CS590D Exam

March 24th, 1998

1. For the following RDBMS scenario, perform inductive logic programming. The goal is to determine the factors influencing the choice of residence (`Fairlands` vs. `West Lafayette`) (5+5+5=15 points). State

   - the hypothesis that you come up with (*i.e.*, do the data mining yourself)
   - some brief sketches on how you would obtain the same with a system like PROGOL
   - if you need ILP for this or can it be achieved by attribute–value mechanisms?

Employee Details	
Name	Salary
John	90,000
Vickie	140,000
Tod	70,000
Paula	90,000
Chad	150,000

Place of Residence	
Name	Place
John	West Lafayette
Tod	Fairlands
Chad	Fairlands

Spouse	
Name	Spouse
John	Paula
Tod	Vickie
Chad	NULL

2. Consider the following table about data obeying Kepler's third law of planetary motion which states that for any planet, the cube of its distance from the sun (D) is proportional to the square of its period (P). Given only columns 1, 2 and 3, explain how you would determine the above mentioned pattern. *i.e.,* State the model that you would use (decision tree/instance based algorithm/neural network etc.) and how you would use it. Notice that we are not looking for a way to verify the pattern (which is what the fourth column is about), but to 'mine' it in the first place (15 points).

Data for Kepler's Third Law			
Planet	Distance (D)	Period(P)	D^3/P^2
Mercury	0.382	0.241	1.00
Venus	0.724	0.616	1.00
Earth	1.0	1.0	1.00
Mars	1.524	1.881	1.00

3. • Mention five advantages of a data warehouse for data mining. (10 points)

- Construct the 3D data cube from the table below to aggregate (sum) sales over all possible dimensions. Is it easy to maintain this cube? (15+5 = 20 points) *Extra Credit: Write down the SQL statement for this expression.*

McDonald's Data			
Type	Year	Subtype	Sales
Breakfast	1994	small	3
Breakfast	1995	big	4
Regular	1994	medium	3
Regular	1994	small	9
Regular	1995	big	8

4. For the following transaction database find all rules with support > 0.35 and confidence > 0.9. What is the negative border and what is the support of each element in the negative border? (14 points)

Transaction	Items							
1	a,	b,	c,	d,		f		
2	a,		c,	d,		f		
3		b,	c,			f		
4							g,	h
5							g	
6	a,							
7				d,			g	
8			c,		e,	f,		
9						f,		h

5. For the following questions, assume the following data: Where data is not available to answer question, provide lower bounds. Frequent itemsets along with their support (minimal support = 0.6) (4x4 =16 points):

```
{a} 0.72
{b} 0.79
{c} 0.65
{a, b} 0.61
{b, c} 0.61

Negative border:
{a, b, c} 0.45
```

- What is the confidence of the rule `a OR b => c` ?

- What is the confidence of the rule `a => b OR c` ?

- Using ONLY the frequent sets, how many transactions contain `a OR b OR c`? (*Hint: use the inclusion-exclusion principle for determining approximate support*)

- What is the error in the above?

6. For the following time series, using a window of two samples ($d = 2$) and a step size of 1, predict the next value using 2 nearest neighbors and linear interpolation between them (10 points):

   `12 14.5 15 13 13.5 12.5 14 ?`