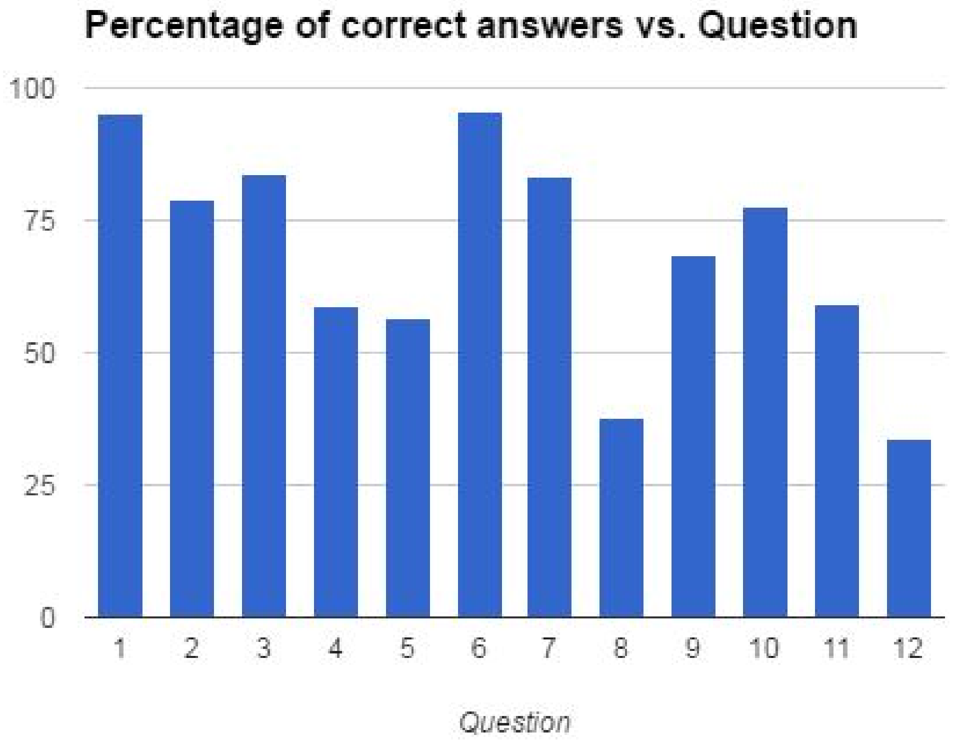
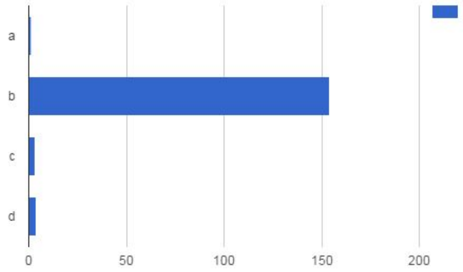
Quiz 4

Solutions


Percentage of correct answers vs. Question
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
Question

Question 1

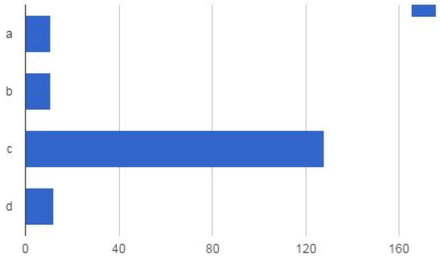


Which of the following statements is **true** for clustering and classification?

- □ 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

week8 classification

Question 2

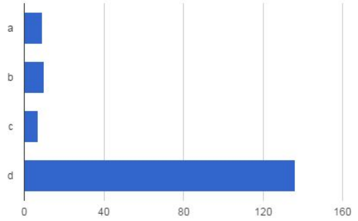


Which of the following statements is true for training set and test set?

- □ 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

a is incorrect: typically larger (week 8 classification slide 5)
b is incorrect: feature combinations of test set depend on the split (week 8 classifiction slide 5)
c is correct: slide 8
d is incorrect: a model that remembers all labels of training set → training error = 0 but poor test error

Question 3

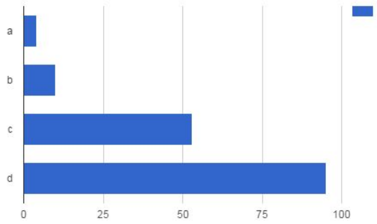


Which is one of the stopping conditions of partitioning in the decision tree induction algorithm?

- □ 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**

week 8 classification slide 13

Question 4



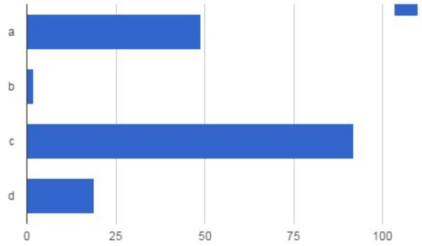
Which is true about entropy?

- □ 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]**

week 8 classification slide 14

The question is about entropy, not information gain

Question 5

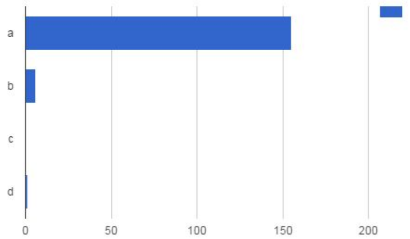


Which is a correct pruning strategy for decision tree induction?

- □ 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

week 8 classification slide 22
a is incorrect: minimum description length

Question 6

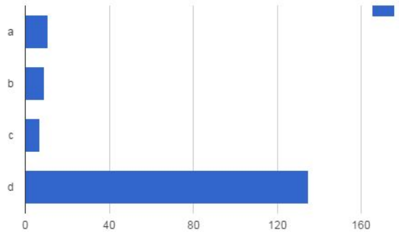


Which is an advantage of using the random forest algorithm?

- □ **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

week 8 classification slide 36

Question 7

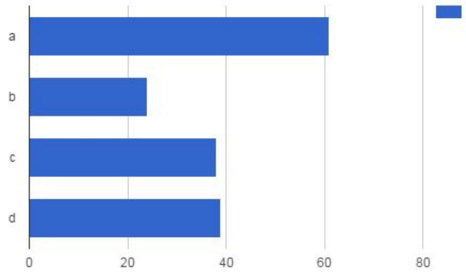


Which is true for social graph community detection?

- □ 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**

week 8 social graph slide 25
number of shortest paths <= total number of paths

Question 8 (not graded)

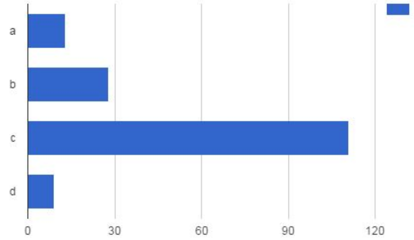


Which is true about crowdsourcing?

- □ 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

b is incorrect: crowdsourcing can be used for multi-label problem (e.g. multiple choice question)
c is incorrect: week 9 slide 20
d is incorrect: if all workers are experts and give all correct answers

Question 9 (not graded)

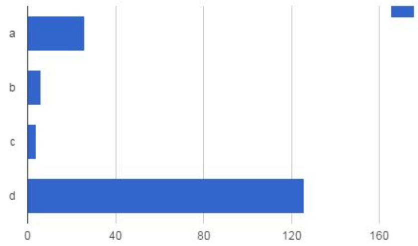


Which is an appropriate method for fighting skewed distributions of class labels in classification?

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

week 9 slide 65

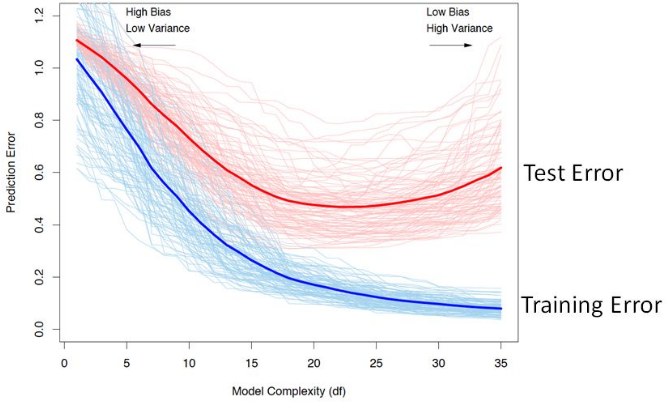
Question 10 (not graded)



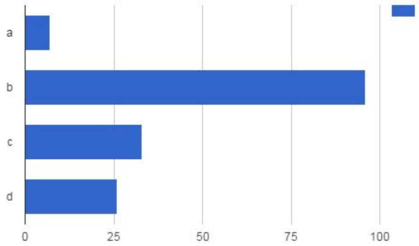
Which is true about errors?

- □ 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**

week 9 slide 69



Question 11

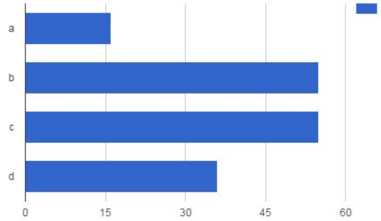


If for the χ2 statistics for a binary feature we obtain P(χ2 | DF = 1) > 0.05 this means

□ 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

week 9 slide 28
the null hypothesis of chi-square test is the independence
p-value > 0.05 → accept null hypothesis → independence

Question 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

spam detection is a classification problem: classify an email as spam (label 1) or not spam (label 0)
a can be solved by clustering: e.g. k-mean
b can be solved by clustering: week 9 slide 25, unsupervised discretization
d can be solved by clustering: documents in the same topic are often similar → use clustering