

Distributed Information Systems: Spring Semester 2017 - Quiz 4 - Answers Marked

Date: May 4 2017. Total number of questions: 12. Single answer!

1. Which of the following statements is **true** for clustering and classification?

- ☐ a. Clustering is only applicable for 2 dimensions
- ☐ b. **Similar items belong to the same cluster**
- ☐ c. To do clustering, we need to know the labels of data
- ☐ d. Classification is an unsupervised machine learning technique

Solution: b. Slide 2,3,4 week8-classification

2. Which of the following statements is **true** for training set and test set?

- ☐ a. The training set must always be larger than the test set
- ☐ b. The test set can only contain feature combinations that occur also in the training set
- ☐ c. **The goal of classification is to maximize the accuracy on the test set**
- ☐ d. The goal of classification model is to maximize the accuracy on the training set, regardless of the test set

Solution: c. Slide 5,8 week8-classification

3. Which is one of the stopping conditions of partitioning in the decision tree induction algorithm?

- ☐ a. Information gain of all attributes is equal
- ☐ b. There is only one attribute left
- ☐ c. The height of the tree is equal to the number of data objects
- ☐ d. **All data objects are in the same class**

Solution: d. Slide 13 week8-classification

4. Which is **true** about entropy?

- ☐ a. Entropy is maximal when it is zero
- ☐ b. We split on the attribute with highest entropy
- ☐ c. The domain value of entropy is $[0, \infty]$
- ☐ d. **The domain value of entropy is $[0,1]$**

Solution: d. Slide 14 week8-classification

5. Which is a correct pruning strategy for decision tree induction?

- ☐ a. Apply Maximum Description Length principle
- ☐ b. Stop partitioning a node when the number of positive and negative samples are equal
- ☐ c. **Build the full tree, then replace subtrees with leaf nodes labelled with the majority class, if classification accuracy does not change**
- ☐ d. Remove attributes with lowest information gain

Solution: c. Slide 22 week8-classification

6. Which is an advantage of using the random forest algorithm?

- ☐ a. **Can be parallelized**
- ☐ b. Uses only a small sample of training data for learning
- ☐ c. Performs always better than deep neural networks
- ☐ d. Produces a human interpretable model

Solution: a. Slide 42 week8-classification

7. Which is **true** for social graph community detection?

- ☐ a. Louvain algorithm is efficient for small networks, while Girvan-Newman is efficient for large networks
- ☐ b. We need to specify the number of clusters in hierarchical clustering
- ☐ c. Louvain algorithm runs in quadratic time, which is better than Girvan-Newman algorithm
- ☐ d. **Edge betweenness is smaller than or equal to the total number of paths passing over the edge**

Solution: a. Slide 25 week8- mining social graphs

8. Which is **true** about crowdsourcing? **(not graded)**

- ☐ a. **Uniform spammers give uniformly random answers**
- ☐ b. Crowd-workers only give yes/no answers
- ☐ c. Honey Pot does not remove sloppy workers, only spammers
- ☐ d. The accuracy of majority voting is never equal to EM

Solution: b. Slide 15 week9-classification pipeline

9. Which is an appropriate method for fighting skewed distributions of class labels in classification? **(not graded)**

- ☐ a. Include an over-proportional number of samples from the larger class
- ☐ b. Use leave-one-out cross validation
- ☐ c. **Construct the validation set with a class label distribution similar to the global distribution of the class labels**
- ☐ d. Generate artificial data points for the most frequent classes

Solution: c. Slide 65, week9-classification pipeline

10. Which is **true** about errors? **(not graded)**

- ☐ a. Training error being less than test error means overfitting
- ☐ b. Training error being greater than test error means underfitting
- ☐ c. Complex models always have smaller test error than simple models
- ☐ d. **Complex models generally have smaller training error than simple models**

Solution: d. Slide 69, week9-classification pipeline

11. If for the χ^2 statistics for a binary feature we obtain $P(\chi^2 \mid DF = 1) > 0.05$ this means

- ☐ 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

Solution: b. $p\text{-value} > 0.05 \rightarrow$ accept null hypothesis \rightarrow independence

12. Which of the following tasks would typically not be solved by clustering

- ☐ a. Community detection in social networks
- ☐ b. Discretization of continuous features
- ☐ c. **Spam detection in an email system**
- ☐ d. Detection of latent topics in a document collection

Solution: c. classify an email as spam or not spam