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

Shared by Daugherty on 12-12-2023

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

Projecting a multi-dimensional dataset onto which vector has the greatest variance?

Options:

A- first principal component

B- first eigenvector

C- not enough information given to answer

D- second eigenvector

E- second principal component

The method based on principal component analysis (PCA) evaluates the features according to the projection of the largest eigenvector of the correlation matrix on the initial dimensions, the method based on Fisher's linear discriminant analysis evaluates. Them according to the magnitude of the components of the discriminant vector.

The first principal component corresponds to the greatest variance in the data, by definition. If we project the data onto the first principal component line, the data is more spread out (higher variance) than if projected onto any other line, including other principal components.

Answer:

A

Question 2

Question Type: MultipleChoice

Google Adwords studies the number of men, and women, clicking the advertisement on search

engine during the midnight for an hour each day.

Google find that the number of men that click can be modeled as a random variable with distribution

Poisson(X), and likewise the number of women that click as Poisson(Y).

What is likely to be the best model of the total number of advertisement clicks during the midnight for an hour ?

Options:

A- Binomial(X+Y,X+Y)

B- Poisson(X/Y)

C- Normal(X+Y(M+Y)1/2)

D- Poisson(X+Y)

The total number of clicks is the sum of the number of men and
women. The sum of two Poisson random variables also follows a Poisson distribution with

rate equal to the sum of their rates.
The Normal and Binomial distribution can approximate the Poisson distribution in certain cases, but the expressions above do not approximate Poisson(X+Y).

Answer:

D

Question 3

Question Type: MultipleChoice

Of all the smokers in a particular district, 40% prefer brand A and 60% prefer brand B. Of those smokers who prefer brand

Options:

A- 30% are females, and of those who prefer brand B. 40% are female. What is the probability that a randomly selected smoker prefers brand A, given that the person selected is a female?
Which of the following is a best way to solve this problem?

A- Bays Theorem

B- Poisson Distribution

C- Binomial Distribution

D- None of the above

Answer:

A, A

Question 4

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 5

Question Type: **MultipleChoice**

A problem statement is given as below

Hospital records show that of patients suffering from a certain disease, 75% die of it. What is the probability that of 6 randomly selected patients, 4 will recover?

Which of the following model will you use to solve it.

Options:

A- Binomial

B- Poisson

C- Normal

D- Any of the above

Answer:

A

Question 6

Question Type: **MultipleChoice**

A website is opened 3 times by a user. What is the probability of he clicks 2 times the advertisement, is best calculated by

Options:

A- Binomial

B- Poisson

C- Normal

D- Any of the above

In a binomial distribution, only 2 parameters, namely n and p, are needed to determine the probability. Where p is the probability of success and q is the probability of failure in a binomial trial, then the expected number of successes in n trials.
This is a binomial distribution because there are only 2 possible outcomes (we get a 5 or we don't).

Answer:

A

Question 7

Question Type: **MultipleChoice**

Question-18. What is the best way to ensure that the k-means algorithm will find a good clustering of a collection of vectors?

Options:

A- Only consider values of k larger than log(N), where N is the number of observations in the data set

B- Run at least log(N) iterations of Lloyd's algorithm, where N is the number of observations in the data set

C- Choose the initial centroids so that they all He along different axes

D- Choose the initial centroids so that they are far away from each other

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining, k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes

This Question-is about the properties that make k-means an effective clustering heuristic which primarily deal with ensuring that the initial centers are far away from each other. This is how modern k-means algorithms like k-means++ guarantee that with high probability Lloyd's algorithm will find a clustering within a constant factor of the optimal possible clustering for each k.

Answer:

D

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

Question 9

Question Type: MultipleChoice

What are the advantages of the mutual information over the Pearson correlation for text classification problems?

Options:

A- The mutual information has a meaningful test for statistical significance.

B- The mutual information can signal non-linear relationships between the dependent and independent variables.

C- The mutual information is easier to parallelize.

D- The mutual information doesn't assume that the variables are normally distributed.

A linear scaling of the input variables (that may be caused by a change of units for the measurements) is sufficient to modify the PCA results. Feature selection methods that are sufficient for simple distributions of the patterns belonging to different classes can fail in classification tasks with complex decision boundaries. In addition, methods based on a linear dependence (like the correlation) cannot take care of arbitrary relations between the pattern coordinates and the different classes. On the contrary, the mutual information can measure arbitrary relations between variables and it does not depend on transformations acting on the different variables.

This item concerns itself with feature selection for a text classification problem and references mutual information criteria. Mutual information is a bit more sophisticated than just selecting based on the simple correlation of two numbers because it can detect non-linear relationships that will not be identified by the correlation. Whenever possible: mutual information is a better feature selection technique than correlation.

Mutual information is a quantification of the dependency between random variables. It is sometimes contrasted with linear correlation since mutual information captures nonlinear dependence.

Correlation analysis provides a quantitative means of measuring the strength of a linear relationship between two vectors of data. Mutual information is essentially the measure of how much 'knowledge' one can gain of a certain variable by knowing the value of another

variable.

Answer:

C