Free Questions for Databricks-Certified-Professional-Data-Scientist

Shared by Nelson on 06-06-2022

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

As a data scientist consultant at ABC Corp, you are working on a recommendation engine for the learning resources for end user. So Which recommender system technique benefits most from additional user preference data?

Options:

A- Naive Bayes classifier

B- Item-based collaborative filtering

C- Logistic Regression

D- Content-based filtering

Item-based scales with the number of items, and user-based scales with the number of users you have. If you have something like a store, you'll have a few thousand items at the most. The biggest stores at the time of writing have around 100,000 items. In the Netflix competition, there were 480,000 users and 17,700 movies. If you have a lot of users: then you'll probably want to go with item-based similarity. For most product-driven recommendation engines, the number of users outnumbers the number of items. There are more people buying items than unique items for sale. Item-based collaborative filtering makes predictions based on users preferences for items. More preference data should be beneficial to this type of algorithm. Content-based filtering recommender systems use information about items or users, and not user preferences, to make recommendations. Logistic Regression, Power iteration and a Naive Bayes classifier are not recommender system techniques.

Answer:

B

Question 2

Question Type: MultipleChoice

Which of the following problem you can solve using binomial distribution

Options:

A- A manufacturer of metal pistons finds that on the average: 12% of his pistons are rejected because they are either oversize or undersize. What is the probability that a batch of 10 pistons will contain no more than 2 rejects?

B- A life insurance salesman sells on the average 3 life insurance policies per week. Use Poisson's law to calculate the probability that in a given week he will sell Some policies

C- Vehicles pass through a junction on a busy road at an average rate of 300 per hour Find the probability that none passes in a given minute.

D- It was found that the mean length of 100 parts produced by a lathe was 20.05 mm with a standard deviation of 0.02 mm. Find the probability that a part selected at random would have a length between 20.03 mm and 20.08 mm

The entire problem can be solved using below method

Binomial: A manufacturer of metal pistons finds that on the average, 12% of his pistons are rejected because they are either oversize or undersize. What is the probability that a batch of 10 pistons will contain no more than 2 rejects?

Poisson: A life insurance salesman sells on the average 3 life insurance policies per week. Use Poisson's law to calculate the probability that in a given week he will sell Some policies

Poisson: Vehicles pass through a junction on a busy road at an average rate of 300 per hour Find the probability that none passes in a given minute.

Normal: It was found that the mean length of 100 parts produced by a lathe was 20.05 mm with a standard deviation of 0.02 mm. Find the probability that a part selected at random would have a length between 20 03 mm and 20.08 mm

Answer:

A

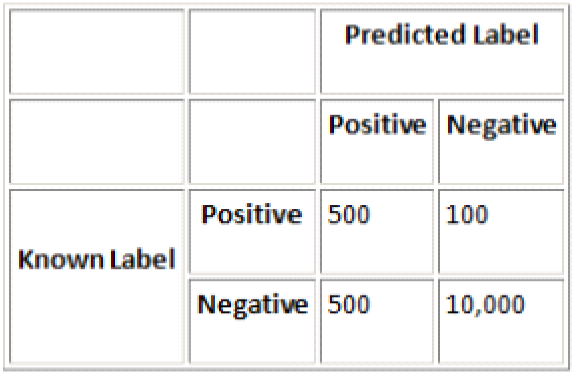
Question 3

Question Type: MultipleChoice

Consider the following confusion matrix for a data set with 600 out of 11,100 instances positive:

In this case, Precision = 50%, Recall = 83%, Specificity = 95%, and Accuracy = 95%.

Select the correct statement



		Predicted Label	
		Positive	Negative
Known Label	Positive	500	100
	Negative	500	10,000

Options:

A- Precision is low, which means the classifier is predicting positives best

B- Precision is low, which means the classifier is predicting positives poorly

C- problem domain has a major impact on the measures that should be used to evaluate a classifier within it

D- 1 and 3

E- 2 and 3

In this case, Precision = 50%, Recall = 83%, Specificity = 95%: and Accuracy = 95%. In this case, Precision is low, which means the classifier is predicting positives poorly. However, the three other measures seem to suggest that this is a good classifier. This just goes to show that the problem domain has a major impact on the measures that should be used to evaluate a classifier within it, and that looking at the 4 simple cases presented is not sufficient.

Answer:

E

Question 4

Question Type: MultipleChoice

Reducing the data from many features to a small number so that we can properly visualize it in two or three dimensions. It is done in_______

Options:

A- supervised learning

B- un-supervised learning

C- k-Nearest Neighbors

D- Support vector machines

The opposite of supervised learning is a set of tasks known as unsupervised learning. In unsupervised learning, there's no label or target value given for the data. A task where we group similar items together is known as clustering. In unsupervised learning, we may also want to find statistical values that describe the data. This is known as density estimation. Another task of unsupervised learning may be reducing the data from many features to a small number so that we can properly visualize it in two or three dimensions

Answer:

B

Question 5

Question Type: MultipleChoice

Suppose there are three events then which formula must always be equal to P(E1|E2,E3)?

Options:

A- P(E1,E2,E3)P(E1)/P(E2:E3)

B- P(E1,E2;E3)/P(E2,E3)

C- P(E1,E2|E3)P(E2|E3)P(E3)

D- P(E1,E2|E3)P(E3)

E- P(E1,E2,E3)P(E2)P(E3)

This is an application of conditional probability: P(E1,E2)=P(E1|E2)P(E2). so

P(E1|E2) = P(E1.E2)/P(E2)

P(E1,E2,E3)/P(E2,E3)

If the events are A and B respectively, this is said to be 'the probability of A given B'

It is commonly denoted by P(A|B): or sometimes PB(A). In case that both 'A' and 'B' are categorical variables, conditional probability table is typically used to represent the conditional probability.

Answer:

B

Question 6

Question Type: MultipleChoice

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has

rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. Which of the following will you use to calculate the probability whether it will rain on the

day of Marie's wedding?

Options:

A- Naive Bayes

B- Logistic Regression

C- Random Decision Forests

D- All of the above

The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. You should consider Bayes' theorem when the following conditions exist.

* The sample space is partitioned into a set of mutually exclusive events {A1, A2,... :An}.

* Within the sample space, there exists an event B: for which P(B) > 0.

* The analytical goal is to compute a conditional probability of the form: P(Ak B).

Answer:

A

Question 7

Question Type: MultipleChoice

Select the correct statement which applies to logistic regression

Options:

A- Computationally inexpensive, easy to implement knowledge representation easy to interpret

B- May have low accuracy

C- Works with Numeric values

D- Only 1 and 3 are correct

E- All 1, 2 and 3 are correct

Depending on the size of the data you are uploading, Amazon S3 offers the following options:

Logistic regression

Pros: Computationally inexpensive, easy to implement knowledge representation easy to interpret Cons: Prone to underfitting, may have low accuracy Works with: Numeric values^ nominal values

Answer:

E

Question 8

Question Type: MultipleChoice

The figure below shows a plot of the data of a data matrix M that is 1000 x 2. Which line represents the first principal component?

Options:

A- yellow

B- blue

C- Neither

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component

accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

The first principal component corresponds to the greatest variance in the data. The blue line is evidently this first principal component, because if we project the data onto the blue line, the data is more spread out (higher variance) than if projected

onto any other line, including the yellow one.

Answer:

B