Student Number_____ Last Name_____ First name_____

3804ICT Quiz 2 – Answer Short Questions

Contents: Mining Frequent Patterns I – II & Machine Learning in Data Mining & Outlier Detection & Time Series and Sequential Data Mining

Instructions: Please list all the detailed calculation steps which help you to get the final solutions.

Question Set 1. Basic concepts (10 points)

Please answer the following questions:

a) What is a frequent pattern? (1 point)
b) Besides support and confidence, what else can be used to measure the interestingness of rules? (1 point)
c) What are the differences between supervised and unsupervised learning methods? (2 points)
d) What are the overfitting problem and the underfitting problem? (2 points)
e) Please describe the challenges of outlier detections. (2 points)
f) What are time series data mining and sequence data mining? (2 points)

Question Set 2. (20 points)

Transaction	Products
1	laptop, camera, hard-drive
2	laptop, DVD
3	DVD, speakers
4	laptop, camera, hard-drive
5	CD, hard-drive
6	DVD, hard-drive
7	CD, DVD
8	laptop, camera, TV
9	TV, speakers
10	laptop, camera

a) Given the transaction dataset above, please find out all the frequent itemsets using Apriori algorithm. (min_support = 0.25) (10 points)
b) Given the frequent k-itemsets (k>1) discovered by (a), find out all the strong association rules. (min_conf = 0.75) (10 points)

Question Set 3. (15 points)

Record ID	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

c) Given the dataset above, estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|−), P(B|−), and P(C|−). Note that P(A|+) represents P(A=1|+). (7 points)

d) Use the estimation of conditional probabilities given in the previous question to predict the class label for a test sample X=(A = 0, B = 1, C = 0) using the naive Bayes approach. (8 points)

Question 4. Sequence Data Mining (5 points)

e) List ten of the 4-subsequences contained in the following data sequence: <{1,3}{3}{2,3}{4,5}>. (2.5 points)

f) List ten of the 3-element sub-sequences contained in the data sequence: <{1,2,3}{2,3}{2,3,5}{4}>. (2.5 points)