**Page 1**

Spark: With the original dataset split into 80% for training and 20% for test, how many of the from the test set were correctly classified?  
A. 19  
B. 10  
C. 1

**Page 3**

What is the purpose of exploring data?  
A. To gain a better understanding of your data.  
B. To gather your data into one repository.  
C. To digitize your data.  
D. To generate labels for your data.

**Page 5**

Which method is used to avoid overfitting in decision trees?  
A. Post-pruning  
B. None of these  
C. Pre-pruning  
D. Pre-pruning and post-pruning

**Page 10**

Suppose we have an email network where the edges of a graph represent the number of emails from one user to another. If I was going to ask if Maria had sent any emails that (either directly or through forwarding from others) reached Julio, I would ask if:  
A. Julio's node was reachable from Maria node  
B. Maria's node was reachable from Julio's node

**Page 15**

All of the following are true about data visualization EXCEPT  
A. Is more important than summary statistics for data exploration  
B. Should be used with summary statistics for data exploration.  
C. Is useful for communicating results.  
D. Provides an intuitive way to look at data.

**Page 20**

Which categories of machine learning techniques are supervised?  
A. classification and regression  
B. regression and association analysis  
C. classification and cluster analysis  
D. cluster analysis and association analysis

**Page 24**

What is the workflow for working with big data?  
A. Extrapolation -> Understanding -> Reproducing  
B. Big Data -> Better Models -> Higher Precision  
C. Theory -> Models -> Precise Advice  
D. Data collection -> Models -> Precise Advice

**Page 28**

Which of the following is an example of big data utilized in action today?  
A. The Internet  
B. Wi-Fi Networks  
C. Social Media  
D. Individual, Unconnected Hospital Databases

**Page 30**

The main steps in the k-means clustering algorithm are:  
A. Assign each sample to the sest centroid, then calculate the new centroid.  
B. Calculate the centroids, then determine the appropriate stopping criterion depending on the number of centroids.  
C. Calculate the distances between the cluster centroids, then find the two sest centroids.  
D. Count the number of samples, then determine the initial centroids.

**Page 39**

What is the minimum no. of variables/ features required to perform clustering?  
A. 0  
B. 1  
C. 2  
D. 3

**Page 59**

What is the most populated county in the state of Texas?  
A. Harris  
B. Dallas  
C. Travis  
D. Bexar

**Page 71**

Which Cypher command launches a Neo4j database search?  
A. MATCH  
B. RETURN  
C. CREATE  
D. None of the others

**Page 72**

What does the 1 represent in the following line of code?  
ssc = StreamingContext(sc, 1)  
A. To create only one partition to manage the stream.  
B. To specific debug output.  
C. To create one single context.  
D. A batch interval of 1 second.

**Page 73**

Rule confidence is used to  
A. Identify frequent item sets  
B. Determine the rule with the most items  
C. Measure the intuitiveness of a rule  
D. Prune rules by eliminating rules with low confidence

**Page 74**

The desired characteristics of a BDMS include (select all that apply):  
A. Narrow range of query sizes  
B. Continuous data ingestion  
C. Support for common "Big Data" data types  
D. Support for ACID  
E. A full query language  
F. A flexible semi-structured data model

**Page 75**

Which of the following are reasons mentioned for why data generated by people are hard to process? (select all that apply):  
A. They cannot be modeled and stored.  
B. The velocity of the data is very high.  
C. Very unstructured data.  
D. Skilled people to analyze the data are hard to come by.

**Page 76**

Which of the Vs BEST describes the result in constant increasing in the number of edges in a graph, sometimes causing challenges in knowing when one has found "an answer" to one's analysis question?  
A. Variety  
B. Volume  
C. Velocity  
D. Valence

**Page 77**

What are examples of preference constraints in the Google Maps application? (choose 2)  
A. Avoid roads under construction  
B. Avoid highways  
C. Include son's school

**Page 78**

What is a way to enable fault tolerance?  
A. Distributed Computing  
B. System Wide Restart  
C. Data Parallel Job Restart  
D. Better LAN Connection

**Page 79**

What does it mean to have a \_id:0 within our query statement?  
A. Grab the first object in the results.  
B. Grab as many objects as possible.  
C. Does not have an effect, simple convention left for compatibility issues.  
D. Tell MongoDB not to return a document id.

**Page 80**

Which of the Vs causes increased interconnectivity of a graph -- which can cause problems in analysis due to density?  
A. Velocity  
B. Variety  
C. Volume  
D. Valence

**Page 81**

Which of the following is a Cypher command used to combine two or more query results?  
A. union  
B. combine  
C. merge  
D. return

**Page 82**

Spark: With the original dataset split into 80% for training and 20% for test, how many of the first 20 samples from the test set were correctly classified?  
A. 19  
B. 10  
C. 1

**Page 83**

What is the difference between low level interfaces and high level interfaces?  
A. Low level deals with storage and scheduling while high level deals with interactivity.  
B. Low level deals with interactivity while high level deals with storage and scheduling.

**Page 84**

KNIME: In configuring the Numeric Binner node, what would happen if the definition for the humidity\_low bin is changed from ]-infinity... 25.0 [ to ]-infinity... 25.0 ] (i.e., the last bracket is changed from [ to ]?  
A. The definition for the humidity\_low bin would change from excluding 25.0 to including 25.0  
B. The definition for the humidity\_low bin would change from having 25.0 as the endpoint to having 25.1 as the endpoint  
C. Nothing would change

**Page 85**

What are the three computing steps of a semi-join?  
A. Project, Ship, Reduce  
B. Project, Decompose, Send  
C. Index, Join, Display  
D. Query, Join, Display

**Page 86**

What is data-parallelism?  
A. Having multiple multiple data pipelines at the same time.  
B. Simultaneously processing input data from multiple cores.  
C. Running the same function simultaneously for the partitions of a data set on multiple cores.  
D. At each step of the data pipeline, process values simultaneously by using multiple cores.

**Page 87**

What is schema-on-read?  
A. The process where formatted data is given structure when read.  
B. Another name for data lakes.  
C. Data is stored as raw data until it is read by an application where the application assigns structure.  
D. The process where data is pre-formatted prior to being read but the schema is loaded on read.

**Page 88**

What are the challenges to data with a high valence?  
A. Complex Data Exploration Algorithms  
B. Difficult to Integrate  
C. Reliability of Data

**Page 89**

In unsupervised approaches,  
A. the target is unlabeled.  
B. the target is unknown or unavailable.  
C. the target is provided.  
D. the target is what is being predicted.

**Page 90**

Trails (paths without repeated edges) can be interesting in which of the following problem applications?  
A. Routing to avoid visiting the same city.  
B. An email network tracing frequency of emails from one person to another.  
C. An email network tracing email replies.  
D. Routing to avoid using the same bridge or road.

**Page 91**

Consider the following generic statement:  
db.<collection>.find(<query filter>, <projection>).<cursor modifier>  
Which part of the statement would reflect that of the SELECT statement in SQL as illustrated in the lecture?  
A. <query filter>  
B. <projection>  
C. <cursor modifier>  
D. <collection>

**Page 92**

What is the purpose of a semi-join?  
A. Another name for join: an operation to combine two tables by column.  
B. Increase the efficiency of sending data across multiple machines.  
C. Increase the speed of the join for trade-off of increased data transmission cost.

**Page 93**

A model that generalizes well means that  
A. The model is overfitting.  
B. The model does a good job of fitting to the noise in the data.  
C. The model performs well on data not used in training.  
D. The model performs well on data used to adjust its parameters.

**Page 94**

Suppose a registration website creates data with the following fields for each person registered (note: if the user does not input a value, NULL is stored instead): Name, Date, Address, and Account Number. Suppose we collect data month by month. Each month, we would have a batch of data containing the fields listed above. At the end of the year, we want to summarize our registrant activities for the entire year, so we would remove redundancies in our data by removing any records with duplicate account numbers from month to month. What type of operation do we use in this scenario?  
A. Join  
B. Not an Operation  
C. Subsetting  
D. Union

**Page 95**

What is the purpose of retrieval and storage; pre-processing; and analysis in order to convert multiple data sources into valuable data?  
A. To enable ETL methods.  
B. Since the multi-layered process is built into the Neo4j database connection.  
C. Designed to work like the ETL process.  
D. To allow scalable analytical solutions to big data.

**Page 96**

Which is the most compelling reason why mobile advertising is related to big data?  
A. Mobile advertising in and of itself is always associated with big data.  
B. Since almost everyone owns a cell/mobile phone, the mobile advertising market is large and thus requires big data to contain all the information.  
C. Mobile advertising benefits from data integration with location which requires big data.  
D. Mobile advertising allows massive cellular/mobile texting to a wide audience, thus providing large amounts of data.

**Page 97**

What is done to the data in the preparation stage?  
A. Retrieve Data  
B. Select Analytical Techniques  
C. Build Models  
D. Identify Data Sets and Query Data  
E. Understanding Nature of Data and Preliminary Analysis

**Page 98**

During construction of a decision tree, there are several criteria that can be used to determine when a node should no longer be split into subsets. Which one of the following is NOT applicable?  
A. The tree depth reaches a maximum threshold.  
B. The number of samples in the node reaches a minimum threshold.  
C. All (or X% of) samples have the same class label.  
D. The value of the Gini index reaches a maximum threshold.

**Page 99**

Which of the following are general requirements for a programming language in order to support big data models? Select all that apply.  
A. Utilize Map Reduction Methods  
B. Support Big Data Operations  
C. Enable Adding of More Racks  
D. Optimization of Specific Data Types  
E. Handle Fault Tolerance

**Page 100**

What are mediated schemas?  
A. Schemas created from customer info.  
B. Schemas created entirely from attribute grouping.  
C. A type of probabilistic schema mapping.  
D. Schema created from integrating two or more schemas.

**Page 101**

The goal of cluster analysis is  
A. To segment data so that differences between samples in the same cluster are maximized and differences between samples of different clusters are minimized.  
B. To segment data so that all samples are evenly divided among the clusters.  
C. To segment data so that all categorical variables are in one cluster, and all numerical variables are in another cluster.  
D. To segment data so that differences between samples in the same cluster are minimized and differences between samples of different clusters are maximized.

**Page 102**

In attribute grouping, how would one evaluate if two attributes should go together? (Choose 2)  
A. Probability of Two Attributes Co-occurring  
B. Integrated Views  
C. Similarity of Attributes  
D. Customer Interaction  
E. Candidate Designs

**Page 103**

In which of the following situations, would you NOT recommend using an agile process?  
A. For a project where change is not expected, requirements are well known, and the solution is both well known and repeatable.  
B. For a project where requirements are well known but the technology and solution are unknown (i.e., changes expected in the solution).  
C. For a project where the customer wants to build an e-commerce website but the development team has never done something like this before. Further, the company knows that they want to build an e-commerce website but are not sure about the functionality at this time.  
D. For a project where the customer has an idea of what they want to build but is not sure of his/her exact needs.

**Page 104**

For two runs of K-Mean clustering is it expected to get same clustering results?  
A. Yes  
B. No

**Page 105**

The feature independence assumption in Naive Bayes simplifies the classification problem by  
A. assuming that the prior probabilities of all classes are independent of one another.  
B. assuming that classes are independent of the input features.  
C. ignoring the prior probabilities altogether.  
D. allowing the probability of each feature given the class to be estimated individually.

**Page 106**

Why are trees useful for semi-structured data such as XML and JSON?  
A. Computers can easily visualize the data with a tree structure.  
B. It is not always the case that XML and JSON can be represented as trees.  
C. Trees take advantage of the parent-child relationship of the data for easy navigation.  
D. They are only useful for XML data as tree-like structure is apparent with tags. While JSON does not contain a tree-like structure as it contains arrays.

**Page 107**

What is the purpose of the acronym BASE?  
A. The same as ACID.  
B. To overcome CAP theorem.  
C. To impose properties on a BDMS in order to guarantee certain results.  
D. Enables stricter enforcement of ACID type design.

**Page 108**

What type of community analytics question is the following?  
Did a community form on twitter around the 2014 World Cup in Brazil?  
A. Static  
B. Prediction  
C. Connection  
D. Evolution

**Page 109**

Which one of the following is the best feature set for your analysis?  
A. Feature set with the smallest set of features that best capture the characteristics of the data for the intended application  
B. Feature set with the smallest number of features  
C. Feature set with the largest number of features  
D. Feature set that contains exclusively re-coded features

**Page 110**

What are the two main components for a data computation framework?  
A. Resource Manager and Container  
B. Applications Master and Container  
C. Node Manager and Applications Master  
D. Resource Manager and Node Manager  
E. Node Manager and Container

**Page 112**

KNIME: In the confusion matrix, what is the difference between low\_humidity\_day and Prediction(low\_humidity\_day)?  
A. low\_humidity\_day is the target class label, and Prediction(low\_humidity\_day) is the predicted class label  
B. low\_humidity\_day is the predicted class label, and Prediction(low\_humidity\_day) is the target class label  
C. There is no difference. The two are the same

**Page 113**

The main steps in the k-means clustering algorithm are:  
A. Assign each sample to the nearest centroid, then calculate the new centroid.  
B. Calculate the centroids, then determine the appropriate stopping criterion depending on the number of centroids.  
C. Calculate the distances between the cluster centroids, then find the two nearest centroids.  
D. Count the number of samples, then determine the initial centroids.

**Page 114**

Which one of the following statements/situations/conversations align with an agile mindset?  
A. A manager providing a status update to a customer says: "We are done with requirements and design so we are 50% done."  
B. Manager: "The customer is suggesting another change in feature X which is complete as per specification. They agree that it was built as we agreed upon but it lacks some functionality and will result in user frustration. Should we implement the change or not?" Developer: "Yes, if it does not fulfill user needs, we should make the change."  
C. A software vendor says to a client: "We must define exactly what the user needs and put that in agreement."  
D. A manager e-mails her team: "Meeting face to face just for a 15-minute discussion is inefficient, let us all just dial-in from our desk."

**Page 115**

Which of the statements below is true?  
A. Dijkstra's algorithm is computationally inefficient (has high computational complexity).  
B. Dijkstra's algorithm is computationally efficient (has low computational complexity).

**Page 116**

What constitutes a community within a graph?  
A. High density of nodes at a certain location.  
B. A neighborhood defined by an integer constant K around a specific node. All K+1 nodes belong in another community.  
C. A dense amount of edge connections between nodes in a community and a few connections across communities.  
D. Many anomalous neighborhoods within the same vicinity.

**Page 117**

Which of these statements is true about samples and variables?  
A. A sample is an instance or example of an entity in your data.  
B. A sample can have many variables to describe it.  
C. A variable describes a specific characteristic of an entity in your data.  
D. All of these statements are true.

**Page 118**

What is a wide transformation?  
A. A transformation that requires data shuffling across node partitions.  
B. Transformations that take a lot of nodes to complete.  
C. A longer time-taking transformation compared to narrow transformations.  
D. The name for the most used transformations.

**Page 119**

Suppose a registration website creates data with the following fields for each person registered (note: if the user does not input a value, NULL is stored instead): Name, Date, Address, and Account Number. Suppose 100 people signup for our system and of the 100 people, 60 of them did not input an address. The system lists the values as NULL for these empty entries in the address field. Would this situation still have structure for our data?  
A. No because the majority of data do not have a specific field filled, thus our originally defined structure is lost.  
B. Yes the data has structure because we have placed a structural constraint on the data, thus the data will always have the originally defined structure.

**Page 120**

What is true between data modeling and the formatting of the data?  
A. There is a one-to-one correspondence between formatting data and data modeling. For every model of data, there is only one way to store the data.  
B. There is always one specific schema for storing model data that is the best and preferred method for the specific data representation.  
C. The data does not necessarily need to be formatted in a way that represents the data model. Just so long as it can be extrapolated.

**Page 121**

What is the minimum no. of variables/features required to perform clustering?  
A. 0  
B. 1  
C. 2  
D. 3

**Page 122**

Which of the following scenarios is NOT an aggregation operation?  
A. Counting the total number of data per type.  
B. Averaging the total number of data per type.  
C. Removing undefined values.  
D. Counting the total number of data.

**Page 123**

What are the 5 key points in order to categorize big data systems?  
A. Execution model, Latency, Scalability, Programming Language, Fault Tolerance  
B. Coordination, Latency, Productivity, Speed, Fault Tolerance  
C. Execution model, Speed, Scalability, Flexibility, Fault Tolerance  
D. Coordination, Latency, Productivity, Flexibility, Fault Tolerance

**Page 124**

Which of the following best describes a way to create and use a validation set to avoid overfitting?  
A. leave-one-out cross-validation  
B. random sub-sampling  
C. k-fold cross-validation  
D. All of the others

**Page 125**

A sample part of the data structure is as follows:  
{\_id:1, userIndex: 10, email: "arealeamil@notreallu.asd", retainRate:2}  
What would be the most likely statement that we would need to grab email info for user indexes greater than 24?  
A. db.userIndex.find({email:{$gt:24}}, {\_id:0})  
B. db.email.find({userIndex:{$gt:24}}, {email:1, \_id:0})  
C. db.userIndex.find({email:{$lte:24}}, {\_id:0})  
D. db.email.find({userIndex:{$lte:24}}, {email:1,\_id:0})

**Page 126**

All of the followings are main benefits of software reuse EXCEPT  
A. Increased dependability  
B. Reduced process risk  
C. Effective use of specialists  
D. All of the others

**Page 127**

What type of algorithm is required for analyzing streaming data?  
A. Accurate and Consistent  
B. Accurate and Memory Efficient  
C. Fast and Complex  
D. Fast and Simple

**Page 128**

In association analysis, an item set is  
A. A transaction or set of items that occur together  
B. A set of transactions that occur a certain number of times in the data  
C. A set of items that two rules have in common  
D. A set of items that infrequently occur together

**Page 129**

What is the main problem with big data information integration?  
A. Pay-as-you-go model  
B. Probabilistic Schema Mapping  
C. Many sources  
D. Mediated Schema

**Page 130**

Which of the Vs results in increased algorithmic complexity (which can cause analyses to not be able to finish running in reasonable amounts of time)?  
A. Valence  
B. Velocity  
C. Volume  
D. Variety

**Page 131**

Of the options below, which query allows you to find the state with the most counties?  
A. source="census.csv" | stats count by CENSUS2010POP | sort count  
B. stats count by STNAME | sort -count  
C. source="census.csv" | stats count by CTYNAME | sort num(count)  
D. source="census.csv" | stats count by STNAME | sort count desc

**Page 132**

What is the out-degree of node B? (Image of a graph is shown)  
A. 0  
B. 1  
C. 2  
D. 3

**Page 133**

A data sample with values that are considerably different than the rest of the other data samples in the dataset is called an/a  
A. Outlier  
B. Invalid data  
C. Noise  
D. Inconsistent data

**Page 134**

Which evaluation metrics are commonly used for evaluating the performance of a classification model when there is a class imbalance problem?  
A. precision and recall  
B. precision and accuracy  
C. accuracy and error  
D. precision and error

**Page 135**

What would we be looking for if we followed the steps below? Note: we have 2 graphs. Create a table for each graph where, for each node, you list the degree of the node. For each graph, create a histogram indicating how many nodes in that graph have a specific degree (e.g., how many nodes have degree 1? 2? etc.). Use advanced approaches (e.g. Euclidean distances) to compare these two histograms.  
A. Connectivity  
B. Centrality  
C. Similarity  
D. Community

**Page 136**

Where does the data for each worker node get sent to after a collect function is called?  
A. Other Worker Nodes  
B. Spark Streaming  
C. Spark Context  
D. None; Stays in the Same Node  
E. Spark SQL

**Page 137**

Updating a graph with a stream of posting information on Facebook is an example of which of the Vs?  
A. Velocity  
B. Volume  
C. Variety  
D. Valence

**Page 138**

Which of the following content would be objects (or nodes) in a graph that represents the activity in a Facebook page (choose 3)?  
A. Created\_post (the action of creating a post)  
B. post text  
C. comment text  
D. location  
E. friends (the action of making someone your friend)