Distributed Information Systems: Spring Semester 2018 - Quiz 5

Student Name: _____
Date: May 17 2018
Student ID: _____
Total number of questions: 8
Each question hopefully has a single answer!

1. Which of the following is **correct** regarding *Louvain* algorithm?
   - ☐ a. It creates a hierarchy of communities with a common root
   - ☐ b. *Clique* is the only topology of nodes where the algorithm detects the same communities, independently of the starting point
   - ☐ **c. If *n* cliques of the same order are connected cyclically with *n-1* edges, then the algorithm will always detect the same communities, independently of the starting point**
   - ☐ d. Modularity is always maximal for the communities found at the top level of the community hierarchy

2. In *User-Based Collaborative Filtering*, which of the following is **correct,** assuming that all the ratings are positive?
   - ☐ a. *Pearson Correlation Coefficient* and *Cosine Similarity* have different value range, but return the same similarity ranking for the users
   - ☐ **b. If the ratings of two users have both variance equal to 0, then their *Cosine Similarity* is maximized**
   - ☐ c. *Pearson Correlation Coefficient* and *Cosine Similarity* have the same value range, but can return different similarity ranking for the users
   - ☐ d. If the variance of the ratings of one of the users is 0, then their *Cosine Similarity* is not computable

3. Which of the following is **correct** regarding *Crowdsourcing*?
   - ☐ a. *Random Spammers* give always the same answer for every question
   - ☐ b. It is applicable only for binary classification problems
   - ☐ c. *Honey Pot* discovers all the types of spammers but not the sloppy workers
   - ☐ d. **The output of *Majority Decision* can be equal to the one of *Expectation-Maximization***

4. Which of the following is **correct** regarding prediction models?
   - ☐ a. Training error being less than test error means overfitting
   - ☐ b. Training error being less than test error means underfitting
   - ☐ **c. Complex models tend to overfit, unless we feed them with more data**
   - ☐ d. Simple models have lower bias than complex models

5. In the χ^2 statistics for a binary feature, we obtain $P(\chi^2 \mid DF = 1) > 0.05$. This means in this case, it is assumed:
   - ☐ a. That the class labels depends on the feature
   - ☐ **b. That the class label is independent of the feature**
   - ☐ c. That the class label correlates with the feature
   - ☐ d. None of the above

6. Which is an appropriate method for fighting skewed distributions of class labels in classification?
   - ☐ a. Include an over-proportional number of samples from the larger class
   - ☐ b. Use leave-one-out cross validation
   - ☐ **c. Construct the validation set such that the class label distribution approximately matches the global distribution of the class labels**
   - ☐ d. Generate artificial data points for the most frequent classes

		Class	
		Fraud	¬Fraud
Classified	Fraud	20	20
	¬Fraud	10	60

7. Considering the results of this fraud classifier, which of the following is **correct**?
   - ☐ **a. The classifier has a precision of 50% and a recall of 66.6%**
   - ☐ b. The classifier has a precision of 75% and a recall of 50%
   - ☐ c. The classifier has a precision of 50% and a recall of 75%
   - ☐ d. The classifier has a precision of 66.6% and a recall of 75%

The following question is cancelled:

8. Which of the following is **correct** regarding community detection?
   - ☐ a. High betweenness of an edge indicates that the communities are well connected by that edge
   - ☐ b. The *Louvain* algorithm attempts to minimize the overall modularity measure of a community graph
   - ☐ c. High modularity of a community indicates a large difference between the number of edges of the community and the number of edges of a null model
   - ☐ d. The *Girvan-Newman* algorithm attempts to maximize the overall betweenness measure of a community graph

Comment: a) Is not correct, yes an edge with high betweenness connects two communities and these two communities will be distinct and won't be connected by many edges. But these communities are not "well-connected" and also answer does not specificy which communities are that edge connects. (Though I must admit this sentence is not well-formed.) b) and d) are not correct.

c) was the intended answer. Suppose that we take one of the community detected by Louvain algorithm. Then randomize the edges of the whole graph, which we call null model, and take the nodes which belong to the community we found. The community extracted from the original graph would have more edges between the nodes that belong to that community, then the community extracted from the null model. However in c, we accidentally meant the whole graph by the null model, which has the same number of edges as the original graph, which would have more edges than a community.