Quiz week 1

Help

The **due date** for this quiz is **Wed 26 Nov 2014 1:30 PM IST**.

☐ **In accordance with the Coursera Honor Code, I (Md Asif Khaleel) certify that the answers here are my own work.**

Question 1

The four V's of big data are Volume, Velocity, Variety and Veracity. Which of these four V's is applicable when we talk about the immense amount of data currently generated?

- ○ Volume
- ○ Variety
- ○ Velocity
- ○ Veracity

Question 2

When we talk about replay, we mean the process where…

- ○ we start from a process model and generate behavior, e.g. traces.
- ○ we start from both a process model and a collection of observed behavior, e.g. traces, and compare these.
- ○ we start from event data and generate a process model, e.g. a Petri net.

Question 3

We would like to learn the influence of someone's weight and drinking behavior on their smoking behavior. What are the response and predictor variables?

drinker	smoker	weight
yes	yes	120
no	no	70
yes	no	72
yes	yes	55
no	yes	94
no	no	62
...	...	...

- Variables drinker and smoker are the response variables and weight is the predictor variable.
- Variable weight is the response variable and drinking and smoker are the predictor variables.
- Variables drinker and weight are the response variables and smoker is the predictor variable.
- Variable smoker is the response variable and drinker and weight are the predictor variables.
- Variables smoker and weight are the response variables and drinking is the predictor variable.
- Variable drinker is the response variable and weight and smoker are the predictor variables.

Question 4

There are two types of learning: supervised and unsupervised. Which of the following statements are true for **unsupervised** learning?

- [ ] An example is to cluster similar data together.
- [ ] The data is labeled such that for each element its class is known
- [ ] The goal is to explain a response variable in terms of the predictor variables.
- [ ] An example is the detection of patterns in the data.
- [ ] An example is classification of data, e.g. learning a decision tree.

Question 5

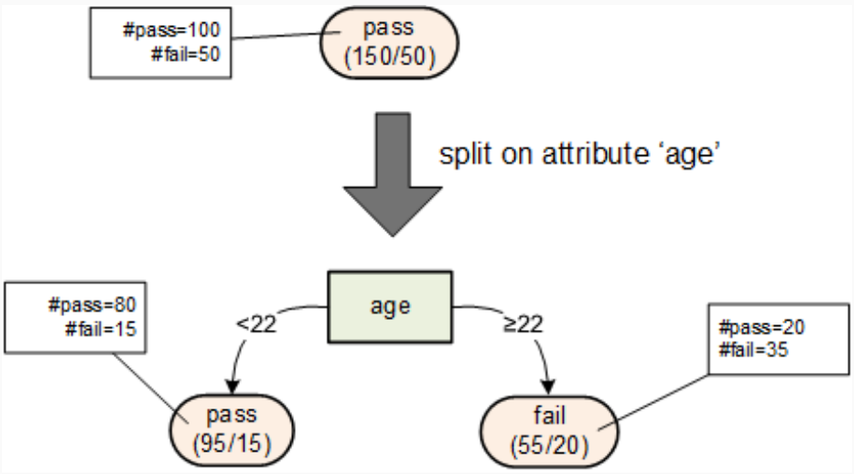
Consider a node in a decision tree with 100 instances of type A and 50 of type B. What is the entropy of this node?

- 0.0817 = $1 - (\frac{100}{150} log_2(\frac{100}{150}) + \frac{50}{150} log_2(\frac{50}{150}))$
- 0.5310 = $1 - (\frac{15}{150} log_2(\frac{15}{150}) + \frac{135}{150} log_2(\frac{135}{150}))$
- 0.4690 = $-(\frac{15}{150} log_2(\frac{15}{150}) + \frac{135}{150} log_2(\frac{135}{150}))$

- 0.6258 = $-\frac{1}{2} \times (\frac{50}{150} log_2(\frac{100}{150}) + \frac{100}{150} log_2(\frac{50}{150}))$
- 0.9183 = $-(\frac{100}{150} log_2(\frac{100}{150}) + \frac{50}{150} log_2(\frac{50}{150}))$
- 0.63500 = $1 - (\frac{25}{150} log_2(\frac{25}{150}) + \frac{125}{150} log_2(\frac{125}{150}))$
- 0.6500 = $-(\frac{25}{150} log_2(\frac{25}{150}) + \frac{125}{150} log_2(\frac{125}{150}))$

Question 6

Consider the two decision trees depicted below (a tree with just one node, and a tree where this node is split based on the age attribute). Does it make sense to split the tree?



- Yes, since the entropy of the entire tree goes from 0.9183 to 0.7453.
- No, since the entropy of the entire tree goes from 0.9183 to 0.7453.
- Yes, since the entropy of the entire tree goes from 0.9183 to 1.1716.
- Yes, since the entropy of the entire tree goes from 0.9183 to 1.7453.

Question 7

What is the formula to calculate the **support** that X implies Y given that

N is the number of instances

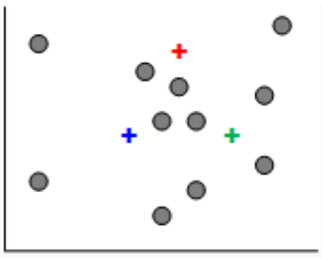
N_X is the number of instances covering X

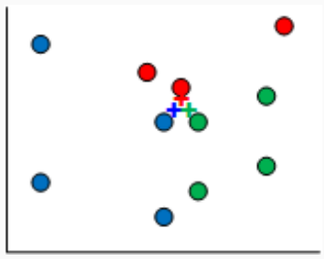
$N_{X \wedge Y} = N_{X \cup Y}$ is the number of instances covering both X and Y

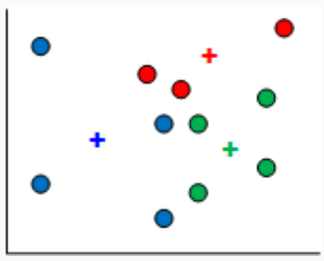
- Support($X \Rightarrow Y$) = $\frac{N_{X \wedge Y}}{N}$ = $\frac{N_{X \cup Y}}{N}$
- Support($X \Rightarrow Y$) = $\frac{N_{X \wedge Y}}{N_X}$ = $\frac{N_{X \cup Y}}{N_X}$
- Support($X \Rightarrow Y$) = $\frac{N_{X \wedge Y}/N}{(N_X/N)(N_Y/N)}$ = $\frac{N_{X \wedge Y} N}{N_X N_Y}$

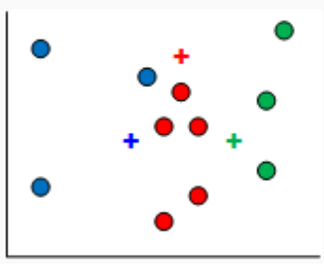
Question 8

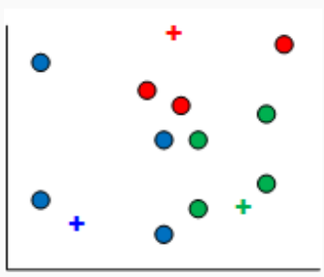
Assume a data set with two variables that we would like to cluster using k-means with k=3. See the following centroids. Which one could be the end results of applying k-means.





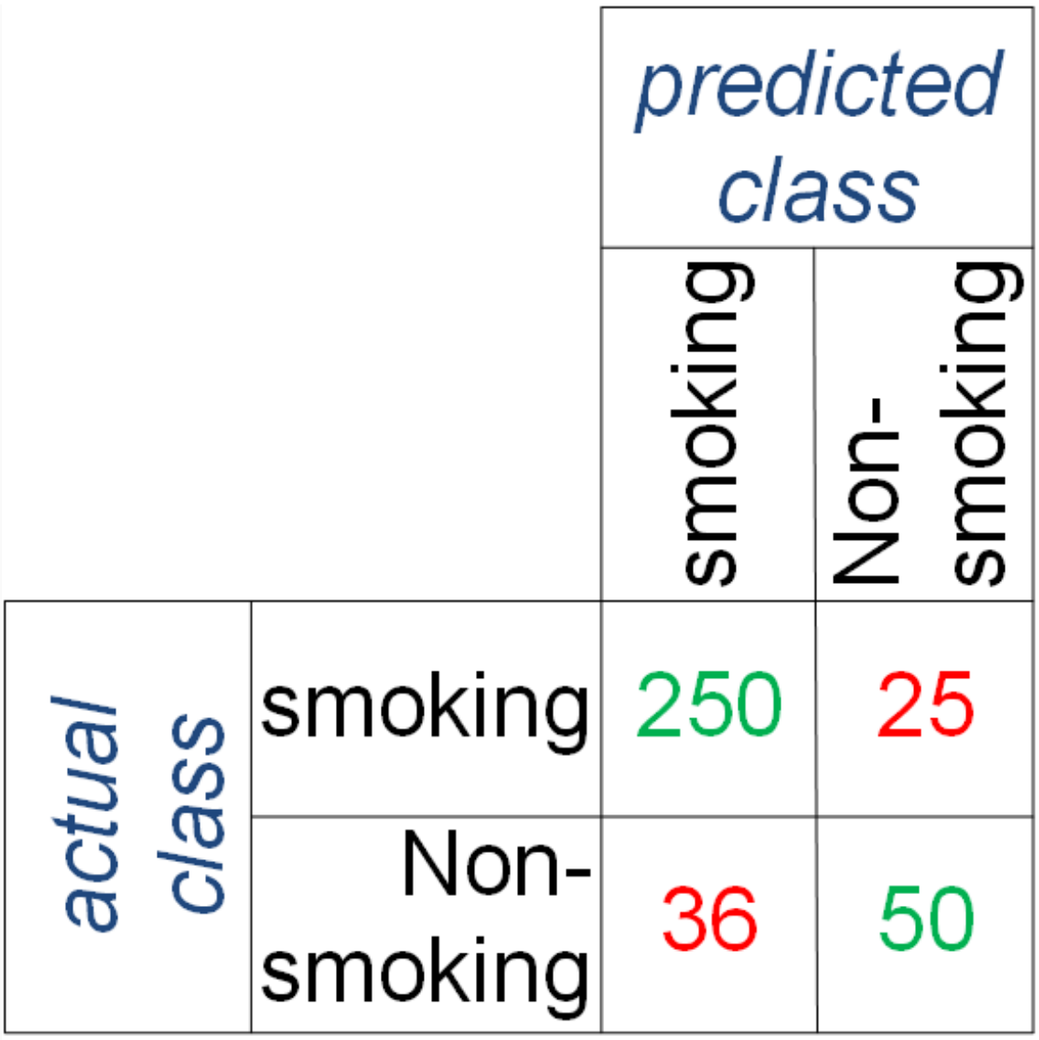






Question 9

Given the classification provided below, what is the corresponding precision?



- 0.1690
- 0.9091
- 0.8741
- 0.8310

Question 10

Please check the statements that are true for k-fold cross validation.

- [ ] The quality of the model learned by the algorithm is evaluated on one data set not used for learning the model.
- [ ] The learning algorithm can only use k-1 data sets during each of the runs.
- [ ] The data set is split into k smaller data sets.
- [ ] The learning algorithm is applied k-1 times on different combinations of training and test data sets.

- [ ] **In accordance with the Coursera Honor Code, I (Md Asif Khaleel) certify that the answers here are my own work.**

Submit Answers Save Answers

You cannot submit your work until you agree to the Honor Code. Thanks!