Distributed Information Systems: Spring Semester 2016

Quiz 4

`Student Name: ________________________________________` `Date: 21 Apr 2016`
`Student ID: __________________________________________` `Time: 11:15AM to 11:30AM`

Total number of questions: 8

Each question has a single answer!

1. Suppose S is a collection of 14 examples of some boolean concepts, including 9 positive and 5 negative examples. Then the entropy of S relative to this binary classification is:

   □ *a)* (9/14)log_2(9/14)-(5/14)log_2(5/14)

   □ *b)* (9/14)log_2(9/14)+(5/14)log_2(5/14)

   ☒ *c)* -(9/14)log_2(9/14)-(5/14)log_2(5/14)

   □ *d)* -(9/14)log_2(9/14)+(5/14)log_2(5/14)

2. The Swiss festivals database has records on all Swiss festivals. In particular, it has a total of 100 records about Montreux festival. Alice queried the database about Montreux festival, and 80 records were retrieved. But she found that, of the 80 records retrieved, only 60 were relevant to her query. What are the recall and precision scores of this query?

   □ *a)* recall = 75%, precision=60%

   ☒ *b)* recall = 60%, precision=75%

   □ *c)* recall = 60%, precision=80%

   □ *d)* recall = 80%, precision=75%

3. What is one **disadvantage** of the filtering technique compared to the wrapping technique in the context of feature selection?

   □ *a)* For each specific classifier, filtering is always more computationally intensive.

   □ *b)* Filtering has to be repeated again for each change of classifier type.

   ☒ *c)* Filtering assumes that features are independent, and this is not always valid.

   □ *d)* Filtering is dependent on the classifier.

4. Which of the following is **false** in the context of discretization of continuous values?

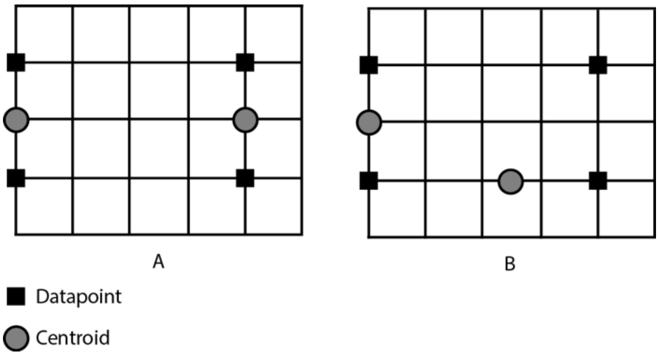
   □ *a)* Clustering allows to perform discretization on all features at once.

   □ *b)* Supervised discretization takes class information into account.

   □ *c)* The equal-frequency method ensures that values are evenly distributed in bins, independent of the values' distribution.

   ☒ *d)* The equal-width method is particularly suitable with non-uniform values' distributions.

5. We consider two initial configurations A and B as shown below. We want to apply k-means clustering on them, with k=2 and the intra-cluster distance as the objective function. Which is the following is **true**?



- ☐ *a*) With initial configuration A, k-means will converge in more steps than with configuration B.
- ☒ *b*) With initial configuration B, applying k-means will result in an optimal clustering (minimal intra-cluster distance).
- ☐ *c*) Applying k-means on initial configurations A and B will lead to different final clusters.
- ☐ *d*) None of the above

6. In the context of crowdsourcing, which of the following is **true**:

- ☒ *a*) Honey pots cannot differentiate sloppy workers from spammers.
- ☐ *b*) There is always a predominant label generated by the users.
- ☐ *c*) With honey pots, good workers can never be misidentified as spammers.
- ☐ *d*) The expectation maximisation aggregation method fully eliminates the impact of answers from spammers and sloppy workers.

7. Which of the following is a **drawback** of using decision trees as classifiers ?

- ☐ *a*) Decision trees can only fit linear models on data.
- ☒ *b*) Without pruning, decision trees tend to overfit on data.
- ☐ *c*) Decision tree models are hard to interpret.
- ☐ *d*) All of the above.

8. Which of the following is **true** about the k-means algorithm?

- ☐ *a*) It handles noisy data well.
- ☒ *b*) It might terminate in local optima.
- ☐ *c*) It is slow to converge.
- ☐ *d*) Its complexity grows quadratically in the number of objects.