.59	59.81
Coração	270	80.00	84.07	83.70
Hepatite	155	81.94	83.23	87.10
Cavalo	368	85.33	78.80	82.61
Ionosfera	351	89.17	82.34	88.89
Iris	150	94.67	95.33	96.00
Trabalho	57	78.95	94.74	92.98
Led7	3200	73.34	73.16	73.56
Linfografia	148	77.03	83.11	86.49
Pima	768	74.35	76.04	76.95
Sonar	208	78.85	69.71	76.92
Jogo da velha	958	83.72	70.04	98.33
Veículo	846	71.04	45.04	74.94
Vinhos	178	94.38	96.63	98.88
Zoológico	101	93.07	93.07	96.04