101 Data Science Interview Questions & Answers (Amended) 2019-10-01

Machine Learning

1. Explain **Logistic Regression**.
2. Explain **Linear Regression.**
3. How do you **split data** between training and validation?
4. Explain **Binary Classification**.
5. Describe **Decision trees**.
6. What **metrics** are used to classify a dataset?
7. What is a **cost function**?
8. What's the difference between **convex** and **non-convex** cost functions?
9. Why is **bias-variance** trade off important?
10. What is **regularization**? What are the differences between **L1** and **L2** regularization?
11. Explain **exploding gradients**
12. Explain **activation functions**
13. How is a **box plot** different from a **histogram**?
14. What is **cross validation**?
15. What are **false positives** and **false negatives**?
16. What is **SVM**?
17. You're asked to implement some new features. What experiment would you run to implement these features?
18. What techniques can be used to evaluate a Machine Learning model?
19. Why is **overfitting** a problem?
20. Describe how to detect **anomalies**.
21. What are the **Naive Bayes** fundamentals?
22. What is an **AUC - ROC** Curve?
23. What is **K-means**?
24. What is **Gradient Boosting**?
25. What are **Support Vector Machines** (SVM)?
26. What is the difference between **bagging** and **boosting**?
27. Why do we need **feature engineering**?
28. What is **unbalanced binary classification**?
29. What is the **ROC curve** and the meaning of **sensitivity**, **specificity**, and **confusion matrix**?
30. Why is **dimensionality reduction** important?
31. What are **hyperparameters**?
32. How do you predict if a customer will do something given income, location, profession, and gender?
33. How do you inspect **missing data**?
34. Design the **heatmap** for Uber drivers to recommend where to wait for passengers.
35. What are **time series forecasting techniques**?
36. How does a **logistic regression** model know what the coefficients are?
37. Explain **Principle Component Analysis** (PCA)
38. Explain **Latent Semantic Analysis** (LSA) and **Latent Dirichlet Allocation** (LDA)
39. <skipped>
40. Why is **gradient checking** important?
41. Is **random weight assignment** better than assigning same weights to the units in a hidden layer?
42. How to find an **F1 score**?
43. Describe common **topic modeling** techniques.
44. How does a neural network with one layer, input and output compare to a logistic regression?
45. Why is a **Rectified Linear Unit (ReLU)** a good activation function?
46. How do you use **Gaussian mixture models (GMMs)**?
47. How to decide whether to double the number of ads in Facebook's Newsfeed?
48. What is **Long short-term memory (LSTM)**?
49. Explain the difference between **generative** and **discriminative** algorithms.
50. What is **MapReduce**?

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

51. How do you select the **threshold** for a binary model?
52. Are **boosting** algorithms better than **decision trees**?
53. What are the important factors in uber's driver/rider assignment algorithm?
54. What is **speech synthesis**?

Data Analysis

55. What are the core steps of the data analysis process?
56. How do you detect if a new observation is an **outlier**?
57. <omitted>
58. <omitted>
59. <omitted>
60. What are **anomaly detection** methods?
61. How do you solve for **multicollinearity**?
62. How to optimize marketing spend between various marketing channels?
63. What metrics would you use to measure if Uber's paid advertising strategy is working?
64. What are the core steps for **preprocessing**?
65. How do you inspect **missing data**?
66. How do you use **caching**?

Stats, Probability, Mathematics

67. How would you **define a representative sample** of search queries from 5 million queries?
68. Discuss how to **randomly select a sample** from a product user population.
69. Describe **Markov Chains**.
70. How do you prove that males are on average taller than females by knowing just gender or height?
71. What is the difference between **Maximum Likelihood Estimation (MLE)** and **Maximum A Posteriori (MAP)**?
72. What does **P-Value** mean?
73. What is the **Central Limit Theorem (CLT)**?
74. There are 6 marbles in a bag, 1 is white. You reach in the bag 100 times. After drawing a marble, it is placed back in the bag. What is the probability of drawing the white marble at least once?
75. Explain **Euclidean distance**.
76. Define **variance**.
77. <omitted>
78. What is the **law of large numbers**?
79. How do you weigh 9 marbles three times on a balance scale to select the heaviest one?
80. You call 3 random friends and ask each if it's raining. Each friend has a 2/3 chance of telling you the truth and a 1/3 chance of lying. All three say "yes". What's the probability it's actually raining?
81. Explain a probability distribution that is not normal and how to apply it.
82. You have 2 dice. What is the probability of getting at least one 4?
83. Draw the curve **log(x+10)**

Programming

84. Write a function to check whether a particular word is a **palindrome**.
85. Write a program to generate a **Fibonacci** sequence.
86. Explain about **string parsing** in R language
87. Write a **sorting algorithm** for a numerical dataset in Python
88. Coding test: moving average Input 10, 20, 30, 10, ... Output: 10, 15, 20, 17.5, ...
89. Write a Python code to **return the count of words** in a string Q
90. Write the code for **finding a percentile.**
91. What is the difference between **Stacks** vs

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

Queues and **Linked Lists** vs **Arrays**?

SQL

92. How should you handle **NULLs** when querying a data set?
93. What is the **JOIN** function in SQL?
94. Select all customers who purchased at least two items on two separate days from Amazon.
95. What is the difference between **DDL**, **DML**, and **DCL**?
96. Why is **Database Normalization** Important?
97. What is the difference between **clustered** and **non-clustered indexes**?

Behavioral

98. What was the most challenging project you have worked on so far? Can you explain your learning outcomes?
99. <omitted>
100. How do you avoid **Selection Bias**?
101. <omitted>

Explain Logistic Regression.

- Logistic Regression is a go-to method for classification. It models the probability of the default class.
- It uses the sigmoid function to map any real-valued number to a probability between 0 and 1 to predict the output class.
- There are two types of logistic regression: Binary (2 categories) and Multinomial (3+ categories).
- Assumptions:
    - Binary logistic regression requires the dependent variable to be binary.
    - Independent variables should be independent of each other. (the model should have little or no multicollinearity.)
    - The independent variables should be linearly related to the log odds.

Explain Linear Regression

- Linear regression finds a relationship between two continuous variables.
- One is the predictor (independent) variable; the other is the response (dependent) variable.
- Assumptions:
    - Linear relationship: Between the dependent and independent variable,
    - Multivariate normality: Multiple regression assumes that the residuals are normally distributed.
    - No or little multicollinearity between the independent variables.
    - No autocorrelation: the correlation between the values of the same variables is based on related objects. It violates the assumption of instance independence, which underlies most of the conventional models.
    - Homoscedasticity: This means the variance around the regression line is the same for all values of the predictor variable.

How do you split your data between training and validation?

- ensure the validation set is large enough to yield statistically meaningful results.
- the validation set should be representative of the entire data set.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- k-folds validation is a good choice. It splits the dataset into training and validation sets. This offers various samples of data and ultimately reduces the chances of overfitting.

Describe Binary Classification.

- Binary classification predicts the class of a set of data points.
- The classes are also known as targets/ labels.
- This method approximates a mapping function (f) from inputs (X) to discrete output variables (y).
- For example, spam detection in email service providers is a binary classification since there are only 2 classes: spam and not spam.

Describe decision trees.

- A decision tree classifier uses a training dataset to stratify the predictor space into multiple regions.
- Each such region has only a subset of the training dataset.
- To predict the outcome for a given observation,
    - 1) determine region it belongs to. Once its region is identified, its outcome class is predicted as being the same as the mode (most common) of the outcome classes of all the training observations that are included in that region.
    - The rules used to stratify the predictor space can be graphically described in a tree-like flow-chart, hence the name of the algorithm.
- Decision tree classifiers can handle qualitative predictors without the need to create dummy variables.
- Missing values are not a problem either.
- Decision trees are used for regression models as well.
    - However, one major problem with decision trees is their high variance.

What metrics are used to classify a dataset?

- The performance metrics for classification problems are as follows:
    - Confusion Matrix;
    - Accuracy;
    - Precision and Recall;
    - F1 Score;
    - AUC-ROC Curve.
- Selecting a performance metric depends on the question and dataset.
    - If the dataset is balanced then accuracy would be a good measure.
    - Confusion matrix is a good alternative if you want to know the cost of False Positives and False Negatives.
- To see examples, click here for a brief tutorial.

What is a cost function?

- Cost functions are used to learn the parameters in an ML model such that the total error is minimized.
- It measures a model's ability to estimate the relationship between dependent and independent variables.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- It is typically expressed as a difference between the predicted and actual values. Every algorithm can have its own cost function depending on the problem.

What's the difference between convex and non-convex cost functions?

- A convex function has one global minimum.
    - In convex, an optimization algorithm won't get stuck in a local (non-global) minimum.
    - An example is x^2. It can easily converge at the global minimum.
- A non-convex function has multiple local minimums.
    - Its shape can be visualized with multiple 'valleys' that depict local minima.
    - Algorithms can get stuck in local minimums - it can take a lot of time to identify whether the problem has no solution or if the solution is global.
    - An example is x^6 + x^3 -x^2.

Why is it important to know bias-variance trade off?

- Bias and Variance are part of model prediction errors.
- A model with high bias pays little attention to training data and oversimplifies the model (underfitting).
- A model with high variance pays too much attention to training data and does not generalize on unseen data (overfitting).
- Underfitting/Bias error is the difference between the expected/average prediction of the model and the true value.
    - The model building/prediction process is repeated with new variations of the data.
    - Due to randomness in the underlying data sets, we will have a set of predictions for each point.
    - Bias measures how much the predictions deviate from the true value we are trying to predict.
- Overfitting/Variance error is the variability of model prediction for a given data point.
    - The model prediction is repeated for various data sets.
    - It's an indicator of a model's sensitivity to small variations.
    - For instance, if a model has high variance then small changes in the training data can result in large prediction changes.
    - There is no analytical way to know when we can achieve the bias-variance tradeoff.
    - You can also read: Detecting Algorithmic Bias and Skewed Decision Making.

What is regularization? What are the differences between L1 and L2 regularization?

- Regularization reduces error by fitting a function on the given training set thereby avoiding overfitting.
- The difference between L1 and L2 regularization is the penalty term.
    - Lasso Regression (Least Absolute Shrinkage and Selection Operator), also known as L1, adds an "absolute value of magnitude" coefficient as the penalty term to the loss function.
    - Ridge regression (L2) adds "squared magnitude" of coefficient as the penalty term to the loss function.
- Another difference is that Ridge sets the weights of some features to small values - Lasso shrinks the less important features coefficient to zero (removing some features altogether).

- This works well for feature selection/dimensionality reduction when we have a huge number of features.

What's the problem of exploding gradients?

- Exploding gradients arise during training when gradients are propagated back through the layers.
  - These gradients are being continuously multiplied via matrix multiplication. If they have values larger than 1 then they will eventually blow up resulting in an unstable network.
  - This will hinder the learning process. The weights can become so large that they overflow which results in NaN/undefined values.

Is it necessary to use activation functions?

- Activation functions add non-linearity to a network. If there is no activation function, the input signal will be mapped to an output using a linear function (a polynomial of one degree).
  - Linear functions are not able to capture complex functional data mappings.
  - However, this is possible with the use of nonlinear functions which have a degree of >1.

How is a box plot different from a histogram?

- Boxplots display the distribution of data based on the Minimum, First Quartile, Median, Third Quartile, and Maximum.
  - Each box on the plot shows the range of values from the first quartile at the bottom of the box to the third quartile at the top of the box.
  - A line in the middle of the box occurs at the median of all the values.
  - Plot “whiskers” display the maximum and minimum. They describe the variability and dispersion of data. They also display outliers and describe the symmetry and skewness of the data.
- Histograms show the frequency distribution of continuous data using rectangles.
  - It is similar to a bar graph, however, the bars are adjacent.
  - Data is split into intervals and the frequency of instances in each interval is plotted.
  - It describes the distribution of a graph, its skewness, and indicates the presence of outliers.

What is cross validation?

- Cross-validation evaluates predictive models. It partitions a dataset into a training set to train the model and a test set to evaluate it.
  - In k-fold cv, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model; the remaining k-1 subsamples are used as training data.
  - The cross-validation process is repeated k times, with each of k subsamples used once as the validation data.
  - The k results from the folds can be averaged to produce a single estimation. This allows all observations to be used for both training and validation, and each observation is used for validation exactly once. It helps in reducing bias because cross validation ensures that every observation from the original dataset has the chance of appearing in training and test set.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

What are false positives and false negatives?

- False positive: when a model incorrectly predicts an outcome to be positive when it should have been negative.
- False negative: when an outcome is incorrectly predicted to be negative.
- Both are used as evaluation metrics for classification algorithms.
- Also read: Type I and Type II Errors: Smoke Detector and the Boy who Cried Wolf

What is SVM?

- SVM is a classification and regression algorithm. It identifies a hyperplane which separates the classes in the data.
- A hyperplane is a geometric entity with a dimension of 1 less than its surrounding (ambient) space.
- If an SVM is asked to classify a two-dimensional dataset, it will use a 1D hyperplane (a line).
- Classes in 3D data will be separated by a 2D plane.
- Nth dimensional data will be separated by a N-1 dimension line.
- SVM is also called a margin classifier because it draws a margin between classes.

You're asked to implement some new features. What type of experiment would you run to implement these features?

- A/B testing can be used to check the response on new features by the general audience.
- A/B testing is a marketing experiment wherein you "split" your audience to test variations of a feature and determine which performs better.
- This short tutorial on A/B testing does a great job of describing and visualizing the process.

What techniques can be used to evaluate a Machine Learning model?

- Regression:
    - Mean Absolute Error
    - Mean Squared Error
    - R square
    - Adjusted R square
    - Root Mean Squared Logarithmic Error
- Classification:
    - Classification Accuracy
    - Logarithmic Loss
    - Precision
    - Recall
    - F1 Score
    - Confusion Matrix
    - Receiver Operating Characteristics (ROC) curve
    - Area under Curve (AUC)
    - Gini coefficient

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

Why is overfitting a problem? What steps can you take to avoid it?

- Overfitting is a phenomenon when a model fits too closely on the training data.
    - aka it has "memorized" the data and performs poorly on unseen data because it's not generalized.
    - An overfitted model learns the details of training data such that noise is picked up and learned as concepts.
    - As a result, these concepts learned by the model are not generalized enough to work with unseen data which reduces the predictive ability.
- Overfitting can be reduced by using the following:
    - Resampling techniques such as k-fold cross validation that creates multiple train-test splits,
    - Using Ensembling techniques that combines predictions from separate models;
    - Using Regularization to add a penalty to the cost function - making models more flexible

Describe a way to detect anomalies.

- Anomaly detection is used to identify unusual patterns that do not conform to expected behavior, called outliers.
- Typically, anomalous data can be connected to some kind of problem or rare event such as e.g. bank fraud, medical problems, structural defects, malfunctioning equipment, etc.
- The simplest approach is to use statistical techniques to flag data that deviates from common statistical properties of a distribution, including mean, median, mode, and quantiles.
- Statistical approaches are difficult in higher dimensions. ML techniques can help in these cases.
    - Isolation Forest (see below)
    - One Class SVM
    - PCA-based Anomaly detection
    - FAST-MCD
    - Local Outlier Factor
- Isolation Forests build a Random Forest in which each Decision Tree is grown randomly.
    - At each node, it picks a feature randomly, then picks a random threshold value (between the min and max value) to split the dataset in two.
    - The dataset gradually gets chopped into pieces this way, until all instances end up isolated from the other instances.
    - Random partitioning produces noticeably shorter paths for anomalies.
    - Hence, when a forest of random trees collectively produces shorter path lengths for particular samples, they are highly likely to be anomalies.

What are the Naive Bayes fundamentals?

- Naive Bayes is a probabilistic model that is used for text classification.
- It learns the probability of an object with a certain feature belonging to a particular group of class.
- The Naive Bayes algorithm is "Naive" because it assumes the occurrence of a feature is independent of the occurrence of other features.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- The classifier is based on the Bayes theorem. It gives us a method to calculate a conditional probability, of an event A based on the previous events.
- There are three types of Naive Bayes:
    - Multinomial: Used when we have discrete data. With respect to text classification, if the words can be represented by their occurrences/frequency count, then use this method.
    - Bernoulli: It assumes input features are binary with only two categories. If you just care about the presence or absence of a word in a document, use Bernoulli classification.
    - Gaussian: It is used with continuous features. For example, the Iris dataset features have sepal width, petal width, sepal length, and petal length.
- To unfold Naive Bayes with Data Science Dojo, click here for part 1 and click here for part 2.

What is an AUC - ROC Curve?

- When we need to evaluate or visualize the performance of the multi-class classification problem, we use AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve.
- ROC is a probability curve and AUC represents the degree of separability.
- It tells how much the model is capable of distinguishing between classes such as spam/not-spam.
- The higher the AUC, the better the model is at predicting spam email as spam and non- spam email as non-spam.
- A highly accurate model has AUC close to 1 which reflets it's good measure of separability. A poor model has AUC near 0 which means it has worst measure of separability.

What is K-means?

- K-means is an unsupervised clustering algorithm.
- K-means randomly identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
- The result is that the input unlabelled data is converted into clusters which are differentiable.

What is Gradient Boosting?

- Gradient Boosting sequentially adds predictors to an ensemble - each predictor is correcting its predecessor.
- imagine the boosting problem as an optimization problem, where we take a loss function and try to optimize it. It has 3 core elements:
    - a weak learner to make predictions,
    - a loss function to be optimized, and
    - an additive model to add to the weak learners to minimize the loss function.

What are Support Vector Machines (SVM)?

- Advantages
    - It has a regularization parameter, which can be tweaked to avoid over-fitting.
    - SVM uses the kernel trick, so you can build a modified version of a model depending on the problem complexity.

    - It works well with less data.
- Disadvantages
    - The choice of kernel according to the problem type is tricky to choose. Kernel models are usually quite sensitive to over-fitting so a lot of knowledge is required to make sound decisions.
    - It is difficult to tune the hyperparameters such that the error is the minimum.

What is the difference between bagging and boosting?

- Bagging and Boosting are both ensemble techniques
    - a set of weak learners are combined to create a strong learner that obtains better performance than a single one.
- In Bagging, each model is trained in parallel and is running independently.
    - The outputs are aggregated at the end without preference to any model.
- Boosting is all about “teamwork”. Each previous model decides the subset of features used by the next model depending on the performance.
    - The choice of the model to use depends on the data.

Why do we need feature selection/engineering?

- The data features used to train your models have a huge influence on the performance of the model.
- Some feature sets will be more influential than others on model accuracy.
- Irrelevant features can increase the complexity of the model and add noise to the data which can negatively impact model performance.
- Features may be redundant if they're highly correlated with another feature. These types of features can be removed from the data set without any loss of information.

- Feature selection methods can identify and remove redundant attributes from data that don't contribute to the accuracy of a predictive model.
- Moreover, variable selection helps in reducing the amount of data that contributes to the curse of dimensionality. Reducing the number of features through feature selection ensures training the model will require minimum memory and computational power, leading to shorter training times and also reducing the common problem of overfitting.

What is unbalanced binary classification?

- Unbalance Binary Classification can be dealt in multiple ways:
    - Under-sampling: Eliminates majority class examples until data is balanced.
    - Over-sampling: Increases number of instances in a minority class by adding copies of those instances. This can be done randomly or synthetically using Synthetic Minority Over-sampling Technique (SMOTE).
- Use suitable performance metrics: Accuracy can be a misleading metric for unbalanced data.
    - Suitable metrics would include Precision, Recall, F1-score, AUC, etc.
- Use suitable algorithms:
    - Algorithms such as Decision Tree and Random Forest work well with unbalanced data.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- Penalizing wrong classification imposes an additional cost on the model for making classification mistakes on rare classes during training more than wrong classifications of the abundant class. These penalties can create bias in the model to pay more attention to and favor of the rare class.

What is the ROC curve and the meaning of sensitivity, specificity, and confusion matrix?

- A Receiver Operating characteristic (ROC) Curve plots True Positive Rate (TPR) vs False Positive Rate (FPR) at different classification thresholds. It tells us how a good a model is in classification. Therefore, curves of different models can be compared directly in general or for different thresholds.

- Sensitivity is a measure of correctly identified positives over all positives. It's also called the TPR and can be explained as:
    - Sensitivity = true positives/(true positive + false negative)

- Specificity is a measure of correctly identified negatives over all negatives. It's also called the TNR (true negative rate) & can be explained as:
    - Specificity = true negatives/(true negative + false positives)

- A confusion matrix is a table that provides summary data of a classification algorithm. It returns combinations of predicted and actual values which include True Positive and False Positive in the 1st row, whereas, False Negative and True Negative in the second row.
- To see an example of the confusion matrix, watch this short tutorial.

Why is dimensionality reduction important?

- Datasets with large number of feature sets (specifically images, sound, and/or textual contents) increase space, add overfitting and slow down the time to train the models.

- Dimensionality reduction is the process of reducing the dimensionality of a feature space by obtaining a set of principal features. This way, it can bootst performance of the learning algorithm resulting in less computational cost with simplification of models.

- It also eliminates redundant features and features with strong correlation between them, therefore, reducing overfitting. Moreover, projection into 2-3 dimensions often used for visualization of high-D data sets, leading to better human interpretations.
- Read more about dimensionality reduction on this blog.

What are hyperparameters, how to tune them | test them | know if they worked?

- While a machine learning model tries to learn parameters from the training data, hyperparameters are are set before the training process begins.
- They describe how a model is supposed to function, for example: the number of trees in a decision tree or the learning rate of a regression model. Hyperparameters directly control the behavior of the training algorithm and have a significant impact on the performance of the model being trained.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- To choose the optimal set of hyperparameters for providing the best results, they can be tuned by running multiple trials. Each trial is a complete execution of our training application with a broad set of values we specify. Some common techniques are: Grid search, Random Search, Bayesian Optimization.

- To find out if the specified hyperparameters work, we need to test them against an evaluation metric based on the nature of our problem. For example, we can choose a set of hyperparameters that give us the best accuracy or the best F1 score.

How do you predict if a customer will do something given income, location, profession, and gender?

- This is a classification algorithm.
- To solve this problem, the dataset will be collected and stored.
    - It will be cleaned and pre-processed for any abnormalities or missing values.
    - Then it will be subjected to feature engineering.
        - Some steps may include dealing with missing values, encoding categorical values, and normalizing numerical values.
- The dataset would then be divided into train and test sets, using K-folds validation with k set to an appropriate number of folds. The train set would be used to fit a classification model such as a Logistic Regression, Decision Tree Classifier or Support Vector Classifier.
- The fitted model is then evaluated against the test set to check how good the model is using an evaluation metric such as accuracy or F1 score.

How will you inspect missing data? When are they important?

- Missing data can be inspected using various techniques depending on the language being used.
- In Python, the isnull function can be used to find the missing data that is marked with NaN.
- In R, missing values can be identified using is.na function.
- Summary statistics could be used to point out missing data in quantitative variables where they might be marked as zero where zero would be an abnormality, for example: 0 in the ‘age of employee’ variable.
- Missing values can reduce the fitness of our model
- Therefore, they can lead to wrong prediction. However, missing data can let us know why certain variables were difficult to collect.
- They can also form correlation with other variables which is considered as missing not at random.
- Missing values can point out if a variable was appropriately gathered, such as asking a personal question in a survey that will be skipped by participants.

Design the heatmap for Uber drivers to recommend where to wait for passengers.

- To design the heatmap, some of the pointers are listed as follows:
- You can use k-means clustering to group previous journeys of the customers in similar area. This will give a fair idea about the preference of the potential rides.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- Perform exploratory data analysis to analyze how long it took for a driver to find the client once they arrived at the pick-up location. Filter out those locations with the minimum pickup time.
- The model can use maps to identify whether it is possible to pick up people at those points in terms of practicality. For instance, it would be inconvenient to pick up people from congested areas so a nearby pickup point should be suggested to ensure efficiency and quick service.

What are time series forecasting techniques?

- Simple moving average (SMA) is the simplest. It adds up the last 'n' period's values, then dividing that number by 'n'. So the moving average value is used as the forecast for the next period.
- Exponential Smoothing: Exponential Smoothing assigns exponentially decreasing weights as the observations get older.
- Autoregressive Integrated Moving Average (ARIMA): The parameters are (P, d, q) - the autoregressive, integrated and moving average parts of the data set, respectively.
    - ARIMA handles the trends, seasonality, cycles, errors and non-stationary aspects of a data set when making forecasts.
- Neural networks: They are also used for time series forecasting.
- tutorials ; introduction

How does a logistic regression model know what the coefficients are?

- First, let's consider the case when the input variable is continuous.
    - The first coefficient is the y-axis intercept.
        - It means that when the input variable/feature is 0 the log(odds of output variable) is equal to the intercept value.
    - The second coefficient is the slope.
        - It represents the change of value in the log(odds of output variable) per unit of x-axis gained.
        - Consider the case when the input variable is discrete. Example: where a mouse is "obese" or "not obese".
            - The independent variable is discrete - whether the mouse has normal or mutated genes.
            - The first coefficient/intercept tells us the log(odds of normal gene) and the second coefficient tells us the logodds ratio which determines, on a log scale, how much having a mutated gene increases/decreases the odds of being obese.

Explain Principle Component Analysis (PCA).

- Principal component analysis is a dimensionality reduction technique for large data sets,
- It transforms a large set of variables into a smaller one that still contains most of the information in the large set. It's often used to make data easy to explore and visualize.
- PCA does not make any explicit assumptions.
- To uncover the maths and theory behind PCA, read this blog.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

Explain Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

- Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are used for topic modelling.
- LSI (also known as Latent Semantic Analysis, LSA) learns latent topics by performing a matrix decomposition on the term-document matrix.
- The objective of LSA is to reduce dimensions for classification in NLP.
- Latent Dirichlet Allocation (LDA) is a “generative probabilistic model” which uses unsupervised learning for topic modeling/classification of topics.

<skipped>

Why is gradient checking important?

- Gradient Checking is used check out the derivatives in back-propagation algorithms.
- Backprop algorithm implementation is prone to bugs and errors.
- It’s necessary before running the neural network on training data to verify that our implementation of back-propagation is correct.
- Gradient checking is a way to do that. It compares back-propagation gradients, which are obtained analytically with a loss function, with numerically obtained gradients for each parameter.
- It ensures that the implementation is correct and increases our confidence in the correctness of our code.
- By numerically checking the computed derivatives, gradient checking eliminates problems that can occur as back-propagation can have subtle bugs.

Is random weight assignment better than assigning same weights to the units in a hidden layer?

- Consider a situation where the weights are assigned equally.
    - Since neural networks use gradient descent to optimize parameters and find the lowest point to reduce the error of the cost function, they need an initialization point.
        - If the starting point is A at the first iteration then it is possible that the network is unable to find a path towards the local minima.
        - Keeping the initialization point consistent each time will lead to the same conclusion.
        - If the starting point is random at each iteration, the network will have a better chance at finding local minima to reduce the error of the cost function.
        - This technique is also known as breaking the symmetry.
        - The initialization is asymmetric so we can find various solutions to the same problem.

How to find the F1 score after a model is trained?

- F1 score is an evaluation metric for classification algorithms.
    - It is derived from precision (ratio of true positives to all positives labeled by algorithm)
    - and recall (ratio of true positives to all positives in reality).
- F1 score is the harmonic mean of precision and recall. It seeks a balance between Precision and Recall in uneven class distributions.
- It can be formulated as 2* ((recall * precision) / (recall+precision))

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- F1 score = 1, when precision and recall are perfect. Whereas, the worst case F1 score is 0. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is.

Describe common topic modeling techniques.

- Latent Semantic Analysis (LSA) uses the context around the words to find hidden concepts.
    - It does that by generating a document-term matrix.
    - Each cell has TD-IDF score which assigns a weight for every term in the document.
    - Using Singular Value Decomposition (SVD), the dimensions of the matrix are reduced to a number of desired topics.
    - The resulting matrices, after decomposition, gives us vectors for every document & term in our data
        - It can then be used to find similar words and similar documents using cosine similarity.

- Probabilistic Latent Semantic Analysis (PLSA) is used to model information under a probabilistic framework instead of SVD.
    - It creates a model P(D,W): for any document d and word w, P(d,w) corresponds to that entry in the document-term matrix.

- Latent Dirichlet Allocation (LDA) automatically discovers topics in documents.
    - LDA represents documents as mixtures of topics that spit out words with certain probabilities.
    - It assumes each document mixes with various topics and every topic mixes with various words.
    - LDA tries to backtrack from the documents to find a set of topics most likely to have generated the collection.
    - It maps all documents to topics such that the words in each document are mostly captured by those topics.

How does a neural network with one layer, input and output compare to a logistic regression?

- Neural networks and logistic regression are both used for classification problems.
- Logistic regression can be described as the simplest form of Neural Network that results in straightforward decision boundaries.
- Neural networks are a superset that includes additional decision boundaries to cater to more complex and large data.
- Logistic regression models cannot capture complex non-linear relationships w.r.t features.
- A neural network with non-linear activation functions enables one to capture highly complex features.

Why is Rectified Linear Unit/ReLU a good activation function?

- ReLUs are better for the training of deep neural networks then the traditional sigmoid or tangent activation functions because they help solve the problem of vanishing gradients.
    - The problem of vanishing gradient occurs while back-propagating weights through the layers which tend to get smaller as we keep moving backwards in the network.
    - Due to this, the learning is very slow for large values of the input as gradient values are small.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- When a neuron's activation saturates close to 0, the gradients at these regions are close to 0.
- During back-propagation, this local gradient will be multiplied with the gradient of the state's output. Hence, if the local gradient is really small, it will make gradients slowly vanish.
- Therefore almost no signal will flow through the neurons to it's weights.
- ReLUs are faster in learning. They are only used for the hidden layers (why?) of the neural networks in deep learning.

How do you use Gaussian mixture models (GMMs)?

- A GMM is a probabilistic model that assumes data was generated from a mixture of several Gaussian distributions with unknown parameters.
- We describe each cluster by its centroid (mean), covariance, and the size (weight) of the cluster.
- Therefore a GMM us applicable when we know that the data points are mixtures of a gaussian distribution and form clusters with different means and standard deviations.

How to decide whether to double the number of ads in Facebook's Newsfeed?

- You can use A/B testing to make a conclusion about the success rate of the ads.
- A/B testing is experimenting and comparing two variations of a campaign such as ad text, a headline, or any element of a marketing campaign.
- For example, one set of the audience can be shown ads that are double the amount they usually see on their newsfeed while the second set will continue to see the existing number of ads.
- Even a small sample size in an A/B test can provide significant, actionable insights.

What is Long short-term memory (LSTM)?

- LSTM is based on a recurrent neural network (RNN) architecture.
- LSTM tackles the problem of long-term dependencies of RNNs in which the RNNs cannot predict the word stored in long-term memory but can give more accurate predictions from recent information.
- LSTM explicitly introduces a memory unit, called the cell, into the network. LSTM can retain the information for a long period of time. (?)
- A common LSTM unit contains a cell, an input gate, an output gate, and a forget gate.
- Each unit makes decisions by considering the current input, previous output and previous memory. It generates a new output and alters its memory.
- LSTM is used for processing, predicting and classifying based on time series data.
- Unlike standard feedforward neural networks, LSTM has feedback connections that make it a general-purpose computer.
- It can process single data points (such as images), and sequences of data (such as speech or video).

Explain the difference between generative and discriminative algorithms.

- Suppose we have a dataset with training input x and labels y.
- A Generative Model explicitly models the actual distribution of each class.
    - It learns the joint probability distribution, p(x, y),
    - and makes predictions using Bayes rules to calculate p(ylx).

    - It then picks the most likely label y.
    - Examples:
        - Naïve Bayes,
        - Bayesian Networks
        - Markov random fields.

- A Discriminative Model learns the conditional probability distribution p(y|x),
    - or learns a direct map from inputs x to the class labels.
    - It models the decision boundary between classes.
    - Some popular discriminative classifiers:
        - Logistic regression,
        - Traditional neural networks
        - Nearest neighbors.

What is MapReduce?

- MapReduce is a data processing tool that enables distributed computations for handling large datasets.
- It is used to split and process data in parallel, achieving quicker results.
- MR makes it easy to scale data processing over multiple computing nodes.
- Processing uses the map and reduce functions.
    - Map converts a dataset into another -individual elements are broken down into tuples.
    - Reduce accepts a map's output & combines those tuples into a smaller set of tuples.

How do you select the threshold for a binary model?

- We need to understand what will happen as a result of selecting a decision boundary.
- You need to know the relative cost of a false positive vs. a false negative.
- A precision-recall curve of your model can be plotted on your validation data.
- Example: it's important to know that if you accidently label a true potential customer as false, this will result in losing customers. This analysis will help in deciding the right threshold for the model.

Are boosting algorithms better than decision trees? Why?

- Yes, they perform better than decision trees.
- Boosting algorithms combine several weak learners into one strong one.
- They create a sequence of models that attempt to fix the errors of previous models.
- Models are added until the training set is predicted perfectly or a maximum number of models are added.
- During this process, if an input is misclassified by a hypothesis, its weight is increased so that the next hypothesis is more likely to classify it correctly.
- This process converts weak learners into better models. The results are combined to create a final output prediction.

What are the important factors in uber's driver/rider assignment algorithm?

- The following list of factors can be used to assign rides to drivers:
    - Drivers who are online at the time of the request.
    - Drivers who have a good reputation (never been rated lower than 3/4 by the passenger making the request).
    - Drivers who are closest to the requesting passenger.
    - Drivers who don't have a destination filter set that excludes the passenger's destination.

How does speech synthesis work?

- Speech synthesis is the process of creating human like speech.
- It's also referred to as a text-to-speech (TTS) system that converts natural language into audible speech.
- It's a form of output where a computer reads words to you out loud in a real or simulated voice.
- This synthesized speech is usually generated by concatenating pieces of recorded speech.
- The entire process could be described as:
    - Pre-processing: Since there's a lot of ambiguity involved in reading text, as words can be read in several ways, the pre-processing tries to eliminate the ambiguity and handles homographs.
    - Words to Phenomes: the algorithm uses phonemes to convert text into sequences of sounds. Phonemes are the sound components used to make spoken words.
    - Phenomes to sound: Last, output techniques are used to mimic human voices read out the text. This can be in 3 forms:
        - Using recordings of humans saying the phonemes
        - Using a computer to generate the phonemes by generating basic sound frequencies
        - Mimicking the mechanism of the human voice

Data Analysis

What are the core steps of the data analysis process?

- Data analysis is a process in which we can change or analyze data to draw a conclusion about a goal.
- It involves inspecting, cleansing, transforming and modelling tasks to discover useful information.
- This way it can be used for creating conclusions and supporting decision-making.
- Data Analysis Process consists of the following iterative phases:
    - Setting of goals: This is the first step of the process.
    - Data gathering: The emphasis is on ensuring accurate and honest collection techniques.
    - Data Processing: This step requires organizing and structuring data in proper format to simplify the workflow. It often involves encoding and standardizing variables for better interpretation
    - Data Cleaning: This is where you'll find, change or remove any incorrect or redundant data. Data scientists correct spelling mistakes, handle missing data and weed out nonsense information. This is the most critical step in the data value chain.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

    - Data Analysis: various techniques can be used to derive conclusions based on the requirements. Here we can explore the data, find correlations among features and identify relevance of each feature to the problem.
    - Result interpretation: It is important to know if the data answers your original question and helps in defending against any objections. For these steps, we can use machine learning algorithms as well as descriptive and inferential statistics.
    - Communication of Results: This is the last step of this process and can be called storytelling. Here we try to communicate this to other teams & management using visualization tools.

How do you detect if a new observation is an outlier?

- Use Boxplot/Whiskers plot to visualize the outlier:
    - Any value that will be more than the upper limit or smaller than the lower limit of the plot will be outliers.
    - Only the data between the Lower and Upper limits is statistically considered normal and can be used for further analysis.
- Standard deviation: Find the points which lie more than 3 times the standard deviation of the data.
- Clustering: Use K-means or Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect outliers.

<omitted>
<omitted>
<omitted>

What are anomaly detection methods?

- Anomaly detection is used to identify patterns that do not conform to expected behavior.
- The simplest approach is to flag data points that deviate from common statistical properties of a distribution, including mean, median, mode, and quantiles.
- Density-Based Anomaly Detection works on the assumption that normal data points occur around a dense neighborhood and abnormalities are far away.
    - The nearest set of data points are evaluated using a score (ex: Euclidean distance).
    - Another technique to detect anomalies is Z-score, which is a parametric outlier detection method.
        - It assumes a Gaussian distribution of the data.
        - The outliers are are in the tails of the distribution and therefore far from the mean.

How do you solve for multicollinearity?

- Multicollinearity occurs when independent variables in a regression model are correlated.
- To solve this issue, remove highly correlated predictors from the model.
- Regularization can be used because it stabilizes the regression coefficients so the effect of multicollinearity is mitigated.
- You can also use PCA to cut the number of correlated predictors.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

How to optimize marketing spend between various marketing channels?

- Choose a set of metrics that will determine which channels get more investment. Read our blog to know what metrics you should be using for email marketing

What metrics would you use to measure if Uber's paid advertising strategy is working?

- Customer acquisition cost (CAC)can be used to track consumers/customers as they progress from interested leads to acquiring customers.

What are the core steps for preprocessing?

- Discovering/Data Acquisition: Gather the data from all the sources and try to understand and make sense of your data.
- Structuring/Data Transformation: Since the data may come in different formats and sizes so it needs to have a consistent size and shape when merged together.
- Cleaning: This step consists of imputing null values and treating outliers/anomalies in the data to make the data ready for further analysis.
- Exploratory Data Analysis: Try to find patterns in the dataset and extract new features from the given data in order to optimize the performance of the applied machine learning model.
- Validating: This stage verifies data consistency and quality.
- Publishing/Modeling: The wrangled data is ready to be processed further by an algorithm or machine learning model.

How do you inspect missing data?

- Imputation of missing values depending on whether the data is numerical or categorical.
- Replacing values with mean, median, mode.
- Using the average value of K nearest neighbours as an imputation estimate.
- Using linear regression to predict values.

How do you use caching in Data science?

- It is often necessary to save data files when loading and/or manipulating data takes considerable time.
- When you want to access some data that is expensive to look up (in terms of time/resources), you cache it so that the next time you need it, it's much less expensive and time efficient.
- Caching also enables content to be retrieved faster because an entire network round trip is not necessary. Caches like the browser cache can make information retrieval nearly instantaneous.

Statistics, Probability and Mathematics

How would you define a representative sample of search queries from 5 million queries?

- Some key features need to be kept in mind:

- Diversity: A sample must be as diverse as the 5 million search queries. It should be sensitive to all the local differences between the search query and should keep those features in mind.
- Consistency: We need to make sure that any change we see in our sample data is also reflected in the true population which is the 5 million queries.
- Transparency: It is extremely important to decide the appropriate sample size and structure so that it is a true representative. These properties of a sample should be discussed to ensure that the results are accurate.

Discuss how to randomly select a sample from a product user population.

- Sampling techniques can be divided into two categories:
  - Probability sampling methods
    - Simple Random Sampling
    - Stratified Sampling
    - Clustered Sampling
    - Systematic Sampling
  - Non-Probability sampling methods
    - Convenience Sampling
    - Snowball Sampling
    - Quota Sampling
    - Judgement Sampling

Describe Markov Chains.

- Markov Chains can be used in marketing analytics.
- They are a stochastic model describing a sequence of possible events that are probabilistically related to each other.
  - The probability of the upcoming event depends only on the present state - not previous states.
  - This is called the Memoryless property. It disregards the events in the past and uses the present information to predict what happens in the next state.
- Imagine you have an online product selling platform:
  - You want to know whether a customer's in the stage where they are considering to "buy a product". (These are the states at which the customer would be at any point in their purchase journey.)
  - Markov Chains provide information about the current state & transition probabilities of moving from one state to another. As a result, we can predict the next stage. In this case, we can predict how likely a customer is going to buy the specified product.

How do you prove that males are on average taller than females by knowing just gender or height?

- We can use the concept of Null and Alternate hypothesis to prove this.
- It is used for statistical significance testing.
- First: compare the sample means of the male heights vs female heights.
- The Null hypothesis: “the mean female height and male height are the same.”

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- The alternate hypothesis: “the mean male height is greater than mean female height.”
- A one tailed hypothesis test can be used to accept or reject the Null Hypothesis.
    - P-value analysis can be used to figure out whether the test is statistically significant or not.

What is the difference between Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP)?

- Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) are methods for estimating a variable in probability distributions or graphical models.
    - MAP usually comes up in Bayesian settings. It works on a posterior distribution - not only the likelihood like MLE.
        - If you have useful prior information, posterior distribution will be more informative than the likelihood function.
    - Comparing both MLE and MAP equation, the only difference is the inclusion of prior P(θ) in MAP - otherwise they are identical.

What does P-Value mean?

- P-Values are used to determine the statistical significance in the Null Hypothesis.
    - It stands for probability value and indicates how likely a result occurred by chance alone.
    - Small p-values indicate the result is unlikely to have occurred by chance alone.
        - These results are known as being statistically significant.
    - Large p-values indicate the result is within chance or normal sampling error.
        - This means nothing happened and the test is not significant.
        - A large p-value beyond a chosen significance level indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.

What is the Central Limit Theorem (CLT)?

- CLT states that the sampling distribution of the sample mean approaches the normal distribution as the sample size gets larger no matter what the initial shape of the population distribution is.
- CLT helps us quantify the probability that the random sample will deviate from the population without having to take any new sample to compare it tp.
- Because of this theorem, we don’t need the characteristics about the whole population to understand the likelihood of our sample being representative of it.
- Confidence intervals, hypothesis testing, and p-value analysis is based on the CLT.
- In a nutshell, CLT can make inferences from a sample about a population.

There are 6 marbles in a bag, 1 is white. You reach in the bag 100 times. After drawing a marble, it is placed back in the bag. What is the probability of drawing the white marble at least once?

- The probability of drawing out at least one marble is the complement of probability of drawing not a single white marble at all. Therefore, we’ll calculate the Probability of drawing all non-white marbles over a hundred times and subtract by 1:

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

P(White at least once) = 1 – [P(Non-white marbles) ^ 100] = 1 - [(5/6) ^ 100]

Explain Euclidean distance.

- Euclidean distance is used to calculate distances between 2 points P and Q. It stems out from the Pythagoras theorem where the distance from point P to Q (in 2-dimensional space) is calculated by considering the line P to Q as hypotenuse of a triangle.
- In n-dimensional space, the Euclidean distance can be generalized using the following formula: d(p,q) = under-root(∑ni=1(pi−qi)2)
- It is used to measure the "similarity" between two vectors.

Define variance.

- Variance is a measure of the variability of data in a distribution.
- It measures how far a set of (random) numbers are spread out from their average value. It can be formulated as the average of the squared differences from the mean.

<omitted>

What is the law of large numbers?

- As the sample size is increased, the sample mean approaches the true population mean.

How do you weigh 9 marbles three times on a balance scale to select the heaviest one?

- Divide 9 marbles into groups of 3. The groups have 4, 4 and 1 marbles respectively.
- Weigh the 2 groups with 4 marbles each.
- If the scale is balanced, then the 1 marble from the last group is the heaviest.
- If one group is heavier, divide it into 2 subgroups of 2 marbles each.
    - Weigh them and find the heaviest group. Now we're left with 2 marbles from the selected group.
    - Weigh those 2 marbles. We are left with the heaviest marble.

You call 3 random friends and ask each if it's raining. Each friend has a 2/3 chance of telling you the truth and a 1/3 chance of lying. All three say "yes". What's the probability it's actually raining?

- We have to find the probability of raining given that all three friends said 'Yes':
  P(rain | yes, yes, yes)

- Using Bayes Theorem, our equation will now be:
  P(rain | yes, yes, yes)
  = P(yes,yes,yes|rain) *P(rain)
  / [P(yes,yes,yes|rain) *P(not rain)
  + P(yes,yes,yes|rain) *P(not rain)]

- We have the following values:
    - P(yes, yes, yes | rain) = 2/3 ^ 3 = 8/27

  - P(yes, yes, yes | not rain) = 1/3 ^ 3 = 1/27
  - P(rain) = R (it is not given in question, so we'll assume R)
  - P(not rain) = 1 - R

- Substituting these values in equation we get:
  - P(rain | yes, yes, yes) = 8P/(7P + 1)

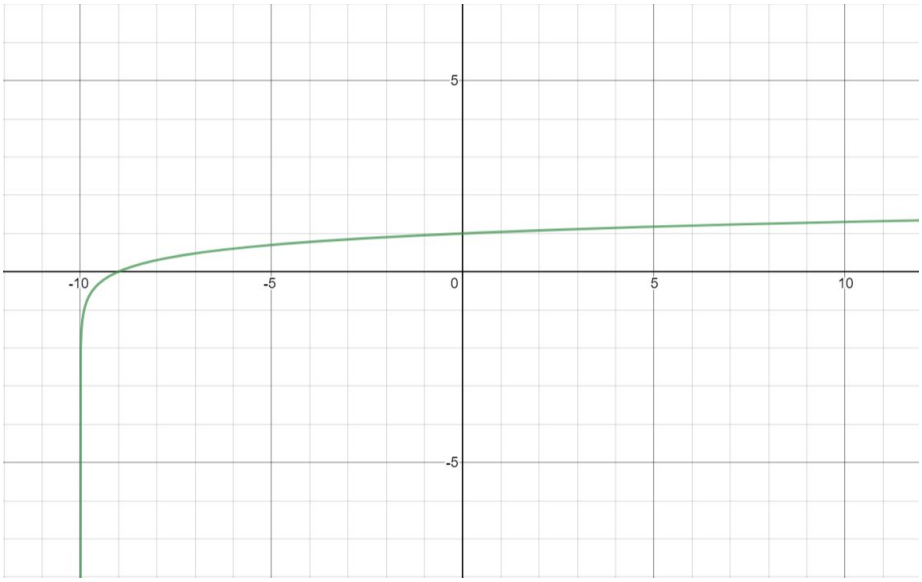
Explain a probability distribution that is not normal and how to apply it.

- A Poisson distribution is a discrete probability distribution that helps to predict the probability of events from happening using how often the event has occurred.
  - It predicts the probability of a given number of events occurring in a fixed interval of time.
  - Examples of Poisson distribution include the number of phone calls received by a call center per hour and the number of decay events per second from a radioactive source.
- The Poisson distribution is applied when:
  - If the event is possible to count and can be counted in whole numbers
  - If the average frequency of occurrence for the time period in question is known
  - When the occurrences are independent

You have 2 dice. What is the probability of getting at least one 4?
Find the probability of getting at least one 4 if you have n dice.

- For 2 die the probability of getting at least one four is:
  - P(at least 1 four) = 1 – P(No four) = 1 – 5/6*5/6 = 1 – (5/6)^2 = 11/36
  - The probability with n dice will be: P(at least 1 four) = 1 – P(No four) = 1 – 5/6n

Draw the curve log(x+10)



101 Data Science Interview Questions & Answers (Amended)
2019-10-01

Programming

Write a function to check whether a particular word is a palindrome.

(Using R)

```
Palindrome <- function(word){
   rawword <- charToRaw(tolower(word))
   if(identical(rawword, rev(rawword)) == 1){
      print("Palindrome")
   }
   else{
      print("Not Palindrome")
   }
}
```

(Using Python)

```
def Palindrome(word):
  reverse = word[::-1]
  if word == reverse:
   print ("Palindrome")
  else:
   print ("Not Palindrome")
```

Write a program to generate a Fibonacci sequence.

(Using R)

```
Fibonacci <- function(n){
   if(n<=1){
     print("Invalid Input")
   }
   else if(n == 2){
      print(0)
      print(1)
   }
   else{
      a <- 0
      b <- 1
      print(a)
      print(b)
      for(i in 0:(n-3)){
         sum <- a+b
         print(sum)
         a <- b
         b <- sum
      }
   }
}
Fibonacci(8)
```

(Using Python)

```
def Fibonacci(n):
 if n<1:
   print('invalid input')
 elif n == 1:
   print(0)
 else:
   a = 0
   b = 1
   print(a)
   print(b)
   for i in range(n-2):
     sum = a + b
     print(sum)
     a = b
     b = sum
Fibonacci(8)
```

101 Data Science Interview Questions & Answers (Amended)
2019-10-01

Explain about string parsing in R language

- In R, you use paste() to concatenate and strsplit() to split.

Write a sorting algorithm for a numerical dataset in Python.

```
def sort(mylist):
  n = len(mylist)
  for i in range(n):
    for j in range(0, n-i-1):
      if mylist[j] > mylist[j+1]:
        mylist[j], mylist[j+1] = mylist[j+1], mylist[j]
    print(mylist)

sort([80, 55, 70])
```

Coding test: moving average Input 10, 20, 30, 10, ... Output: 10, 15, 20, 17.5, ...

(Using R)

```
moving_avg <- function(mylist){
    mysum <- 0
    for (i in 1:length(mylist)){
        mysum <- mylist[i] + mysum
        avg <- mysum/(i)
        print(avg)
    }
}
moving_avg(c(10, 20, 30, 10))
```

(Using Python)

```
def moving_avg(mylist):
  mysum = 0
  for i in range(len(mylist)):
    mysum += mylist[i]
    avg = mysum/(i+1)
    print(avg)
moving_avg([10, 20, 30, 10])
```

Write a Python code to return the count of words in a string Q

```
def count_words(my_string):
        count = len(my_string.split())
        print ("The number of words in string are : " +  str(count))

count_words(Q)
```

Write the code for finding a percentile.

101 Data Science Interview Questions & Answers (Amended)
2019-10-01

(Using R)	(Using Python)

percentile <- function(data, score){
  size <- length(data)
  sorted <- sort(data)
  score_index <- match(score, sorted) -1
  perc <- (score_index/size)*100
  print(perc)
}
 	
percentile(c(80, 55, 70, 44, 33, 21, 65, 90, 12, 18), 55)
	def percentile(data, score):
	  size = len(data)
	  score_index = sorted(data).index(score)
	  perc = (score_index/size)*100
	  print(perc)
	
	percentile([80, 55, 70, 44, 33, 21, 65, 90, 12, 18], 55)

What is the difference between Stacks vs Queues?

- A stack is a linear data structure in which elements can be inserted and deleted only from one side of the list, called the top.
    - Stack is a LIFO (last in first out) data structure.
    - We track of the last element present in the list with a pointer called top.
- Queue is a FIFO (first in first out) data structure.
    - Clements can be inserted only from one side of the list (the rear).
    - Elements can be deleted only from the front.

What is the difference between Linked lists and Arrays?

- The difference is how they allocate memory.
- A Linked List is an ordered collection of elements of the same type which are connected to each other using pointers.
    - The address of the new element's memory location is stored in the previous node of the linked list, forming a link between the two nodes/elements.
    - Linked lists have a dynamic size, but random access isn't allowed.
- Whereas an array is a random-access data structure, where an array consumes contiguous memory locations allocated at compile time. An array has a fixed size, but random access is permissible via an index.

Structured Query Language (SQL)

How should you handle NULLs when querying a data set?

- In a relational database, null means that no entry has been made for that cell.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

- Either the values exist but is unknown, or there is no information about the existence of value.
- A null is not the same as 0 or blank. Databases such as SQL reserves the NULL keyword to denote an unknown or missing value.
- It is extremely important to handle null values when doing some arithmetic operations because if a null value is used in any of these operations, the answer always remains null which is hard to demystify.

What is the JOIN function in SQL?

- SQL handles queries across multiple tables with JOINs.
- JOINs are clauses in SQL statements that link two tables together, usually based on the common keys that define the relationship between those two tables.
- There are several types of JOINs:
    - INNER: It selects all rows from both tables that meet the required condition.
    - LEFT: This returns all the rows of the table on the left side of the join and matching rows for the table on the right side of join.
        - In no matches on the right side, the result will contain nulll.
    - RIGHT: This returns all the rows of the table on the right side of the join and matching rows for the table on the left side.
        - In no matches on the left side, the result will contain null.
    - FULL: It returns the combined result of both LEFT and RIGHT JOIN - it will contain all the rows from both tables.
        - In case of no matching, the result will contain null.

Select all customers who purchased at least two items on two separate days from Amazon.

```
SELECT Customer_ID,
COUNT(DISTINCT Item_ID) as ‘item’,
COUNT(DISTINCT Purchase_Date) as ‘date’
FROM Purchase_List
GROUP BY Customer_ID
HAVING ‘date’ >= 2 AND ‘item’ >= 2
```

What is the difference between DDL, DML, and DCL?

- DDL: Data Definition Language. It describes commands such as CREATE, DROP, ALTER, and TRUNCATE which can be applied on data.
- DML: Data Manipulation Language. It describes commands such as SELECT, INSERT, UPDATE, and DELETE.
- DCL: Data Control Language. It describes commands so you can GRANT or REVOKE access rights of someone over the database.

Why is Database Normalization Important?

- Database normalization is used to organize a database. The goals are:
    - All the data is stored in one place ensuring consistency
    - Removes duplicate records

- Minimizes data modification issues
- Querying the database is simplified

What is the difference between clustered and non-clustered indexes?

- The purpose of indexes is to speed-up query process in an SQL Server.
- Clustered Indexes:
  - A clustered index defines the order in which data is physically stored in a table.
  - Since the data can be sorted in only one order, there can be only one clustered index per table.
  - It is faster to read than non-clustered index as data is physically stored in index order.
- Non-Clustered Indexes:
  - A non-clustered index doesn't sort the data inside the table.
  - A non-clustered index is stored at one place and table data is stored in another place.
  - This allows for multiple non-clustered indexes per table. This method is quicker to insert and update operations (more writes than reads) than a clustered index.

Situational/Behavioral Questions

What was the most challenging project you have worked on so far? Can you explain your learning outcomes?

- It will be time consuming to create a well thought-out example. Keep the following points in mind:
  - Choose an appropriate example: Pick a project that's relevant to the responsibilities of the job.
  - Be Specific: Take the hiring manager through the process of the project. Break down the project into goals and milestones and explain how you were able to achieve those and describe your responsibilities. If you were managing a group project, make sure to mention about your communication and group management skills.
  - Explain Your Position Clearly: highlight the outcomes of the project and your role. Align your learnings with the aims of the company you're applying for. The hiring manager should know your challenges through the project phase and how you overcame them.

<omitted>

How do you avoid Selection Bias?

- Selection bias occurs during population sampling.
- It's defined as when a selected sample does not represent the characteristics of the population.
- The following are three types of selection bias:
  - Undercoverage: Happens when some members of the population are inadequately represented. This problem usually occurs while doing convenience sampling.
  - Voluntary Response Bias: Happens when members are self-selected volunteers who are strongly opinionated. The resulting sample tends to overrepresent these individuals.

101 Data Science Interview Questions & Answers (Amended) 2019-10-01

    - Nonresponse Bias: Happens when there's a significant difference between those who responded to the survey and those who did not. This may happen for a variety of reasons such as some people refused to participate or some people simply forgot to return the surveys.
- To avoid selection bias, use random sampling. The following are some of the choices for sampling:
    - Simple Random Sampling
    - Stratified Random Sampling

<omitted>