Universidade do Minho
Departamento de Informática
Mestrado em Informática e Curso de Especialização
Mestrado em Sistemas de Dados e Processamento Analítico
Extracção de Conhecimento em Bases de Dados

Exame de Época Especial **20 de Dezembro de 2007**

Duração: 2.00 h

Leia atentamente as questões propostas que se seguem. Tente ser claro e objectivo nas suas respostas. Clarifique sempre que possível com exemplos.

1. Considere a árvore de decisão da figura 1:

Apresente as previsões obtidas nas árvores (original e simplificada) para o novo caso:

```
physician fee freeze=u; water project cost sharing=u; mx missile=y
```

Descreva as regras de decisão que podem ser obtidas das duas árvores usadas nas previsões anteriores. Indique o suporte e confiança das regras. Qual lhe parece ser a árvore com melhor desempenho? Justifique.

2. Considere o Algoritmo Apriori. Aplique o algoritmo ao seguinte log de um website (clickstream data):

TID	Produto
1	111
1	121
2	131
2	141
2	151
3	111
3	121
3	131
4	121
4	131
4	151

usando $minconf = 0.8$ e $minsup = 0.25$. Apresente todos os passos intermédios bem como o resultado final i.e. as regras de associação.

3. Discuta as questões inerentes ao tratamento de atributos numéricos na geração de regras de associação. Enumere as questões da definição do tamanho/formato dos intervalos definidos no processo de discretização destes atributos. Descreva como a proposta de *Partial Completeness* de Srikant & Agrawal tenta resolver estes problemas. Discuta em particular o problema de lidar com atributos numéricos no consequente das regras de associação.

```
C4.5 [release 8] decision tree generator        Tue Dec 18 14:51:54 2007
---------------------------------------

    Options:
        File stem <vote>


Read 300 cases (16 attributes) from vote.data



Decision Tree:

physician fee freeze = n:
|   adoption of the budget resolution = y: democrat (151.0)
|   adoption of the budget resolution = u: democrat (1.0)
|   adoption of the budget resolution = n:
|   |   education spending = n: democrat (6.0)
|   |   education spending = y: democrat (9.0)
|   |   education spending = u: republican (1.0)
physician fee freeze = y:
|   synfuels corporation cutback = n: republican (97.0/3.0)
|   synfuels corporation cutback = u: republican (4.0)
|   synfuels corporation cutback = y:
|   |   duty free exports = y: democrat (2.0)
|   |   duty free exports = u: republican (1.0)
|   |   duty free exports = n:
|   |   |   education spending = n: democrat (5.0/2.0)
|   |   |   education spending = y: republican (13.0/2.0)
|   |   |   education spending = u: democrat (1.0)
physician fee freeze = u:
|   water project cost sharing = n: democrat (0.0)
|   water project cost sharing = y: democrat (4.0)
|   water project cost sharing = u:
|   |   mx missile = n: republican (0.0)
|   |   mx missile = y: democrat (3.0/1.0)
|   |   mx missile = u: republican (2.0)
```

```
Simplified Decision Tree:


physician fee freeze = n: democrat (168.0/2.6)
physician fee freeze = y: republican (123.0/13.9)
physician fee freeze = u:
|   mx missile = n: democrat (3.0/1.1)
|   mx missile = y: democrat (4.0/2.2)
|   mx missile = u: republican (2.0/1.0)


Evaluation on training data (300 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

          25    8( 2.7%)      7   13( 4.3%)    ( 6.9%)   <<
```

Árvore de decisão para o dataset vote.data
Figura 1

FIM