

Feedback — Quiz week 1

[Help Center](#)

Thank you. Your submission for this quiz was received.

You submitted this quiz on **Wed 14 Oct 2015 6:47 AM CEST**. You got a score of **4.78** out of **5.00**.

Please note that quiz questions might change between attempts!

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 problem that you cannot be sure that the data is fully accurate?

Your Answer	Score	Explanation
<input type="radio"/> Variety		
<input type="radio"/> Velocity		
<input checked="" type="radio"/> Veracity	✓ 0.50	Veracity refers to the fact that you cannot be fully sure that the data is fully accurate.
<input type="radio"/> Volume		
Total	0.50 / 0.50	

Question Explanation

The meaning of each of the four V's is explained in the first lecture of week 1: 'Data science and big data'.

Briefly, they are as follows:

1. *Volume* refers to the incredible amount of data we currently generate.
2. *Velocity* refers to data continuously being added and things change very rapidly.
3. *Variety* refers to the fact that there often is not one type of data but many types (text, image, audit trails) exist and are often combined.
4. *Veracity* refers to the fact that you cannot be completely sure that the data is fully accurate.

Question 2

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

Your Answer	Score	Explanation
<input checked="" type="radio"/> we start from both a process model and a collection of observed behavior, e.g. traces, and compare these.	✓ 0.50	Replay is when we start from both a process model and a collection of observed behavior, e.g. traces, and compare these.
<input type="radio"/> we start from a process model and generate behavior, e.g. traces.		
<input type="radio"/> we start from event data and generate a process model, e.g. a Petri net.		
Total	0.50 / 0.50	

Question Explanation

The notions of play in, play out and replay are discussed in Week 1 lecture 'Different types of process mining'.


Briefly, they are as follows:

- *Play-Out* is when we start from a process model and generate behavior, e.g. traces.
- *Play-In* is when we start from event data and generate a process model, e.g. a Petri net.
- *Replay* is when we start from both a process model and a collection of observed behavior, e.g. traces, and compare these by replaying the traces on the process model.

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
...

Your Answer	Score	Explanation
<input type="radio"/> Variables drinker and smoker are the response variables and weight is the predictor variable.		
<input checked="" type="radio"/> Variable smoker is the response variable and drinker and weight are the predictor variables.	 0.50	We would like to predict someone's smoking behavior, so 'smoker' is the response variable. The other two variables are predictor variables.
<input type="radio"/> Variables drinker and weight are the response variables and smoker is the predictor variable.		
<input type="radio"/> Variables smoker and weight are the response variables and drinking is the predictor variable.		
<input type="radio"/> Variable weight is the response variable and drinking and smoker are the predictor variables.		
<input type="radio"/> Variable drinker is the response variable and weight and smoker are the predictor variables.		
Total	0.50 / 0.50	

Question Explanation

We would like to predict someone's smoking behavior, so smoking is the response variable. The

other two variables are predictor variables. The notions of predictor and response variables, and other aspects of predicting, are explained in more detail in the lecture 'How process mining relates to data mining' in week 1.

Question 4

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

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> The goal is to explain a response variable in terms of the predictor variables.	✓ 0.10	The labels which variables are the response or predictor variables is also provided as input.
<input checked="" type="checkbox"/> An example is classification of data, e.g. learning a decision tree.	✓ 0.10	Using the response and predictor variables data can be classified.
<input type="checkbox"/> The data is labeled such that for each element its class is known	✗ 0.00	The labels contain additional information added to the data,
<input type="checkbox"/> An example is the detection of patterns in the data.	✓ 0.10	This is not the goal of supervised learning but of unsupervised learning.
<input type="checkbox"/> The goal is to cluster similar data together.	✓ 0.10	This is not the goal of supervised learning but of unsupervised learning.
Total	0.40 / 0.50	

Question Explanation

In supervised learning additional information is added to the data by labeling the data. Furthermore, it is indicated what variables are (possibly) depending on others (the response variables depend on the predictor variables). An example of a supervised learning technique is learning a decision tree.

The differences between supervised and unsupervised learning are explained in the lecture 'How Process Mining Relates to Data Mining' in week 1.

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?

Your Answer	Score	Explanation
-------------	-------	-------------

☐ 0.63500 =

$$1 - \left(\frac{25}{150} \log_2 \left(\frac{25}{150} \right) + \frac{125}{150} \log_2 \left(\frac{125}{150} \right) \right)$$

☐ 0.6500 =

$$- \left(\frac{25}{150} \log_2 \left(\frac{25}{150} \right) + \frac{125}{150} \log_2 \left(\frac{125}{150} \right) \right)$$

☐ 0.5310 =

$$1 - \left(\frac{15}{150} \log_2 \left(\frac{15}{150} \right) + \frac{135}{150} \log_2 \left(\frac{135}{150} \right) \right)$$

☐ 0.4690 =

$$- \left(\frac{15}{150} \log_2 \left(\frac{15}{150} \right) + \frac{135}{150} \log_2 \left(\frac{135}{150} \right) \right)$$

☒ 0.9183 =

$$- \left(\frac{100}{150} \log_2 \left(\frac{100}{150} \right) + \frac{50}{150} \log_2 \left(\frac{50}{150} \right) \right)$$



0.50

This is indeed the correct formula and the correct assignment of values to variables.

☐ 0.6258 =

$$- \frac{1}{2} \times \left(\frac{50}{150} \log_2 \left(\frac{100}{150} \right) + \frac{100}{150} \log_2 \left(\frac{50}{150} \right) \right)$$

☐ 0.0817 =

$$1 - \left(\frac{100}{150} \log_2 \left(\frac{100}{150} \right) + \frac{50}{150} \log_2 \left(\frac{50}{150} \right) \right)$$

Total

0.50 /

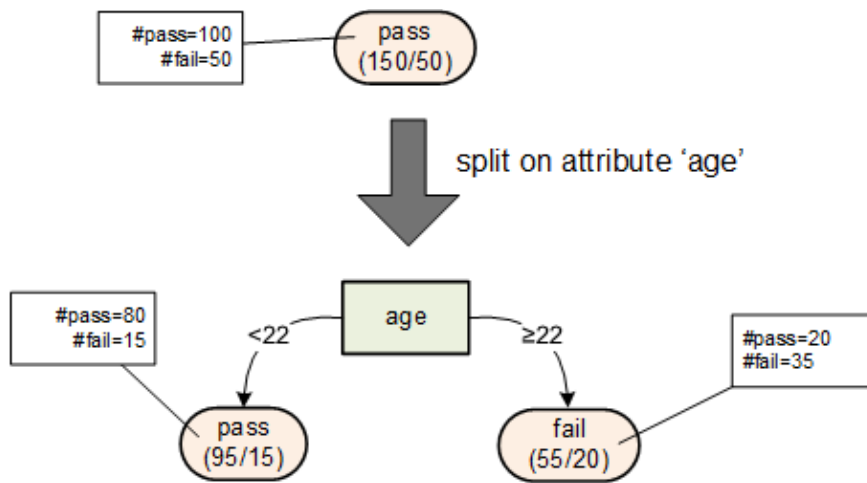
0.50

Question Explanation

In the lecture 'Learning Decision Trees' of week 1 the notion of entropy is explained. Entropy is calculated by filling in the following formula: $E = - \sum_{i=1}^k p_i \log_2(p_i)$ where k are all possible values (in our case 2: 'A' and 'B'), $p_i = \frac{c_i}{n}$ is the fraction of elements having value i with $c_i \geq 1$ is the number of i values and $n = \sum_{i=1}^k c_i$. When we fill in this formula for our example we obtain $k = 2$ such that $E = - \sum_{i=1}^k p_i \log_2(p_i) = -(p_A \log_2(p_A) + p_B \log_2(p_B)) = -(\frac{c_A}{n} \log_2(\frac{c_A}{n}) + \frac{c_B}{n} \log_2(\frac{c_B}{n})) = -(\frac{100}{150} \log_2(\frac{100}{150}) + \frac{50}{150} \log_2(\frac{50}{150})) = -(\frac{2}{3} \log_2(\frac{2}{3}) + \frac{1}{3} \log_2(\frac{1}{3})) = -(\frac{2}{3} \times -0.58496250072 + \frac{1}{3} \times -1.58496250072) = -(-0.38997500048 - 0.528320833573) = 0.91829583405 = 0.9183$.

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?



Your Answer	Score	Explanation
<input checked="" type="radio"/> Yes, since the entropy of the entire tree goes from 0.9183 to 0.7453.	<input checked="" type="checkbox"/> 0.50	The entropy of the single-node tree is 0.9183. That of the splitted tree is $\frac{95 \times 0.62925 + 55 \times 0.94566}{150} = 0.74527$ which is an increase in the information gain.
<input type="radio"/> No, since the entropy of the entire tree goes from 0.9183 to 0.7453.		
<input type="radio"/> Yes, since the entropy of the entire tree goes from 0.9183 to 1.1716.		
<input type="radio"/> Yes, since the entropy of the entire tree goes from 0.9183 to 1.7453.		
Total	0.50 / 0.50	

Question Explanation

The entropy of the single-node tree is 0.9183. That of the splitted tree is That of the splitted tree is $\frac{95 \times 0.62925 + 55 \times 0.94566}{150} = 0.74527$ which is an increase in the information gain hence splitting is a good idea. More information about entropy, information gain and constructing a decision tree can be found in lectures 'Learning Decision Trees' and 'Applying Decision Trees' of week 1.

Question 7

What is the formula to calculate the **confidence** 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

Your Answer	Score	Explanation
<input type="radio"/> $\text{Confidence}(X \Rightarrow Y) = \frac{N_{X \wedge Y}}{N}$ $= \frac{N_{X \cup Y}}{N}$		
<input checked="" type="radio"/> $\text{Confidence}(X \Rightarrow Y) = \frac{N_{X \wedge Y}}{N_X}$ $= \frac{N_{X \cup Y}}{N_X}$	0.50	This formula indeed expresses confidence, which estimates the fraction of times Y holds when X holds.
<input type="radio"/> $\text{Confidence}(X \Rightarrow Y) = \frac{N_{X \wedge Y}/N}{(N_X/N)(N_Y/N)} = \frac{N_{X \wedge Y}N}{N_X N_Y}$		
Total	0.50 / 0.50	

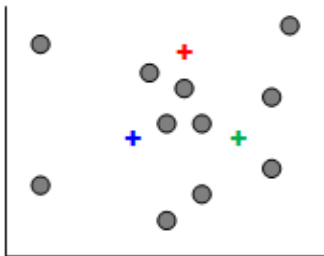
Question Explanation

Confidence expresses the number of times X and Y both hold given that X holds, and is therefore expressed as $\text{Confidence}(X \Rightarrow Y) = \frac{N_{X \wedge Y}}{N_X} = \frac{N_{X \cup Y}}{N_X}$.

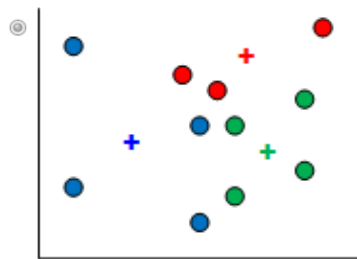
More information regarding support, confidence and lift is provided in the lecture "Association Rule Learning" in week 1.

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.

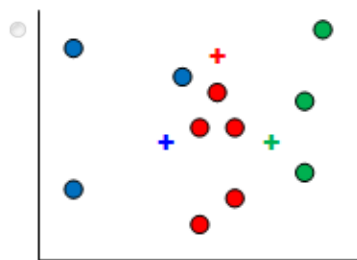
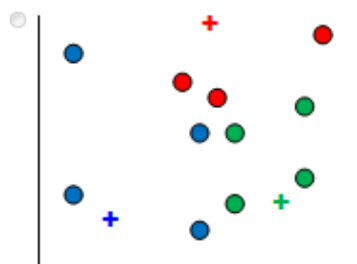
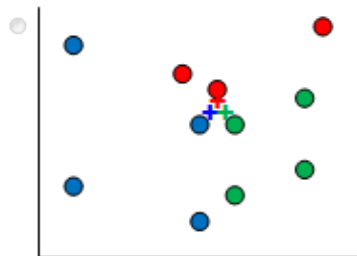


Your Answer	Score	Explanation
-------------	-------	-------------



✓ 0.50

This is indeed the correct answer.

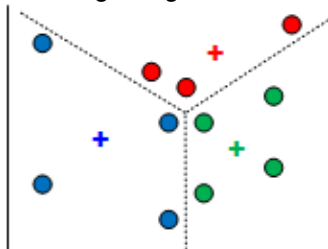


Total

0.50 / 0.50

Question Explanation

The k-means algorithm starts by assigning elements to the closest centroid. The centroids are then moved to the center of the elements assigned to their cluster. In our example the centroids move only slightly and no elements are moved from one cluster to another. This results in the following assignment and centroid positions after one iteration:



If you would like to learn more about k-means clustering, please refer to the 'cluster analysis' lecture in week 1.

Question 9

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

		<i>predicted class</i>	
		smoking	Non-smoking
<i>actual class</i>	smoking	250	25
	Non-smoking	36	50

Your Answer

Score

Explanation

☐ 0.8310

☐ 0.8741

☐ 0.9091

☒ 0.1690



0.50

This is indeed the corresponding error value.

Total

0.50 / 0.50

Question Explanation

The four measures of error, accuracy, precision and recall are explained in the lecture 'Evaluating mining results' in week 1. Using the information provided there and the classification above we can derive TP=250, TN=50, FN=25, FP=36 and K = 361.

Using these values we can fill in the following formulas:

$$\text{Error} = \frac{FP+FN}{K} = \frac{36+25}{361} = 0.1690$$

$$\text{Accuracy} = \frac{TP+TN}{K} = \frac{250+50}{361} = 0.8310$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{250}{250+36} = 0.8741$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{250}{250+25} = 0.9091$$

Question 10

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

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> The data set is split into k smaller data sets.	✓ 0.12	
<input checked="" type="checkbox"/> Within a run, the quality of the model learned by the algorithm is evaluated on the one data set not used for learning the model.	✓ 0.12	
<input type="checkbox"/> The learning algorithm can only use k-1 data sets during each of the runs to learn the model from.	✗ 0.00	
<input type="checkbox"/> The learning algorithm is applied k-1 times on different combinations of training and test data sets.	✓ 0.12	This is not correct since the learning algorithm is applied k times on the data sets, where each data set is exactly once used as training set for evaluation.
Total	0.38 / 0.50	

Question Explanation

K-fold cross validation is explained in more detail in the 'evaluating mining results' lecture in week 1.

