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

Shared by Kidd on 29-01-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

You have used k-means clustering to classify behavior of 100, 000 customers for a retail store. You decide to use household income, age, gender and yearly purchase amount as measures. You have chosen to use 8 clusters and notice that 2 clusters only have 3 customers assigned. What should you do?

Options:

A- Decrease the number of measures used

B- Increase the number of clusters

C- Decrease the number of clusters

D- Identify additional measures to add to the analysis

kmeans uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. You can control the details of the minimization using several optional input parameters to kmeans, including ones for the initial values of the cluster centroids, and for the maximum number of iterations.

Clustering is primarily an exploratory technique to discover hidden structures of the data: possibly as a prelude to more focused analysis or decision processes. Some specific applications of k-means are image processing^ medical and customer segmentation. Clustering is often used as a lead-in to classification. Once the clusters are identified,

labels can be applied to each cluster to classify each group based on its characteristics. Marketing and sales groups use k-means to better identify customers who have similar behaviors and spending patterns.

Answer:

C

Question 2

Question Type: MultipleChoice

Which analytical method is considered unsupervised?

$y1, y2, y3, \ldots, yn\text{-}1, yn$

may have a trend component that is quadratic in nature. Which pattern of data will indicate that the trend in the time series data is quadratic in nature?

Options:

A- Naive Bayesian classifier

B- Decision tree

C- Linear regression

D- K-means clustering

kmeans uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. You can control the details of the minimization using several optional input parameters to kmeans, including ones for the initial values of the cluster centroids, and for the maximum number of iterations.

Clustering is primarily an exploratory technique to discover hidden structures of the data, possibly as a prelude to more focused analysis or decision processes. Some specific applications of k-means are image processing, medical and customer segmentation. Clustering is often used as a lead-in to classification. Once the clusters are identified,

labels can be applied to each cluster to classify each group based on its characteristics. Marketing and sales groups use k-means to better identify customers who have similar behaviors and spending patterns.

Answer:

D

Question 3

Question Type: MultipleChoice

The method based on principal component analysis (PCA) evaluates the features according to

Options:

A- The projection of the largest eigenvector of the correlation matrix on the initial dimensions

B- According to the magnitude of the components of the discriminate vector

C- The projection of the smallest eigenvector of the correlation matrix on the initial dimensions

D- None of the above

Feature Selection:

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 discriminate analysis evaluates. Them according to the magnitude of the components of the discriminate vector.

Answer:

A

Question 4

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.

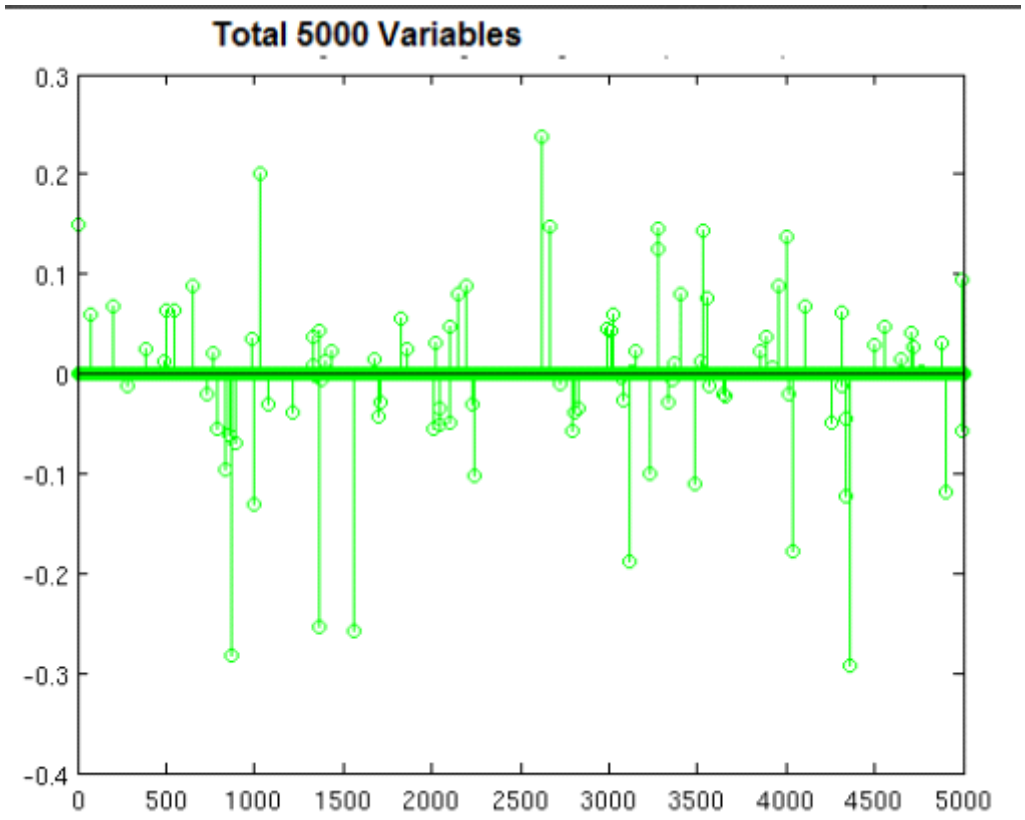
Answer:

E

Question 5

Question Type: MultipleChoice

You are building a classifier off of a very high-dimensiona data set similar to shown in the image with 5000 variables (lots of columns, not that many rows). It can handle both dense and sparse input. Which technique is most suitable, and why?



Options:

A- Logistic regression with L1 regularization, to prevent overfitting

B- Naive Bayes, because Bayesian methods act as regularlizers

C- k-nearest neighbors, because it uses local neighborhoods to classify examples

D- Random forest because it is an ensemble method

Logistic regression is widely used in machine learning for classification problems. It is well-known that regularization is required to avoid over-fitting, especially when there is a only small number of training examples, or when there are a large number of parameters to be learned. In particular L1 regularized logistic regression is often used for feature selection, and has been shown to have good generalization performance in the presence of many irrelevant features. (Ng 2004; Goodman 2004) Unregularized logistic regression is an unconstrained convex optimization problem with a continuously differentiate objective function. As a consequence, it can be solved fairly efficiently with standard convex optimization methods, such as Newton's method or conjugate gradient. However, adding the L1 regularization makes the optimization

problem com-putationally more expensive to solve. If the L1 regulariza-tion is enforced by an L1 norm constraint on the parameLogistic regression is a classifier and L1 regularization tends to produce models that ignore dimensions of the input that are not predictive. This is particularly useful when the input contains many dimensions, k-nearest neighbors classification is also a classification technique, but relies on notions of distance. In a high-dimensional space, most every data point is 'far' from others (the curse of dimensionality) and so these techniques break down. Naive Bayes is not inherently regularizing. Random forests represent an ensemble method; but an ensemble method is not necessarily more suitable to high-dimensional data. Practically, I think the biggest reasons for regularization are 1) to avoid overfitting by not generating high coefficients for predictors that are sparse. 2) to stabilize the estimates especially when there's collinearity in the data.

1) is inherent in the regularization framework. Since there are two forces pulling each other in the objective function, if there's no meaningful loss reduction, the increased penalty from the regularization term wouldn't improve the overall objective function. This is a great property since a lot of noise would be automatically filtered out from the model. To give you an example for 2), if you have two predictors that have same values, if you just run a regression algorithm on it since the data matrix is singular your beta coefficients will be Inf if you try to do a straight matrix inversion. But if you add a very small regularization lambda to it, you will get stable beta coefficients with the coefficient values evenly divided between the equivalent two variables. For the difference between L1 and L2, the following graph demonstrates why people bother to have L1 since L2 has such an elegant analytical solution and is so computationally straightforward. Regularized regression can also be represented as a constrained regression problem (since they are Lagrangian equivalent). The implication of this is that the L1 regularization gives you sparse estimates. Namely, in a high dimensional space, you got mostly zeros and a small number of non-zero coefficients. This is huge since it incorporates variable selection to the modeling problem.

In addition, if you have to score a large sample with your model, you can have a lot of computational savings since you don't have to compute features(predictors) whose coefficient is 0. I personally think L1 regularization is one of the most beautiful things in machine learning and convex optimization. It is indeed widely used in bioinformatics and large scale machine learning for companies like Facebook, Yahoo, Google and Microsoft.

Answer:

A

Question 6

Question Type: MultipleChoice

Which of the following is not a correct application for the Classification?

Options:

A- credit scoring

B- tumor detection

C- image recognition

D- drug discovery

Classification : Build models to classify data into different categories credit scoring, tumor detection, image recognition Regression: Build models to predict continuous data, electricity load forecasting, algorithmic trading, drug discovery

Answer:

D

Question 7

Question Type: MultipleChoice

Question-34. Stories appear in the front page of Digg as they are "voted up" (rated positively) by the community. As the community becomes larger and more diverse, the promoted stories can better reflect the average interest of the community members. Which of the following technique is used to make such recommendation engine?

Options:

A- Naive Bayes classifier

B- Collaborative filtering

C- Logistic Regression

D- Content-based filtering

One scenario of collaborative filtering application is to recommend interesting or popular information as judged by the community. As a typical example, stories appear in the front page of Digg as they are 'voted up' (rated positively) by the community. As the community becomes larger and more diverse, the promoted stories can better reflect the average interest of the community members.

Answer:

B