TEXATA WORLD CHAMPIONSHIP
WORLD'S BIG DATA SHOWDOWN

TEXATA 2014 – Round 1

Case Study: Big Data Theory MCQs

Question Appeared in: TEXATA 2014 Round 1

Time Allocated: 60 minutes

Question 1

A salesman offers you a choice of three boxes, one containing a million dollars and two containing fifty dollars and tells you to pick one. He then shows you fifty dollars in one of the other two boxes and asks you if you want to change your choice to the remaining box that you have neither picked nor seen inside. What do you do?

a. Change to the other box

b. Stay with the one you picked originally

c. It doesn't matter, so do nothing

d. You don't have enough information to figure out whether you should change, so do nothing

Question 2

Which of the following is one of the reasons for not using a logistic activation function in a neural network?

a. Computation speed

b. Interpretability of the results

c. Presence of outliers in sample data

d. None of the above

Question 3

You are given a dataset of 200 patients with 4,000 variables including an indicator of whether or not they had developed cancer in the past year. Which of the following steps might you reasonably want to take next in modeling the incidence of cancer in your dataset?

a. Remove any variables you think won't matter so that the model can function effectively

b. Run logistic regression on the data and report the results

c. Look for a commonly-occurring observation and report it as being a good example of someone with cancer

d. None of the above

Question 4

A Gibbs sampler is useful for:

a. Situations when the likelihood function is smooth and convex

b. Situations when you don't really know the distribution of the parameters in your model

c. Situations when your model is experiencing convergence issues

d. Situations when you don't have enough data for a modeling task and want to generate more

Question 5

One of the problems with a Support Vector Machine approach to modeling is:

a. It runs in quadratic time

b. It runs in exponential time

c. It automatically changes the scale of your data

d. It isn't as good as logistic regression at using a large number of variables

Question 6

You run an experiment to see whether there is a difference in length of visit between customers viewing your old website and your new website. One way of confirming this is:

a. Calculating the difference in mean length of visit

b. Performing a t-test on the mean length of visit for the two groups

c. Performing linear regression on the observations of length of visit, with group membership as the dependent variable

d. Performing a chi-squared test on the mean length of visit for the two groups

Question 7

Linear Regression and the Nearest Neighbours model are different in that:

a. Linear Regression is used for continuous variables and Nearest Neighbours is not

b. Linear Regression is more likely to suffer from high bias and Nearest Neighbours from high variance

c. Linear Regression is more likely to suffer from high variance and Nearest Neighbours from high bias

d. Linear Regression is not as good at handling a large number of variables as Nearest Neighbours is

Question 8

The difference between L1 and L2 regularization is:

a. L1 regularization is useful for performing variable selection prior to running a model, whereas L2 regularization is not

b. L1 regularization cannot be combined with other forms of model tuning, whereas L2 can

c. L1 regularization uses the absolute value of the model parameters, while L2 regularization uses the squared distance of the parameter vector

d. L1 regularization uses the squared distance of the parameter vector, while L2 regularization uses the absolute value of the model parameters

Question 9

One of the drawbacks of a large recurrent neural network (RNN) is:

a. It can't handle very long time series

b. It takes a long time to create predictions on new data

c. The initial choice of weights needs to be made very carefully

d. It doesn't provide a serious increase in performance over a standard feed-forward network

Question 10

You are attempting to model the price of cars at auction and you find that your model has trained well, but subsequently does poorly on new data. This could be because:

a. Human behaviour isn't a good subject for modeling

b. Your model is overfitting

c. Your model is underfitting

d. None of the above

Question 11

The Dirichlet distribution:

a. Is a variant of the logistic distribution that underpins modeling counts instead of binary incidence (i.e. 1/0 data)

b. Is useful for discovering hidden structure in high-dimensional data such as free text

c. Is used for modeling high-frequency data such as stock prices

d. None of the above

Question 12

The marketing department of your company is looking for a way to call customers who are likely to churn and persuade them to stop. When you are building a model to do this, one of the best ways to assess its usefulness is:

a. The recall score, as this gives the best indication of the strike rate per call

b. The precision score, as this gives the best indication of the whether the model is successfully identifying customers about to churn

c. The precision score, as this gives the best indication of the strike rate per call

d. None of the above

Question 13

Which is one of the advantages of a standard feed-forward neural network over a Hidden Markov Model (HMM)?

a. The neural network does not require as many hidden units or states to retain a similar amount of information as a HMM

b. The neural network does not require knowledge of the output variable, whereas the HMM does

c. The HMM is not as easily interpreted as the neural network

d. All of the above

Question 14

Boosting is:

a. Designed for situations where you want a model to overfit

b. Designed for situations in which data with spurious variables are being used to fit a model

c. Designed for situations in which a classification model with a very irregular decision boundary is being fit

d. Designed for situations in which you know part of your data has been misclassified

Question 15

The degrees of freedom problem:

a. Is important in the context of modeling because it relates the amount of candidate solutions the model has to choose between

b. Is important in the context of modeling because it relates how accurate results can be

c. Is not important in the context of modeling as long as the data has been scaled

d. None of the above

Question 16

You are thinking of constructing an ensemble model but are not sure about the best way to go about it. Which parts of the process are amenable to parallelism?

a. The train stage

b. The test stage

c. Neither

d. Both

Question 17

You have been set a modeling task and it seems that the required level of accuracy is not being achieved. What might be the logical next step?

a. Start over because you have probably missed something

b. Use a more sophisticated technique for modeling

c. Explain to your stakeholders that the problem you're working with probably isn't perfectly suited for modeling

d. Assess the data and results you do have and look for regularities that might indicate why the model is not performing well

Question 18

Which of the following is a potential pitfall you could encounter when analyzing social media data?

a. The 'silent majority' problem

b. Problems with establishing user sentiment

c. Self-selection of users

d. All of the above

Question 19

Which of the following is not a visualization tool

a. Flare

b. D3.js

c. Sqoop

d. None of the above

Question 20

True or False: Stochastic Gradient Descent is an ideal candidate for optimizing a model in parallel

a. True, it runs very quickly and you can distribute your data between workers easily

b. True, it normally doesn't run quickly but does when run in parallel

c. False, running Stochastic Gradient Descent in parallel does not offer any speed gain

d. False, it is not possible to perform Stochastic Gradient Descent in parallel

Question 21

Stochastic Gradient Descent and Online Learning are similar in that:

a. Stochastic Gradient Descent is a special form of Online Learning

b. Online Learning is a special case of Stochastic Gradient Descent

c. Both are well-suited to execution in parallel

d. None of the above

Question 22

An n-gram is:

a. A way of storing large amounts of text data

b. A collection of words that frequently appear in text or speech data

c. A contiguous group of words extracted from text or speech data

d. None of the above

Question 23

The Naive Bayes method is called “Naive” because:

a. It makes a very strong assumption about the relationship between dependent variables

b. Bayes Rule is generally quite a simplistic approach to complex modeling tasks

c. It doesn't provide very good results

d. All of the above

Question 24

In time-series data, a process is said to be explosive if:

a. It grows exponentially

b. The variance increases over time

c. All of the unit roots are positive

d. All of the above

Question 25

The difference between likelihood and the posterior odds:

a. Is approximately the same as the difference between joint and conditional probability

b. Comes down to the technique you use to estimate a model

c. Is nothing, they're actually the same

d. Isn't important in the context of modeling because you generally compute both

Question 26

You have run a job in parallel but are finding that the speed gain from doing so is not very large. This may be because:

a. You are performing a simple operation repeatedly that requires a large amount of data to complete each time

b. You are performing a complicated operation repeatedly that doesn't require much data to complete each time

c. You are performing an operation repeatedly that does not require knowledge of the results from previous or subsequent iterations, making parallelism meaningless

d. None of the above

Question 27

You have taken charge of a model that has been run previously on a small scale every day but now needs to be run on millions of customers daily and given to analysts who will proactively contact customers. Which of the following tasks do you prioritize?

a. Altering the model to decrease the training error

b. Altering the model to make the results more easily interpreted by the analysts

c. Altering the model to ensure that it can run in the same amount of time

d. None of the above

Question 28

Your CEO is worried that your company is not very responsive to social media and wants a data science-driven way of fixing this. Which of these options do you NOT suggest?

a. Counsel caution because people's opinions on social media are often different to real life

b. Suggest that it should be fairly doable because there are now a number of modeling and visualization options available for this kind of problem

c. Counsel caution because modeling doesn't work very well in social media

d. Suggest it's possible to achieve things on social media, but not without a concrete goal

Question 29

Which of the following principles is a useful guide for visualizing data?

a. Try to use bright colours because they will help your viewer engage better with the data

b. Try to present your data in a way that guides the viewer to what you think is the answer

c. Try to avoid giving different views of the same data because it will confuse the viewer

d. Try to present your data in an unstructured way so that people can draw their own conclusions

Question 30

You work in a company that is only just starting to come to grips with the challenges of data science. Which is NOT a good way to enable this process?

a. Point out that the world's biggest companies are all embracing the challenge

b. Explain that the company needs to move on from outdated techniques to stay competitive

c. Try to explain that it's mostly about making better use of existing data

d. Explain that it's very difficult to misinterpret findings, so there isn't much risk involved

Question 31

You are given two models produced by your company and told to implement one of them. How do you decide between them?

a. Use the one that uses the most sophisticated model – this will more likely perform better

b. Use the one that is simplest – this will be implemented quicker

c. Look at their training accuracy – the model that does well in training has done a better job of learning about the data

d. None of the above

Question 32

Which of the following is not a common method of representing a network?

a. Adjacency list

b. Edgelist

c. Node graph

d. Adjacency matrix

Question 33

Which of the following would not represent raw data?

a. Weather station readings

b. Summarised demographics of website visitors

c. Web server logs

d. Single band satellite imagery

Question 34

You are analyzing a cohort of visitors to two websites. You know that the cohort in website A consists of 576 male visitors and 768 female visitors. Assuming all visitors also visit website B, what is the probability that a randomly selected visitor is female?

a. 3/4

b. 4/3

c. 3/7

d. 4/7

Question 35

Your customer demographic data stores whether a person is female, male or other; aged 12-25, 26-35 or 35+; and whose maximum education level is high school, undergraduate or postgraduate. Assuming a person can only have one value for each property and individuals are uniformly distributed across all combinations of categories, what is the probability that a randomly selected customer is Female, aged 35+ and has a maximum education level of undergraduate?

a. 1/27

b. 1/9

c. 1/6

d. 1/7

Question 36

Assuming a fair coin, what are the chances of flipping a heads, tails, tails and heads, in that order?

a. 1/2

b. 1/4

c. 1/8

d. 1/16

Question 37

Assuming a prior probability of 0.3 that a given marketing communication will result in a website visit, how many marketing emails were likely opened if you received 60 website visits?

a. 20

b. 200

c. 18

d. 180

Question 38

In social network analysis, the measure of node centrality which is calculated by the number of links each node has is:

a. Eigenvector centrality

b. Degree centrality

c. Betweenness centrality

d. Closeness centrality

Question 39

In social network analysis, what is the definition of a clique?

a. A set of nodes which are all linked to each other

b. The set of nodes which have the highest centrality

c. The largest connected component of the graph

d. A connected component which is not a member of the giant component

Question 40

In social network analysis, the measure of node centrality which incorporates the centrality of each node in the context of its neighbours is:

a. Eigenvector centrality

b. Degree centrality

c. Betweenness centrality

d. Closeness centrality

Question 41

Why might a study of government unemployment rates produce unreliable data?

a. Because of the social processes involved necessary to produce these data

b. There might be figures that are not reported by the government

c. Not everybody reports that they are unemployed

d. All of the above

Question 42

An O(n2) algorithm always takes longer to run than an O(log n) algorithm.

a. True

b. False

Question 43

If you double the size of a hash table, to be efficient, you must change the hash function.

a. True

b. False

Question 44

Running the merge sort algorithm on an already-sorted array takes O(n) time.

a. True

b. False

Question 45

Inserting an element into a binary search tree of size n and height h takes:

a. O(n) time

b. O(logn) time

c. O(h) time

d. O(h**2) time

Question 46

The insertion sort algorithm takes:

a. O(1) time

b. O(log n) time

c. O(n) time

d. O(n2) time**

Question 47

Breadth-first search always obtains the shortest path to each vertex (i.e. the path using the minimum number of edges).

a. True

b. False

Question 48

Which of the following statements about binary search trees is false?

a. They always have multiple links per node

b. They can be sorted efficiently

c. They always have the same shape for a particular set of data

d. They are nonlinear

Question 49

Which of the following is true of concurrent programs?

a. Data structure invariants have to hold any time the lock protecting the data structure is not held

b. No special considerations when accessing two interdependent data structures

c. Memory is stable unless explicitly updated

d. Deadlock can't happen

Question 50

Which of the following is not a potential cause of a race condition?

a. Two threads accessing the same global variable

b. Two threads have processed a request but have not communicated their success

c. Network packets are dropped between coordinating nodes

d. Memory conditions are not accessible from more than one thread

Answers

1	A
2	A
3	D
4	C
5	A
6	B
7	B
8	C
9	C
10	B
11	B
12	C
13	A
14	C
15	A
16	D
17	D
18	D
19	C
20	D
21	B
22	C
23	A
24	B
25	A
26	A
27	C
28	C
29	D
30	D
31	D
32	C
33	B
34	D
35	A
36	D
37	B
38	B
39	A
40	A
41	D
42	B
43	A
44	B
45	C
46	D
47	A
48	C
49	A
50	D