Universidade do Minho
Departamento de Informática
Mestrado de Informática
Mineração de Dados
(UCE30 - SSD)

Teste Teórico **20 de Junho de 2008**

Duração: 2.00 h

Indicações da Correcção deste teste com indicações de resposta

1. Considere a árvore de decisão da página 3, geradas pelo sistema C4.5. Considerando este modelo responda:

   i. Diga qual a previsão derivada desta árvore para o novo caso:

   $$\mathsf{A9} = \mathsf{t},\ \mathsf{A15} = 200,\ \mathsf{A11} = 7$$

   Qual é a proporção de erros para a previsão anterior ? Qual a razão para este número ser decimal ? E qual é o número de casos de treino que cobrem o caminho percorrido pela previsão anterior ?

   Resposta: o caminho seguida para a previsão nesta árvore corresponde ao último ramo (última linha da pag 3). A previsão seria classe '+'. 10 casos de treino chegam a esta folha dos quais 2.4 estão errados. Proporção de erros = 2.4 /10. O valor de erros é decimal porque (duas razões): ou há nulos no treino, ou há pruning o que leva a estimação do erro nas folhas. Número de casos de treino que cobrem previsão anterior = 10 - 2.4.

   ii. Apresente na forma de regra de classificação o ramo da árvore usado para classificar o exemplo anterior. Não esquecer de apresentar o suporte e confiança dessa regra.

   REs: A9 = t & A15 > 5 & A15 < 228 & A11 > 3 → classe = '+'. sup = (10 - 2.4) / 490, conf = (10 - 2.4) / 10. Há simplificação das condições em A15.

   iii. Se o valor do atributo A15 = ? (nulo) que procedimento teriamos de usar ? Explique sucintamente o procedimento de previsão para este caso de teste.

   REs: explicar sucintamente o procedimento de distribuição de classes para previsões onde um teste na árvore não tem valor de comparação i.e. é nulo.

iv. Explique sucintamente o porquê do termo simplified usado na designação desta árvore. Apresente argumentos para defender o uso da aplicação desta simplificação.

Res: simplificação da árvore i.e. pruning. explicar a capacidade do pruning em combater overfitting.

v. Por decomposição do erro associado a esta árvore obtém-se os seguintes valores: Bias = 0.009, Var = 0.04, Noise = 0.0. Aplicando *Naive Bayes* no mesmo conjunto de treino obtém-se: Bias = 0.04, Var = 0.02, Noise = 0.0. Que conclusões podemos tomar sobre estes resultados ?

Res: Estamos perante um problema de viez pois em comparação o erro (soma das 2 componentes) é maior no NB e neste algoritmo o viez é alto. No entanto a variância baixa para metade quando mudamos de C4.5 para NB. Ou seja, a maior componente é viez o que faz explicar a degradação do erro quando mudamos de C4.5. para NB.

vi. Estranhamente, o valor do erro obtido com o modelo after pruning é maior que o de before pruning. Aponte as principais razões para o facto.

Res: a razão destes resultados deve-se ao facto de a avaliação ser feita com o conj. de treino. O overfitting permite obter melhores resultados na árvore completa quando comparada com a a¿vore podada.

2. Para um conjunto de teste T dois modelos A e B têm os seguintes desempenhos (em relação às classes):

```
          Modelo A
Precision   Recall  F-Measure   Class
  1         0.98      0.99      c1
  0.94      0.94      0.94      c2
  0.941     0.96      0.95      c3

          Modelo B
Precision   Recall  F-Measure   Class
  1         1         1         c1
  0.902     0.92      0.911     c2
  0.918     0.9       0.909     c3
```

Que conclusões pode tirar sobre estes resultados ?

Res: apresentar um comparativo por classe do desempenho dos 2 modelos. Interpretar o significado de precision e recall. Apontar qual o melhor modelo por cada classe.

3. Considere as seguintes regras de associação (assuma que são geradas a partir de um dataset de 1000 registos). O consequente de cada regra refere-se ao valor de colesterol no sangue da população:

```
(s=0.02,cf=0.89)    Col=high <--  diabetes=yes & A=1 & blood_type=2
(s=0.05,cf=0.90)    Col=high <--  diabetes=yes & blood_type=2
(s=0.10,cf=0.88)    Col=high <--  blood_type=2
(s=0.40,cf=0.40)    Col=high <--
(s=0.03,cf=0.91)    Col=med  <--  blood_type=1 & A=2
(s=0.01,cf=0.95)    Col=med  <--  X=2 & A=2 & blood_type=1
(s=0.03,cf=0.99)    Col=med  <--  X=2 & A=2
```

3.1. Descreva formas de eliminar redundância entre as regras descritas e apresente as regras finais a considerar. Como se caracteriza a população com alto teor de colesterol no sangue?

Res: Aplicar improvement sobre as regras. Eliminar regras não productivas e.g. 1º regra redundante em relação à 2º porque desce o valor da cf. Com as regras que passam o filtro de improvement (são só 2 para col=high) caracterizar essa subpopulação e.g. são diabéticos e tem sangue tipo 2.

3.2 Sabendo que o *teste de Fisher* entre a segunda regra e a terceira regra dá um $pvalue = 0.05556709$. Que conclusões deve tirar deste teste? Que acção deve ser tomada?

Res: A 2º regra não passa o teste (pvalue $>$ 0.05). Esta deve ser descartada (eliminada).

3.3. Apresente algumas vantagens da medida de interesse conviction em relação à confiança. Calcule o valor da conviction para a primeira regra e comente esse resultado.

Nota: $conv(A \rightarrow C) = \frac{1-s(C)}{1-conf(A \rightarrow C)}$

Res: Têm em conta sup do consequente e captura independência entre antecedente e consequente. Tenta representar "implicação".

conv = (1 - 0.40) / (1 - 0.89).

4. Um analista de dados decidiu tratar um problema de previsão numérica como um problema de classificação. A conversão foi obtida pela discretização do atributo objectivo. Elabore um pequeno ensaio (não mais que uma página) sobre as consequências de tal decisão, nomeadamente em termos de precisão do modelo gerado, caracteristica do problema em análise, sobreajustamento, medidas de avaliação, etc.

Deve ser referido o problema da definição dos intervalos do atributo classe numérico. Referir a necessidade de usar outras medidas de erro. Sobreajustamento pode surgir com intervalos muito pequenos. Deve tb ser descrito como é efectuada a previsão. Previsão é o ponto médio dos intervalos.

```
Read 490 cases (15 attributes) from Data/crx.data
Simplified Decision Tree:

A9 = f: - (239.0/19.4)
A9 = t:
|   A15 > 228 : + (106.0/3.8)
|   A15 <= 228 :
|   |   A11 <= 3 :
|   |   |   A4 = l: + (0.0)
|   |   |   A4 = t: + (0.0)
|   |   |   A4 = u:
|   |   |   |   A7 = h: + (18.0/1.3)
|   |   |   |   A7 = j: - (1.0/0.8)
|   |   |   |   A7 = n: + (0.0)
|   |   |   |   A7 = z: + (1.0/0.8)
|   |   |   |   A7 = dd: + (0.0)
|   |   |   |   A7 = ff: - (1.0/0.8)
|   |   |   |   A7 = o: + (0.0)
|   |   |   |   A7 = v:
|   |   |   |   |   A14 <= 110 : + (18.0/2.5)
|   |   |   |   |   A14 > 110 :
|   |   |   |   |   |   A15 > 8 : + (4.0/1.2)
|   |   |   |   |   |   A15 <= 8 :
|   |   |   |   |   |   |   A6 = c: - (4.0/2.2)
|   |   |   |   |   |   |   A6 = d: - (2.0/1.0)
|   |   |   |   |   |   |   A6 = cc: + (2.0/1.8)
|   |   |   |   |   |   |   A6 = i: - (0.0)
|   |   |   |   |   |   |   A6 = j: - (0.0)
|   |   |   |   |   |   |   A6 = k: - (2.0/1.0)
|   |   |   |   |   |   |   A6 = r: - (0.0)
|   |   |   |   |   |   |   A6 = x: - (0.0)
|   |   |   |   |   |   |   A6 = e: - (0.0)
|   |   |   |   |   |   |   A6 = ff: - (0.0)
|   |   |   |   |   |   |   A6 = m:
|   |   |   |   |   |   |   |   A13 = g: + (2.0/1.0)
|   |   |   |   |   |   |   |   A13 = p: - (0.0)
|   |   |   |   |   |   |   |   A13 = s: - (5.0/1.2)
|   |   |   |   |   |   |   A6 = q:
|   |   |   |   |   |   |   |   A12 = t: + (4.0/1.2)
|   |   |   |   |   |   |   |   A12 = f: - (2.0/1.0)
|   |   |   |   |   |   |   A6 = w:
|   |   |   |   |   |   |   |   A12 = t: - (2.0/1.0)
|   |   |   |   |   |   |   |   A12 = f: + (3.0/1.1)
|   |   |   |   |   |   |   A6 = aa:
|   |   |   |   |   |   |   |   A2 <= 41 : - (3.0/1.1)
|   |   |   |   |   |   |   |   A2 > 41 : + (2.0/1.0)
|   |   |   |   A7 = bb:
|   |   |   |   |   A14 <= 164 : + (3.4/1.5)
|   |   |   |   |   A14 > 164 : - (5.6/1.2)
|   |   |   A4 = y:
|   |   |   |   A13 = p: - (0.0)
|   |   |   |   A13 = s: + (2.0/1.0)
|   |   |   |   A13 = g:
|   |   |   |   |   A14 <= 204 : - (16.0/2.5)
|   |   |   |   |   A14 > 204 : + (5.0/2.3)
|   |   A11 > 3 :
|   |   |   A15 <= 4 : + (25.0/1.3)
|   |   |   A15 > 4 :
|   |   |   |   A15 <= 5 : - (2.0/1.0)
|   |   |   |   A15 > 5 : + (10.0/2.4)

Evaluation on training data (490 items):
         Before Pruning           After Pruning
        ----------------   --------------------------
        Size      Errors   Size      Errors   Estimate

          90   19( 3.9%)     58   24( 4.9%)    (11.9%)   <<
```