

Apple

1. Suppose you're given millions of users that each have hundreds of transactions and these millions of transactions are for tens of thousands of products. How would you group the users together in meaningful segments?

This is a user segmentation analysis. Since the number of users and items are so large, I would do dimension reduction first. Utilizing PCA to reduce the user attributes (items buy/not bouy) to around 50 dimension, then use that to train a k-means clustering model in order to group the users.

Microsoft

1. Describe a project you've worked on and how it made a difference.

The project I worked on using random forest to predict the medication adherence.

**2. How would you approach a categorical feature with high-cardinality?
(nominal variables with more than 100 distinct values "high-cardinality attributes)**

- transforming the nominal variable to a single continuous one, whose values are correlated with the target label
 - One-hot coding (creating k dummy variables)
 - Group together then one-hot coding (52 States -> Southeast, Southwest, Northeast, Northeast, midwest...)
3. What would you do to summarize a Twitter feed?
 - Tf-idf
 - Topic
 - Sentiment
 - Location
 - Based on the user
 4. What are the steps for wrangling and cleaning data before applying machine learning algorithms?
 - Checking data table summary
 - Check extreme values and outliers
 - Parse the date and other unstructured
 - 5. How do you measure distance between data points?**
 - L1/L2 distance
 - Categorical distance: entropy
 - Cosine similarity
 6. Define variance.
 - How far a set of (random) numbers are spread out from their mean.
 - $\text{Var}(X) = E((X-E(X))^2)$

7. Describe the differences between and use cases for box plots and histograms.

- A histogram is highly useful when wide variances or very little variance exist among the observed frequencies for a particular data set.
- Boxplot is useful to identify median, quartiles; there is moderate variation among the observed frequencies, which causes the histogram to look ragged and non-symmetrical due to the way the data is grouped.

Twitter

1. What features would you use to build a recommendation algorithm for users?
 - User-item table (collaborative filtering)
 - User attribute table
 - Item attribute table

Uber

1. Pick any product or app that you really like and describe how you would improve it.
 - Evernote: very useful in cross-platform note taking. I can improve it by adding Markdown features and more flexible online collaborative editing features.
2. **How would you find an anomaly in a distribution ?**
 - Scatter plot
 - Histogram
 - Statistical Anomaly Detection
 - Clustering-Based Anomaly Detection
 - Density-Based Anomaly Detection
 - Proximity-Based Anomaly Detection

http://trevorwhitney.com/data_mining/anomaly_detection

3. How would you go about investigating if a certain trend in a distribution is due to an anomaly?

- Use moving average
 - Use control chart
 - Comparative Analysis
 - Basic Statistical Analysis (Mean $\pm$ 3*SD)
4. How would you estimate the impact Uber has on traffic and driving conditions?
 - Flexible commute allow Uber cars transport more passengers than usual, therefore the need for private cars and other taxis will be reduced. As a result, the traffic will be a little bit better thanks to the new efficient way of transportation
 -
 5. What metrics would you consider using to track if Uber's paid advertising strategy to acquire new customers actually works? How would you then approach figuring out an ideal customer acquisition cost?

LinkedIn

1. Big Data Engineer Can you explain what **REST** is?
REST (Representational State Transfer)

REST is an architectural style for designing distributed systems. It is not a standard but a set of constraints, such as being stateless, having a client/server relationship, and a uniform interface.

Principles of REST

- Resources expose easily understood directory structure URIs.
- Representations transfer JSON or XML to represent data objects and attributes.
- Messages use HTTP methods explicitly (for example, GET, POST, PUT, and DELETE).
- Stateless interactions store no client context on the server between requests. State dependencies limit and restrict scalability. The client holds session state.

Machine Learning Questions

Google

1. Why do you use feature selection?
 - simplification of models to make them easier to interpret by researchers/users
 - shorter training times
 - to avoid the [curse of dimensionality](#)
 - enhanced generalization by reducing [overfitting](#)
2. What is the effect on the coefficients of logistic regression if **two predictors are highly correlated? What are the confidence intervals of the coefficients?**
 - The estimates for these highly correlated coefficients are unstable since $0.5x + 0.5y = .7x + .3y$. The model prediction should be influenced
 - **CI?**
3. What's the difference between Gaussian Mixture Model and K-Means?
 - K-means define hard clusters, the samples are to be associated to the groups (subpopulations).
 - Mixture models would let you determine these subpopulations, without associating each sample with a cluster.
4. How do you pick k for K-Means?
 - Start with $\sqrt{N/2}$
 - Visualization (scatter plot)
 - Plot within cluster sum of square against k
5. How do you know when **Gaussian Mixture Model** is applicable?
6. Assuming a clustering model's labels are known, how do you evaluate the performance of the model?
 - **Misclassification rate?**

Microsoft

1. What's an example of a machine learning project you're proud of?
2. Choose any machine learning algorithm and describe it.

3. Describe how Gradient Boosting works.
4. Data Mining Describe the decision tree model.
5. Data Mining What is a neural network?
6. Explain the Bias-Variance Tradeoff
7. How do you deal with unbalanced binary classification?
8. What's the difference between L1 and L2 regularization?

Uber

1. What sort features could you give an Uber driver to predict if they will accept a ride request or not? What supervised learning algorithm would you use to solve the problem and how would compare the results of the algorithm?

LinkedIn

1. Name and describe three different kernel functions and in what situation you would use each.
2. Describe a method used in machine learning.
3. How do you deal with sparse data?

IBM

1. How do you prevent overfitting?
 - Add more data
 - Limit the model complexity (Regularization : Lasso,)
 - Cross-validation
2. How do you deal with outliers in your data?
 - Delete outliers
 - Build a different model for outliers
 -
3. How do you analyze the performance of the predictions generated by regression models versus classification models?
 - Regression model: MSE
 - Classification: Log-loss, one-other accuracy/precision/recall/AUC
4. How do you assess logistic regression versus simple linear regression models?
5. What's the difference between supervised learning and unsupervised learning?
6. What is cross-validation and why would you use it?
7. What's the name of the matrix used to evaluate predictive models?
8. What relationships exist between a logistic regression's coefficient and the Odds Ratio?
9. What's the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)

10. If you had a categorical dependent variable and a mixture of categorical and continuous independent variables, what algorithms, methods, or tools would you use for analysis?
11. Business Analytics What's the difference between logistic and linear regression? How do you avoid local minima?

Salesforce

1. What data and models would you use to measure attrition/churn? How would you measure the performance of your models?
2. Explain a machine learning algorithm as if you're talking to a non-technical person.

Capital One

1. How would you build a model to predict credit card fraud?
2. How do you handle missing or bad data?
3. How would you derive new features from features that already exist?
4. If you're attempting to predict a customer's gender, and you only have 100 data points, what problems could arise?
5. Suppose you were given two years of transaction history. What features would you use to predict credit risk?
6. Design an AI program for Tic-tac-toe

Zillow

1. Explain overfitting and what steps you can take to prevent it.
 - Overfitting means the model fit the random noise instead of the true underlying relationship
2. Why does SVM need to maximize the margin between support vectors?

Hadoop

Twitter

1. How would you use Map/Reduce to split a very large graph into smaller pieces and parallelize the computation of edges according to the fast/dynamic change of data?
2. Data Engineer Given a list of followers in the format: 123, 345234, 678345, 123... Where column one is the ID of the follower and column two is the ID of the followee. Find all mutual following pairs (the pair 123, 345 in the example above). How would you use

Map/Reduce to solve the problem when the list does not fit in memory?

In-memory

```
def follow(lst):
    res = []
    dct = {}
    for i,j in lst:
        if tuple(sorted([i,j])) in dct and dct[tuple(sorted([i,j]))] != (i,j):
            res.append((i,j))
        else:
            dct[tuple(sorted([i,j]))] = (i,j)
    return res
```

Map-Reduce

```
from mrjob.job import MRJob

class find_pair(MRJob):

    def mapper(self,_,line):
        users = line.split(',')
        yield tuple(sorted(users)), users == sorted(users)

    def reducer(self, key, values):
        if len(set(values)) == 2:
            yield key,1

if __name__ == '__main__':
    find_pair.run()
```

Capital One

1. Data Engineer What is Hadoop serialization?

Serialization is the process of converting structured objects into a byte stream. It is done basically for two purposes one, for transmission over a network(interprocess communication) and for writing to persistent storage.

2. Explain a simple Map/Reduce problem.

Hive

LinkedIn

1. Data Engineer Write a Hive UDF that returns a sentiment score. For example, if good = 1, bad = -1, and average = 0, then a review of a restaurant states "Good food, bad service," your score might be $1 - 1 = 0$.

Spark

Capital One

1. Data Engineer Explain how RDDs work with Scala in Spark

Resilient Distributed Datasets (RDD) is a simple and immutable distributed collection of objects. Each RDD is split into multiple partitions which may be computed on different nodes of the cluster. In spark all function are performed on RDDs only.

RDDs can be created in two different ways:

- Referencing an external dataset in an external storage system, such as a shared file system, HDFS, HBase, or any data source offering a Hadoop Input Format.
- By parallelizing a collection of objects(a list or a set) in the driver program.

Life cycle of a Spark program:

1. Some input RDDs are created from external data or by parallelizing the collection of objects in the driver program.
2. These RDDs are lazily transformed into new RDDs using transformations like filter() or map().
3. Spark caches any intermediate RDDs that will be needed to reused.
4. Actions such as count() and collect are launched to kick off a parallel computation which is then optimized and then executed by Spark.

Statistics & Probability Questions

Google

1. Explain Cross-validation as if you're talking to a non-technical person.
2. Describe a non-normal probability distribution and how to apply it.

Microsoft

1. Data Mining Explain what heteroskedasticity is and how to solve it

Twitter

1. Given Twitter user data, how would you measure engagement?

Uber

1. What are some different Time Series forecasting techniques?
2. Explain Principle Component Analysis (PCA) and equations PCA uses.
3. How do you solve Multicollinearity?
4. Analyst Write an equation that would optimize the ad spend between Twitter and Facebook.

Facebook

1. What's the probability you'll draw two cards of the same suite from a single deck?

$$52 \times 12 / 52 \times 51 = 12 / 51$$

IBM

1. What are p-values and confidence intervals?

Capital One

1. Data Analyst If you have 70 red marbles, and the ratio of green to red marbles is 2 to 7, how many green marbles are there?
2. What would the distribution of daily commutes in New York City look like?
3. Given a die, would it be more likely to get a single 6 in six rolls, at least two 6s in twelve rolls, or at least one-hundred 6s in six-hundred rolls?

PayPal

1. What's the Central Limit Theorem, and how do you prove it? What are its applications?

Programming & Algorithms

Google

1. Data Analyst Write a program that can determine the height of an arbitrary binary tree

Microsoft

1. Create a function that checks if a word is a palindrome.

Twitter

1. Build a power set.
2. How do you find the median of a very large dataset?

Uber

1. Data Engineer Code a function that calculates the square root (2-point precision) of a given number. Follow up: Avoid redundant calculations by now optimizing your function with a caching mechanism.

Facebook

1. Suppose you're given two binary strings, write a function adds them together without using any builtin string-to-int conversion or parsing tools. For example, if you give your function binary strings 100 and 111, it should return 1011. What's the space and time complexity of your solution?
2. Write a function that accepts two already sorted lists and returns their union in a sorted list.

LinkedIn

1. Data Engineer Write some code that will determine if brackets in a string are balanced
2. How do you find the second largest element in a Binary Search Tree?
3. Write a function that takes two sorted vectors and returns a single sorted vector.
4. If you have an incoming stream of numbers, how would you find the most frequent numbers on-the-fly?
5. Write a function that raises one number to another number, i.e. the pow() function.
6. Split a large string into valid words and store them in a dictionary. If the string cannot be split, return false. What's your solution's complexity?

Salesforce

1. What's the computational complexity of finding a document's most frequently used words?
2. **If you're given 10 TBs of unstructured customer data, how would you go about finding extracting valuable information from it?**

Capital One

1. **Data Engineer** How would you 'disjoin' two arrays (like JOIN for SQL, but the opposite)?
2. Create a function that does addition where the numbers are represented as two linked lists.

```
def addTwoNumbers(self, l1, l2):
    """
    :type l1: ListNode
    :type l2: ListNode
    :rtype: ListNode
    """
    dummy = ListNode(0)
    current, carry = dummy, 0

    while l1 or l2:
        val = carry
        if l1:
            val += l1.val
            l1 = l1.next
        if l2:
            val += l2.val
            l2 = l2.next
        carry, val = val / 10, val % 10
        current.next = ListNode(val)
        current = current.next

    if carry == 1:
        current.next = ListNode(1)

    return dummy.next
```

3. Create a function that calculates matrix sums.

```
for i in range(len(X)):
    # iterate through columns
    for j in range(len(X[0])):
        result[i][j] = X[i][j] + Y[i][j]
```

4. How would you use Python to read a very large tab-delimited file of numbers to count the frequency of each number?

```

from collections import Counter
def count_num(f):
    cnt = Counter()
    with open(f,'r'):
        for line in f:
            for num in line.split("\t"):
                cnt[num] += 1
    return cnt

```

PayPal

1. Write a function that takes a sentence and prints out the same sentence with each word backwards in O(n) time.

```

def print_backward(sentence):
    return ' '.join([word[::-1] for word in sentence[:-1].split(' ')]) + sentence[-1]

```

```

def print_backward(sentence):
    start = 0
    res = []
    for i in xrange(len(sentence[:-1])):
        if sentence[i] == ' ':
            res.append(sentence[start:i][::-1])
            start = i+1
    res.append(sentence[start:i][::-1])
    return ' '.join(res) + sentence[-1]

```

```

print_backward('I like apple pie.')

```

2. Write a function that takes an array, splits the array into every possible set of two arrays, and prints out the max differences between the two array's minima in O(n) time.

```

def max_diff(lst):
    max_num = -1<<31
    min_num = 1<<31 -1
    for num in lst:
        max_num = max(max_num,num)
        min_num = min(min_num,num)
    return max_num - min_num

```

3. Write a program that does merge sort.

```

def merge_sort(arr):
    if len(arr)<=1:
        return arr

```

```

mid = len(arr)/2
left = merge_sort(arr[:mid])
right = merge_sort(arr[mid:])
return merge(left,right)

```

```

def merge(left,right):
    if not left:
        return right
    if not right:
        return left
    if left[0]<right[0]:
        return [left[0]] + merge(left[1:],right)
    else:
        return [right[0]] + merge(left,right[1:])

```

SQL Questions

Microsoft

1. Data Analyst Define and explain the differences between clustered and non-clustered indexes.
2. Data Analyst What are the different ways to return the rowcount of a table?

Facebook

1. Data Engineer If you're given a raw data table, how would perform ETL (Extract, Transform, Load) with SQL to obtain the data in a desired format?
2. How would you write a SQL query to compute a frequency table of a certain attribute involving two joins? What changes would you need to make if you want to ORDER BY or GROUP BY some attribute? What would you do to account for NULLS?

LinkedIn

1. Data Engineer How would you improve ETL (Extract, Transform, Load) throughput?

Brain Teasers & Word Problems

Google

1. Suppose you have ten bags of marbles with ten marbles in each bag. If one bag weighs differently than the other bags, and you could only perform a single weighing, how would you figure out which one is different?

Facebook

1. You are about to hop on a plane to Seattle and want to know if you should carry an umbrella. You call three friends of yours that live in Seattle and ask each, independently, if it's raining.

Each of your friends will tell you the truth $\frac{2}{3}$ of the time and mess with you by lying $\frac{1}{3}$ of the time. If all three friends answer “Yes, it’s raining,” what is the probability that it is actually raining in Seattle?

Uber

1. Imagine you are working with a hospital. Patients arrive at the hospital in a Poisson Distribution, and the doctors attend to the patients in a Uniform Distribution. Write a function or code block that outputs the patient’s average wait time and total number of patients that are attended to by doctors on a random day.

Facebook

1. Imagine there are three ants in each corner of an equilateral triangle, and each ant randomly picks a direction and starts traversing the edge of the triangle. What’s the probability that none of the ants collide? What about if there are N ants sitting in N corners of an equilateral polygon?

Total ways the ants can go = $2 * 2 * 2$

Ways in which they will not collide = 2 (all clockwise or all anti)

Probability they don’t collide = $\frac{1}{4} = 0.25$

Alternatively :

The ants will not collide if they all go in the same direction

Probability they don’t collide = $P(\text{all go left}) + P(\text{all go right})$

$$= .5 * .5 * .5 + .5 * .5 * .5 = .125 + .125 = .25$$

2. How many trailing zeros are in 100 factorial (i.e. 100!)?

LinkedIn

1. Imagine you’re climbing a staircase that contains n stairs, and you can take any number k steps. How many distinct ways can you reach the top of the staircase? (This is a modification of the original stair step problem)