Analytics and Data Science Interview Questions - Tech

Saturday, November 21, 2020 10:28 AM

Machine Learning and Stats Questions:

1. What is the difference between boosting and bagging?
2. Explain SVM. Explain about C value in SVM
3. Explain Hypothesis Testing to a newbie
4. When to use log transformation?
5. When to use cosine distance and when to use Euclidean,manhatten,jaccard?
6. What is overfitting and underfitting
7. Difference in supervised and unsupervised learning
8. How will you use VIF to remove multicollinearity in data? What values of VIF would you use to remove?
9. What other ways can you use to remove multicollinearity in data?
10. Difference Between ridge and lasso? What other ways can you reduce overfitting? - more data points
11. Explain difference time series analysis techniques
12. What is LSTM? When is it useful and have you used it?
13. What was the most challenging part of your project? How was the model useful to marketing team? How did you identify most important variables in the model?
    What changes would you make if given a chance to your model?
14. What is central limit theorem and why is it important?
15. What is sampling and how many sampling methods do you know?
16. Difference in type 1 and type 2 error
17. Explain Linear Regression? What do terms p-value, coefficient and r-squared value mean and what is their significance?
18. What are the assumptions of Linear Regression : http://r-statistics.co/Assumptions-of-Linear-Regression.html
19. What is selection bias?
20. What is naïve bayes?
21. What is normal distribution?
22. Explain p-value in layman terms
23. Explain Hierarchical clustering
24. How does dropout work?
25. How would you improve a classification model that suffers from low precision? - increase probability threshold (https://www.kdnuggets.com/2016/12/4-reasons-machine-learning-model-wrong.html)
26. Why do we have L1 and L2 regularization but not L0 and L4? - No intuitive meaning : https://stats.stackexchange.com/questions/269298/why-do-we-only-see-l-1-and-l-2-regularization-but-not-other-norms/269407
27. How do you handle class imbalance?
28. Imp Statistics Questions : https://towardsdatascience.com/40-statistics-interview-problems-and-answers-for-data-scientists-6971a02b7eee
29. Imp Data Science Questions : https://towardsdatascience.com/amazon-data-scientist-interview-practice-problems-15b9b86e86c6
30. *How do you find thresholds for a classifier?*
31. *What's the difference between logistic regression and support vector machines? What's an example of a situation where you would use one over the other?*
32. *What is the relationship between PCA with a polynomial kernel and a single layer autoencoder? What if it is a deep autoencoder?*
33. *What is "random" in random forest? If you use logistic regression instead of a decision tree in random forest, how will your results change? - Logistic Regression gives Linear boundary, RF gives non-linear boundary*
34. *What is the interpretation of an ROC area under the curve as an integral?*
35. Let's say you have a categorical variable with thousands of distinct values, how would you encode it? (https://www.interviewquery.com/blog-amazon-machine-learning-interview-questions-solutions)
36. How does K-means work? What kind of distance metric would you choose? What if different features have different dynamic range?
37. What are generative and discriminative algorithms? What are their strengths and weaknesses? Which type of algorithms are usually used and why?
38. How does a logistic regression model know what the coefficients are?
39. Difference between convex and non-convex cost function; what does it mean when a cost function is non-convex?
40. Is random weight assignment better than assigning same weights to the units in the hidden layer?
41. Given a bar plot and imagine you are pouring water from the top, how to qualify how much water can be kept in the bar chart?
42. Why is gradient checking important?
43. Describe Tree, SVM, Random forest and boosting. Talk about their advantage and disadvantages.
44. How do you weight 9 marbles three times on a balance scale to select the heaviest one?
45. Describe the criterion for a particular model selection. Why is dimension reduction important?
46. If you can build a perfect (100% accuracy) classification model to predict some customer behaviour, what will be the problem in application?
47. What's the difference between MLE and MAP inference?
48. How many topic modeling techniques do you know of?
49. How do you deal with sparse data?
50. What are some situations where a linear model fails?
51. Do you think 20 decision trees are better than a large one? Why?
52. 100 Data Science Questions : https://towardsdatascience.com/over-100-data-scientist-interview-questions-and-answers-c5a66186769a
53. *What's your favorite kernel function?*

54. *Why use a bias term in Linear Regression? Bias Term helps us to capture the base case. Eg, if we want to predict the sales based on advertising spend, it is highly unlikely that the sales will be zero if we do not spend on advertising. Possible reasons for this might include past advertising, existing customer relationships etc*
55. *Define sampling distribution and standard error.*
    Sampling distribution : Distribution of all possible sample estimates. It is normal in nature when samples are large (Central Limit Theorem)
    Standard Error : On average how far does an estimate move from True population mean
    Confidence Interval : sample estimate + margin of error
    Hypothesis Test : Tests if the sample statistic in the population is significantly larger than, smaller than or different from some hypothesized value
56. *Why do we find out confidence interval? - CI shows how much uncertainty there is with any particular statistic. It tells you how confident you can be that the results from a sample reflect what you would expect to find if it were possible to survey the entire population. More here*
57. *https://github.com/kojino/120-Data-Science-Interview-Questions*
58. *What is the beta coefficient of a multivariate regression? How do you derieve it and what is the non-closed form?*
59. *Can you explain how to interpret the confidence interval of a logistic regression model?*
60. What is hypothesis testing?
61. What is a null hypothesis?
62. What is a random sample?
63. What is a mean?
64. What is a standard deviation?
65. What is a p-value?
66. What is a confidence interval? - The band which tells us the uncertainty around a point estimate
67. How do you reject a null hypothesis?
68. How does sample size affect the p-value v/s the confidence interval? As n increases, SE decreases, thus Conf Interval Decreases. For the same reason the p-value will be a very small value now.
69. Suppose you have 2 colors to test on a landing page. How would you evaluate which color to use for the sign-up button? For an A/B testing, define an evaluation metric for experiments that aligns with its strategic goals
70. *How would you select a representative sample of search queries from six million?*
71. *Find the maximum of sub sequence in an integer list?*
72. *Give an example of a scenario where you would use Naive Bayes over another classifier?*
73. *How would you explain what MapReduce does as concise as possible?*
74. *What is the ROC curve and the meaning of sensitivity, specificity, confusion matrix?*
75. *The autocomplete feature: How would you implement it and can you highlight the flaws in this tool today?*
76. *Describe efficient ways to merge a given k sorted arrays of size n each.*

SQL Problems :

1. Create a new column by extracting last 2 characters of a column containing ID of 7 digit long and in string datatype
2. SQL use-case for practise : https://towardsdatascience.com/sql-case-study-investigating-a-drop-in-user-engagement-510b27d0cbcc
3. Find the cumulative sum of top 10 most profitable products of the last 6 month for customers in Seattle.
4. SQL problems on Mode : https://mode.com/sql-tutorial/a-drop-in-user-engagement/
5. Challenging SQL problem : https://www.youtube.com/watch?v=sJTa7HNFN2I , https://www.youtube.com/watch?v=1gziHPyvAAk
6. Provided a table with user_id and dates they visited the platform, find the top 100 users with the longest continuous streak of visiting the platform as of yesterday.
7. Given 2 tables, one with the phone numbers that Facebook sends the confirmation message to and another one with the phone numbers that confirmed the verification, write a SQL query to calculate the confirmation percentage.
8. Given a table containing date, post_id, relationship (e.g. Friend, Group, Page), interaction (like, share etc.), and a table containing poster id and post id, calculate: how many likes were made on friend posts yesterday.
9. Given a table with detailed customer complaint tickets of different types, calculate the share of processed tickets within each type.
10. Provided a table with page_id, event timestamp and a flag for a state (which is on/off), find the number of pages that are currently on.
11. Write an SQL query that makes recommendations using the pages that your friends liked. Assume you have two tables: a two-column table of users and their friends, and a two-column table of users and the pages they liked. It should not recommend pages you already like

Theoretical SQL :

1. When will ROW_NUMBER and RANK give different results? Give an example.

2. Is it possible for LEFT JOIN and FULL OUTER JOIN to produce the same results? Why or why not?

3. Why would I use DENSE_RANK instead of RANK? What about RANK instead of DENSE_RANK?

4. What happens if I GROUP BY a column that is not in the SELECT statement? Why does this happen?

5. LAG and LEAD are especially useful in what type of scenarios?

6. For dealing with NULL values, why would I choose to use IFNULL vs. CASE WHEN?

7. Do temp tables make your code cleaner and faster, one of the two, or none? Why?

8. When is a subquery a bad idea? A good idea?

Python

1. Right a python code for recognizing if entries to a list have same characters or not. Then what is the computation complexity of it?
2. *Given an Array of numbers & a target value, return indexes of two numbers such that their Absolute difference is equal to the target*
3. *Given two dates D1 & D2. count number of days, months?*
4. *Find 1st missing positive number (must do in O(1) memory & O(n) time)*
5. *Given an array a, return the indices i,j that minimize |a_i -a_j|*
6. *Write a function to sample from a multinomial distribution.*
7. *Given an array of words and a max width parameter, format the text such that each line has exactly X characters.*
8. *Write a query to randomly sample a row from a table with 100 million rows.*
9. *What's the probability that you roll at least two 3s when rolling three die?*

Use-cases :

1. Market-mix modelling in python : https://towardsdatascience.com/building-a-simple-marketing-mix-model-with-ols-571ac3d5b64f
2. How would the change of prime membership fee would affect the market?
3. When users are navigating through the Amazon website, they are performing several actions. What is the best way to model if their next action would be a purchase?
4. *Due to engineering constraints, the company can't AB test a feature before launching it. How would you analyze how the feature is performing?*
5. *Kaggle questions*
6. *How do you figure out when a small uptick in sales is a fad? At what point should you consider it a trend? How does a trend take off?*
7. *We've all heard the tale that supermarkets have daily necessities at the back of the store to get customers to walk through the store and buy more things than they need. Is such a layout actually more profitable?*
8. *If you placed bread and butter next to each other, would you get higher sales than if you placed them on opposite shelves? What about in adjacent aisles?*
9. *If you want to trial a VR interactive experience in your retail stores, how do you pick the stores?*
10. *How do you derive more value out of flagging customer relationships?*
11. *How do you make an email newsletter more relevant?*
12. *How do you assess when an employee's expense receipts amount to an attempt to defraud the company? How do you curb such behaviour?*
13. *Is Black Friday an inconvenient tradition, or does it actually add any value to a customer-focused company?*
14. *Facebook news feed ranking algorithm? : https://www.geeksforgeeks.org/facebook-news-feed-algorithm/*
15. *Lyft rider cancel rates are up 5%. What could be the reason?*
16. *You launch a NEW product in a New region and how you predict it will grow/fail ?*
    a) If your product is not growing, how will find what are the factors impacting it?
    b) How will you compare it against the similar products?

List of preparatory material - solve sql questions(given at end) from here for practise
https://medium.com/better-programming/the-data-science-interview-study-guide-c3824cb76c2e

ML Glossary for Revision : https://ml-cheatsheet.readthedocs.io/en/latest/
Market Basket Analysis Revision : https://medium.com/swlh/a-tutorial-about-market-basket-analysis-in-python-predictive-hacks-497dc6e06b27

Web Analytics :
https://www.kaushik.net/avinash/impact-matrix-digital-analytics-framework/
https://www.kaushik.net/avinash/sitemap/
https://www.thinkwithgoogle.com/marketing-strategies/data-and-measurement/business-advertising-metrics/

ETL process info :

https://medium.com/hashmapinc/etl-understanding-it-and-effectively-using-it-f827a5b3e54d

SQL Fine-tuning
https://www.sisense.com/blog/8-ways-fine-tune-sql-queries-production-databases/