Data Mining Exam – June 2016

Name:.....

Specialization:.....

Marking rule: 0.5 points/ question

Remark: the multiple choice questions can have 1,2 or several correct answers

1. Let us suppose that we have a log file containing data on the access of various users to a web server. We are interested in identifying some user profiles. To which category belongs this task? (a) classification; (b) forecasting; (c) clustering; (d) association rules; (e) regression.
2. Let us consider the following set of transactions:
   T1: {milk, bread, meat, water}
   T2: {bread, water}
   T3: {bread, butter, meat, water}
   T4: {water}
   and the rule: IF bread and meat THEN water. Compute the values for the rule support: ..... and for the rule confidence .....
3. Let us consider a numerical attribute taking values in the interval [a,b). We are interested in discretizing the attribute by using an equi-depth discretization approach in such a way that the discretized attribute takes values in a set with N elements (v_1,v_2,..., v_N). Which interval of values is mapped to the i-th discretized value (v_i)? (a) [a+i*(b-a)/N, a+(i+1)*(b-a)/N); (b) [a+(i-1)*(b-a)/N, a+(i+1)*(b-a)/N); (c) [a+(i-1)*(b-a)/N, a+i*(b-a)/N); (d) [b-(i+N)*(b-a)/N, b-(i+N-1)*(b-a)/N); (e) [b-(i+N-1)*(b-a)/N, b-(i+N-2)*(b-a)/N);
4. Let us consider the following confusion matrix provided by a binary classifier

	Predicted C1	Predicted C2
True C1	25	15
True C2	5	55

   Compute the accuracy of the classifier: .....

5. Let us consider the following probability distributions correponding to three splitting variants (in the context of the induction decision trees for a binary classification problem): (a) (0.5,0.5); (b) (1,0); (c) (0.25,0.75); (d) (0.75,0.25). Which of these distributions have the smallest entropy?