Distributed Information Systems: Spring Semester 2018 - Quiz 4

Student Name: ______________________________________
Date: April 26 2018
Student ID: ________________________________
Total number of questions: 8
Each question has a single answer!

____

1. In a FP tree, the leaf nodes are the ones with:
   a. Lowest confidence
   b. Lowest support
   c. Least in the alphabetical order
   d. None of the above

2. Suppose that an item in a leaf node N exists in every path. Which one is correct?
   a. N co-occurs with its prefix in every transaction.
   b. For every node p that is a parent of N in the fp tree, confidence(p->n) = 1
   c. N's minimum possible support is equal to the number of paths.
   d. The item N exists in every candidate set.

Note: (look at slide 65. E is in every path but it does occur with c in every transaction, c can be in a transaction without e)

3. Fundamentally, why clustering is considered an unsupervised machine learning technique?
   a. Number of clusters are not known.
   b. The class labels are not known.
   c. The features are not known.
   d. The clusters can be different with different initial parameters.

4. Which of the following is true for a density based cluster C:
   a. Any two points in C must be density reachable. Each point belongs to one, and only one cluster
   b. Any two points in C must be density reachable. Border points may belong to more than one cluster
   c. Any two points in C must be density connected. Border points may belong to more than one cluster
   d. Any two points in C must be density connected. Each point belongs to one, and only one cluster

5. Suppose that q is density reachable from p. The chain of points that ensure this relationship are {t,u,g,r} Which one is FALSE?
    a. {t,u,g,r} have to be all core points.
    b. p and q will also be density-connected
    c. p has to be a core point
    d. q has to be a border point

6. What is a correct pruning strategy for decision tree induction?
    a. Apply Maximum Description Length principle
    b. Stop partitioning a node when either positive or negative samples dominate the samples of the other class
    c. Choose the model that maximizes L(M) + L(M|D)
    d. Remove attributes with lowest information gain

    Solution: b (week 8 classification slide 21)
    c is incorrect: L(M) + L(D|M) (week 8 classification slide 21)
    a is incorrect: minimum description length (week 8 classification slide 22)

7. Given the distribution of positive and negative samples for attributes A1 and A2, which is the best attribute for splitting?

A1	P	N
a	7	0
b	1	4
A2	P	N
x	5	1
y	3	3

    a. A1
    b. A2
    c. They are the same
    d. There is not enough information to answer the question

You can use this website to compute the binary entropy:
http://www.wolframalpha.com/widgets/view.jsp?id=4e095a8fa96257fbf9da2529e930ccd3

Solution:
H(A1) = 7/24 * H(7,0) + 5/24 * H(1,4) = 7/24 * 0 + 5/24 * 0.811 = 0.169
H(A2) = 6/24 * H(5,1) +6/24 * H(3,3) = 6/24 * 0.65 + 6/24 * 1 = 0.4125
→ H(16,8) - H(A1) > H(16,8) - H(A2)
→ A1 is the best attribute for splitting due to better information gain

8. When using bootstrapping in Random Forests, the number of different data items used to construct a single tree is:
   a. smaller than the size of the training data set, with high probability
   b. of order square root of the size of the training set, with high probability
   c. the same as the size of the training data set
   d. subject to the outcome of the sampling process, and can be both smaller or larger than the training set