ICMC-USP

Lista de Exercícios - Capítulo 8 [1]
SCC-630 - Inteligência Artificial

1o. Semestre de 2011 - Prof. João Luís

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
http://www.icmc.usp.br
Departamento de Ciências de Computação

1. Seja breve na resposta às seguintes questões:
   (a) o que você entende por Aprendizado de Máquina (AM)?
   (b) qual é a diferença fundamental entre AM simbólico e não-simbólico?
   (c) caracterize AM supervisionado, não supervisionado e semissupervisionado.
   (d) caracterize AM incremental e não incremental.
   (e) dado um conjunto de exemplos de treinamento, que se entende por erro majoritário de um classificador?
   (f) qual é a diferenca fundamental entre a taxa de erro aparente e a estimativa da taxa de erro verdadeira de um classificador?

2. Defina quando uma hipótese induzida por um algoritmo de AM supervisionado é:
   (a) Completa e Consistente
   (b) Completa e Inconsistente
   (c) Incompleta e Consistente
   (d) Incompleta e Inconsistente

References

[1] Monard, M.C.; Metz, J. (PAE), Listas de Exercícios 5 e 6 - SCC-0630 Inteligência Artificial. 2010.

[2] Rezende, S. O. (org.), *Sistemas Inteligentes - Fundamentos e Aplicações*. Editora Manole, 2003.

As próximas 5 páginas contêm a lista de exercícios sobre Aprendizado de Máquina da Profa. Maria Carolina Monard [1]. Utilize a referência [2], capítulos 4 e 5.

Continua na próxima página...

Instituto de Ciências Matemáticas e de Computação

Curso: **SCC-0630 Inteligência Artificial**
http://agora.tidia-ae.usp.br
Responsável: **Profa. Maria Carolina Monard** mcmonard@icmc.usp.br
Estagiário PAE: **Jean Metz** metzz@icmc.usp.br
Semestre: 1^{ro} **de 2010**

Lista 5

1. É requerido diagnosticar um problema na linha de produção automobilística de uma empresa, na qual a maior causa do problema é o motor defeituoso. Um motor defeituoso é determinado pelos atributos e valores mostrados na Tabela 1. A classe identifica o estado do motor, que neste exemplo é $classe = \{bom,\ ruim\}$.

Atributo	Vvalor
Rotação do Motor	$\{baixa, normal, alta\}$
Tempo de Uso	$\{pouco, medio, muito\}$
Temperatura	$\{baixa, normal, alta\}$

Tabela 1: Atributos e valores para diagnóstico do estado de um Motor

A Tabela 2 mostra o conjunto de observações (exemplos de treinamento), ou seja, eventos passados observados que relacionam os valores dos atributos ao estado do motor.

#Id.	Rotação do Motor	Tempo de Uso	Temperatura	Classe
1	baixo	muito	alta	ruim
2	baixo	muito	normal	ruim
3	normal	pouco	normal	bom
4	normal	muito	alta	ruim
5	alto	muito	alta	ruim
6	alto	pouco	normal	bom
7	normal	pouco	normal	bom
8	baixo	pouco	alta	ruim
9	baixo	pouco	alta	ruim

Tabela 2: Conjunto de exemplos de treinamento - Motor

A Tabela 3 mostra um conjunto de exemplos de treinamento utilizados por um algoritmo de AM *i.e.* o número de exemplos (#Ex.), número e porcentagem de exemplos duplicados (que aparecem mais de uma vez) e/ou conflitantes (mesmos valores dos atributos mas classificações diferentes), número de atributos (#Atrib.) contínuos e nominais, distribuição das classes, erro majoritário e se o conjunto de dados tem pelo menos um valor desconhecido (Valores Desconh.)

Conj. de Dados	# Ex.	Duplic. ou Conflit.	# Atrib. (cont.,nom.)	Classe	% Classe	Erro Majoritário	Valores Desconh.
mistério	500	0 (0%)	15 (5,10)	"classe 1"	30.0%	30.0%	N
				"classe 2"	70.0%		

Tabela 3: Resumo de características de um conjunto de dados

(a) Descreva o conjunto de treinamento — Tabela 2 — de maneira semelhante à mostrada na Tabela 3

(b) Utilizando como linguagem de descrição a linguagem de Árvores de Decisão (AD) induza a hipótese correspondente utilizando ganho de informação para selecionar os atributos em cada nó da árvore.

(c) Escreva as regras de decisão disjuntas que representam a AD induzida.

(d) Calcule o erro aparente do classificador, *i.e.* utilizando o conjunto de treinamento — Tabela 2

(e) Considerando que a classe *bom* corresponde a exemplos + e *ruim* a exemplos − do conceito a ser aprendido, determine o tipo de hipótese encontrada considerando a cobertura da AD no conjunto de exemplos de treinamento da Tabela 2. Ou seja, se a hipótese induzida é:

i. Completa e Consistente
ii. Completa e Inconsistente
iii. Incompleta e Consistente
iv. Incompleta e Inconsistente

2. Conseguir predizer a atuação dos usuários é uma das aplicações de comercio eletrônico. Considere o conjunto de dados da Tabela 4 usado para predizer se um usuário solicitará mais informação (Mais-info) levando em conta se o acesso está sendo realizado desde um domínio relacionado com educação (Edu.), se esta é a primeira visita (Primeira), se já tem realizado compras com uma companhia afiliada (Comprou) e se já visitou um determinado site para compras on-line (Visit.)

#Id.	**Comprou**	**Edu.**	**Primeira**	**Visit.**	**Mais-info**
1	false	true	false	false	true
2	true	false	true	false	false
3	false	false	true	true	true
4	false	false	true	false	false
5	false	false	false	true	false
6	true	false	false	true	true
7	true	false	false	false	true
8	false	true	true	true	false
9	false	true	true	false	false
10	true	true	true	false	true
11	true	true	false	true	true
12	false	false	false	false	true

Tabela 4: Conjunto de exemplos de treinamento- Comercio eletrônico

Utilizando ganho de informação para selecionar os atributos em cada nó da árvore de decisão, encontre:

(a) A melhor AD com apenas um nó. Qual o erro aparente dessa AD?

(b) A melhor AD com apenas a raíz e os filhos da raíz. Qual o erro aparente dessa AD?

(c) A melhor AD sem restrição na altura da árvore. Qual o erro aparente dessa AD?

3. Considere a definição de *matriz de confusão* a seguir:

> Uma matriz de confusão mostra o número de classificações corretas em oposição ao número de classificações preditas para uma classe. Considerando problemas de classificação binária, ou seja, problemas de classificação com apenas duas classes que geralmente são rotuladas como "+" e "−", as escolhas são estruturadas para predizer a ocorrência ou não-ocorrência de um simples evento ou hipótese.

Quando apenas duas classes são consideradas, os dois erros possíveis são denominados de *falsos positivos* e *falsos negativos*. A Tabela 5 ilustra a matriz de confusão para problemas de classificação com duas classes onde Tp é o número de exemplos corretamente classificados como exemplos positivos, Fn é o número de exemplos errôneamente classificados como positivos, Tn é o número de exemplos corretamente classificados como exemplos negativos, Fn é o número de exemplos errôneamente classificados como negativos e $n = (Tp + Fn + Fp + Tn)$ é o total de exemplos.

Ainda considerando a Tabela 5, quatro situações podem ocorrer:

- O exemplo pertence a classe C_+ e é predito pelo classificador como da classe C_+. Neste caso, o exemplo é um *verdadeiro positivo*.
- O exemplo pertence a classeC_- e é predito pelo classificador como da classe C_-. Neste caso, o exemplo é um *verdadeiro negativo*.
- O exemplo pertence a classe C_- e é predito pelo classificador como da classe C_+. Neste caso, o exemplo é um *falso positivo*.
- O exemplo pertence a classe C_+ e é predito pelo classificador como da classe C_-. Neste caso, o exemplo é um *falso negativo*.

Classe	preditos como C_+	preditos como C_-	Precisão da Classe	Precisão Total
	$Verdadeiros\ positivos$	$Falsos\ negativos$	$\frac{Tp}{Tp+Fn}$	
C_+	Tp	Fn		$\frac{Tp+Tn}{n}$
	$Falsos\ positivos$	$Verdadeiros\ negativos$		
C_-	Fp	Tn	$\frac{Tn}{Fp+Tn}$	

Tabela 5: Matriz de confusão para problemas de classificação binária

Seja o conjunto de exemplos de treinamento ilustrado na Tabela 6. Esse conjunto de dados contém informações sobre medições diárias sobre as condições meteorológicas e uma classificação se o dia é apropriado a uma visita à fazenda ("go") ou não ("dont_go"). Cada exemplo é composto pelos seguintes atributos:

- outlook: assume os valores *sunny*, *overcast* ou *rain*;
- temperature: um valor numérico que indica a temperatura em graus *Celsius*;
- humidity: um valor numérico que indica a umidade relativa do ar;
- windy: assume os valores *yes* ou *no*.

Outlook	Temperature	Humidity	Windy	Voyage?
rain	27	95	no	go
rain	20	70	yes	dont_go
rain	23	80	yes	dont_go
rain	25	81	no	go
rain	21	80	no	go
sunny	25	72	yes	go
sunny	21	79	yes	dont_go
sunny	26	70	no	go
sunny	27	92	no	dont_go
sunny	30	88	no	dont_go
overcast	23	90	yes	go
overcast	29	78	no	go
overcast	19	65	yes	dont_go
overcast	26	75	no	go
overcast	20	87	yes	dont_go

Tabela 6: Conjunto de exemplos de treinamento *voyage*

Utilizando o conjunto de dados na Tabela 6 como conjunto de treinamento, foi induzida uma AD descrita pelas regras mostradas na Tabela 7.

```
IF outlook = overcast
THEN CLASS = go

IF outlook = sunny
            AND humidity ≤ 78
THEN CLASS = go

IF outlook = sunny
            AND humidity > 78
THEN CLASS = dont_go

IF outlook = rain
            AND windy = yes
THEN CLASS = dont_go

IF outlook = rain
            AND windy = no
THEN CLASS = go

CLASS = go
```

Tabela 7: Classificador AD - Regras induzidas *voyage*

(a) Descreva o conjunto de treinamento — Tabela 6 — de maneira semelhante à mostrada na Tabela 3

(b) Construa a matriz de confusão usando os exemplos de treinamento. Qual o erro desse classificador sobre o conjunto de treinamento (Tabela 6)?

(c) Construa a matriz de confusão usando os exemplos de teste. Qual o erro desse classificador sobre o conjunto de teste (Tabela 8)?

4. Suponha que você é informado que a estimativa do erro **verdadeiro** de um classificador é de 2%. Com somente essa informação é possível afirmar que o erro é pequeno, ou seja, que o algoritmo de AM que induziu esse classificador aprendeu o conceito com uma boa

Outlook	Temperature	Humidity	Windy	Voyage?
sunny	25	72	yes	go
sunny	28	91	yes	dont_go
sunny	22	70	no	go
sunny	23	95	no	dont_go
sunny	30	85	no	dont_go
overcast	23	90	yes	go
overcast	29	78	no	go
overcast	19	65	yes	dont_go
overcast	26	75	no	go
overcast	20	87	yes	go
rain	22	95	yes	go
rain	19	70	yes	dont_go
rain	23	80	yes	dont_go
rain	25	81	no	go
rain	21	80	no	go

Tabela 8: Conjunto de exemplos de teste *voyage*

precisão, ou é fundamental mais informação? Qual(is) é(são) essa(s) informação(ões)? Justifique sua resposta.

5. Responda às seguintes perguntas relacionadas a métodos de avaliação em AM.

   (a) Qual o problema em se estimar o erro de um classificador usando o conjunto de treinamento (o qual foi usado para obter o classificador)?

   (b) Por que métodos de avaliação como *cross-validation* e *leave-one-out* são necessários?

   (c) Além de avaliar o desempenho dos classificadores produzidos por um algoritmo de aprendizado, você consegue pensar em outras características de um algoritmo de aprendizado que são passíveis de avaliação?