120 DATA SCIENCE INTERVIEW QUESTIONS

COMPILED AND CREATED BY:
CARL SHAN, MAX SONG, HENRY WANG, AND WILLIAM CHEN

INTRODUCTION

This guide is meant to bridge the gap between the knowledge of a recent graduate and the skillset required to become a data scientist. By reading this guide and learning how to answer these questions, recent graduates will equip themselves with the expected knowledge and skills of a data scientist.

To help readers with these goals, we've gathered 120 interview questions in product metrics, programming and databases, probability, experimentation and inference, data analysis, and predictive modeling. These questions are all either real data science interview questions or inspired by real data science interview questions, and should help readers develop the skills needed to succeed in a data science role.

The role of a data scientist is highly malleable and company dependent. However, the general skillset needed is similar. Candidates need:

- Technical skills - data analysis and programming
- Business/product intuition - metrics and identifying opportunities for impact
- Communication ability - clarity in explaining findings and insights

To prepare for your interview, you may want to brush up by reviewing some probability, data analysis, SQL, coding, and experimental design. The questions in this guide should help you do so. The background of data science applicants varies wildly, so interviews may generally be more holistic and test your intuition, analytic, and communication abilities rather than focusing on specific technical concepts.

Prepare to discuss your past work involving analyzing large and complicated datasets, defending your approaches and communicating what you learned during your project. Expect questions involving how to measure "goodness" of a feature on the company's product, and be sure to approach these problems in a scientific and principled way. You have a good chance of getting a product metrics or experimentation question based on some actual questions the company is tackling at this time.

Check up on your company's engineering / data blog and see if anything's relevant. Be familiar with A/B testing and common metrics that companies similar to the one you are interviewing for may use. Brush up on your Python (especially iPython notebook) and/or R abilities to prepare for a potential live data analysis problem.

And finally, of course, follow the general interview advice. Prepare to elaborate on related projects from your resume. Be enthusiastic. Share your thoughts with your interviewer as you're going through a problem or doing a piece of analysis. And be sure to answer the question!

You have our best wishes!
Carl, Max, Henry, and William

Please feel free to reach out to us with questions, comments and suggestions at www.datasciencehandbook.me

CONTENTS

PREDICTIVE MODELING

1. *(Given a Dataset)* Analyze this dataset and give me a model that can predict this response variable.

2. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

3. What are some ways I can make my model more robust to outliers?

4. What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?

5. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

6. What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What's the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)

7. What is regularization and where might it be helpful? What is an example of using regularization in a model?

8. Why might it be preferable to include fewer predictors over many?

9. Given training data on tweets and their retweets, how would you predict the number of retweets of a given tweet after 7 days after only observing 2 days worth of data?

10. How could you collect and analyze data to use social media to predict the weather?

PRO TIP

If asked to predict a response variable during your interview, you should favor simpler models that run quickly and which you can easily explain. If the task is specifically a predictive modeling task, you should try to do, or at least mention, cross-validation as it really is the golden standard to evaluate the quality of one's model. Talk about and justify your approach while you're doing it, and leave some time to plot and visualize the data.

PREDICTIVE MODELING

11 How would you construct a feed to show relevant content for a site that involves user interactions with items?

12 How would you design the people you may know feature on LinkedIn or Facebook?

13 How would you predict who someone may want to send a Snapchat or Gmail to?

14 How would you suggest to a franchise where to open a new store?

15 In a search engine, given partial data on what the user has typed, how would you predict the user's eventual search query?

16 Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

17 You're Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?

18 How would you build a model to predict a March Madness bracket?

19 You want to run a regression to predict the probability of a flight delay, but there are flights with delays of up to 12 hours that are really messing up your model. How can you address this?

PRO TIP

Variations on ordinary linear regression can help address some problems that come up working with read data. LASSO helps when you have too many predictors by favoring weights of zero. Ridge regression can help with reducing the variance of your weights and predictions by shrinking the weights to 0. Least absolute deviations or robust linear regression can help when you have outliers. Logistic regression is used for binary outcomes, and Poisson regression can be used to model count data.

PROGRAMMING

1 Write a function to calculate all possible assignment vectors of *2n* users, where *n* users are assigned to group 0 (control), and *n* users are assigned to group 1 (treatment).

2 Given a list of tweets, determine the top 10 most used hashtags.

3 Program an algorithm to find the best approximate solution to the knapsack problem[1] in a given time.

4 Program an algorithm to find the best approximate solution to the travelling salesman problem[2] in a given time.

5 You have a stream of data coming in of size *n*, but you don't know what *n* is ahead of time. Write an algorithm that will take a random sample of *k* elements. Can you write one that takes *O(k)* space?

6 Write an algorithm that can calculate the square root of a number.

7 Given a list of numbers, can you return the outliers?

8 When can parallelism make your algorithms run faster? When could it make your algorithms run slower?

9 What are the different types of joins? What are the differences between them?

10 Why might a join on a subquery be slow? How might you speed it up?

11 Describe the difference between primary keys and foreign keys in a SQL database.

PRO TIP

Traditional software engineering questions may show up in data science interviews. Expect those questions to be easier, less about systems, and more about your ability to manipulate data, read databases, and do simple programming tasks. Review your SQL and prepare to do common operations such as JOIN, GROUP BY, and COUNT. Review ways to manipulate data and strings (we suggest doing this in Python), so you can answer questions that involve sifting through numerical or string data.

1 See http://en.wikipedia.org/wiki/Knapsack_problem
2 See http://en.wikipedia.org/wiki/Travelling_salesman_problem

12 Given a `COURSES` table with columns `course_id` and `course_name`, a `FACULTY` table with columns `faculty_id` and `faculty_name`, and a `COURSE_FACULTY` table with columns `faculty_id` and `course_id`, how would you return a list of faculty who teach a course given the name of a course?

13 Given a `IMPRESSIONS` table with `ad_id`, `click` (an indicator that the ad was clicked), and `date`, write a SQL query that will tell me the click-through-rate of each ad by month.

14 Write a query that returns the name of each department and a count of the number of employees in each:

`EMPLOYEES` containing: `Emp_ID` (Primary key) and `Emp_Name`

`EMPLOYEE_DEPT` containing: `Emp_ID` (Foreign key) and `Dept_ID` (Foreign key)

`DEPTS` containing: `Dept_ID` (Primary key) and `Dept_Name`

PROBABILITY

1 Bobo the amoeba has a 25%, 25%, and 50% chance of producing 0, 1, or 2 offspring, respectively. Each of Bobo's descendants also have the same probabilities. What is the probability that Bobo's lineage dies out?

2 In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

3 How can you generate a random number between 1 - 7 with only a die?

4 How can you get a fair coin toss if someone hands you a coin that is weighted to come up heads more often than tails?

5 You have an 50-50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?

6 Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?

7 A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

8 You have a group of couples that decide to have children until they have their first girl, after which they stop having children. What is the expected gender ratio of the children that are born? What is the expected number of children each couple will have?

9 How many ways can you split 12 people into 3 teams of 4?

PRO TIP

Important concepts to review from an introductory probability class include the Law of Total Probability, Bayes' Rule, and Expectation. You can learn many of these topics (and important topics regarding hypothesis testing and inference) with intro-level courses in probability and inference.

10 Your hash function assigns each object to a number between 1:10, each with equal probability. With 10 objects, what is the probability of a hash collision? What is the expected number of hash collisions? What is the expected number of hashes that are unused.

11 You call 2 UberX's and 3 Lyfts. If the time that each takes to reach you is IID, what is the probability that all the Lyfts arrive first? What is the probability that all the UberX's arrive first?

12 I write a program should print out all the numbers from 1 to 300, but prints out Fizz instead if the number is divisible by 3, Buzz instead if the number is divisible by 5, and FizzBuzz if the number is divisible by 3 and 5. What is the total number of numbers that is either Fizzed, Buzzed, or FizzBuzzed?

13 On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Alice and Bob randomly pick adjectives, what is the probability that they form a match?

14 A lazy high school senior types up application and envelopes to *n* different colleges, but puts the applications randomly into the envelopes. What is the expected number of applications that went to the right college

15 Let's say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?

16 What's the expected number of coin flips until you get two heads in a row? What's the expected number of coin flips until you get two tails in a row?

PRO TIP

Many Bayes' Rule questions can be solved quickly with the odds form of Bayes Rule, which says that prior odds times likelihood ratio is the posterior odds. For problem 18, the prior odds is 999:1 and the likelihood ratio is 1/1024:1 (10 heads has a 1/1024 probability with a fair coin and a 1 probability with a biased coin), which means the posterior odds is about 1:1. For problem 19, the prior odds is 1:1 and the likelihood ratio is 1/4:9/16, so the posterior odds is 4:9.

17 Let's say we play a game where I keep flipping a coin until I get heads. If the first time I get heads is on the *n*th coin, then I pay you 2*n*-1 dollars. How much would you pay me to play this game?

18 You have two coins, one of which is fair and comes up heads with a probability 1/2, and the other which is biased and comes up heads with probability 3/4. You randomly pick coin and flip it twice, and get heads both times. What is the probability that you picked the fair coin?

19 You have a 0.1% chance of picking up a coin with both heads, and a 99.9% chance that you pick up a fair coin. You flip your coin and it comes up heads 10 times. What's the chance that you picked up the fair coin, given the information that you observed?

STATISTICAL INFERENCE

1. In an A/B test, how can you check if assignment to the various buckets was truly random?
2. What might be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?
3. What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?
4. What would be some issues if blogs decide to cover one of your experimental groups?
5. How would you conduct an A/B test on an opt-in feature?
6. How would you run an A/B test for many variants, say 20 or more?
7. How would you run an A/B test if the observations are extremely right-skewed?
8. I have two different experiments that both change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?
9. What is a p-value? What is the difference between type-1 and type-2 error?
10. You are AirBnB and you want to test the hypothesis that a greater number of photographs increases the chances that a buyer selects the listing. How would you test this hypothesis?
11. How would you design an experiment to determine the impact of latency on user engagement?
12. What is maximum likelihood estimation? Could there be any case where it doesn't exist?

PRO TIP

Proper A/B testing practices are often a common discussion, especially because it easily becomes more complicated than anticipated in practice. Multiple variants and metrics, simultaneous conflicting experiments, and improper randomization will complicate experiments. Most people do not have a formal academic background on experimental design.

STATISTICAL INFERENCE

13 What's the difference between a MAP, MOM, MLE estimator? In which cases would you want to use each?

14 What is a confidence interval and how do you interpret it?

15 What is unbiasedness as a property of an estimator? Is this always a desirable property when performing inference? What about in data analysis or predictive modeling?

PRO TIP

Important concepts to know include randomization, Simpson's paradox, and multiple comparisons. Advanced concepts to know that may impress interviewers includes alternatives to A/B testing (such as multi-armed bandit strategies), or alternatives to t-tests and z-tests (e.g. non-parametric methods, bootstrapping)

DATA ANALYSIS

1 *(Given a Dataset)* Analyze this dataset and tell me what you can learn from it.

2 What is R^2? What are some other metrics that could be better than R^2 and why?

3 What is the curse of dimensionality?

4 Is more data always better?

5 What are advantages of plotting your data before performing analysis?

6 How can you make sure that you don't analyze something that ends up meaningless?

7 What is the role of trial and error in data analysis? What is the the role of making a hypothesis before diving in?

8 How can you determine which features are the most important in your model?

9 How do you deal with some of your predictors being missing?

10 You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?

11 Let's say you're given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?

12 Now you have a feasible amount of predictors, but you're fairly sure that you don't need all of them. How would you perform feature selection on the dataset?

PRO TIP

Some concepts that are important in data analysis and common in the field, include overfitting, regression towards the mean, curse of dimensionality, importance of visualization, and inductive bias. These questions test your knowledge and experience with some of the hazards of blind data analysis and your ability to distinguish a significant result from a spurious one.

DATA ANALYSIS

13 Your linear regression didn't run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?

14 You run your regression on different subsets of your data, and find that in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

15 What is the main idea behind ensemble learning? If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse?

16 Given that you have wifi data in your office, how would you determine which rooms and areas are underutilized and overutilized?

17 How could you use GPS data from a car to determine the quality of a driver?

18 Given accelerometer, altitude, and fuel usage data from a car, how would you determine the optimum acceleration pattern to drive over hills?

19 Given position data of NBA players in a season's games, how would you evaluate a basketball player's defensive ability?

20 How would you quantify the influence of a Twitter user?

21 Given location data of golf balls in games, how would construct a model that can advise golfers where to aim?

22 You have 100 mathletes and 100 math problems. Each mathlete gets to choose 10 problems to solve. Given data on who got what problem correct, how would you rank the problems in terms of difficulty?

PRO TIP

If asked to analyze a dataset during the interview, the interviewer is looking to learn about your comfort with your statistical software and your ability to generate interesting insights in a short period of time. We recommend making visualizations first, to show that you know good practices, prevent future missteps, and identify possible transformations needed. Be sure to talk about your procedure and anticipate questions about your approach.

DATA ANALYSIS

23 You have 5000 people that rank 10 sushis in terms of saltiness. How would you aggregate this data to estimate the true saltiness rank in each sushi?

24 Given data on congressional bills and which congressional representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most liberal? Most republican? Most bipartisan?

25 How would you come up with an algorithm to detect plagiarism in online content?

26 You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters to include?

27 Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach this problem?

PRO TIP

Consider asking your interviewer how data scientists extract and wrangle data at the company, what tools the team uses to do its exploratory analysis, and how the company shares its findings internally. Most of the work is not the analysis. In fact, data scientists spend most of their time just getting, cleaning, and processing the data.

PRODUCT METRICS

1 What would be good metrics of success for an advertising-driven consumer product? (Buzzfeed, YouTube, Google Search, etc.) A service-driven consumer product? (Uber, Flickr, Venmo, etc.)

2 What would be good metrics of success for a productivity tool? (Evernote, Asana, Google Docs, etc.) A MOOC? (edX, Coursera, Udacity, etc.)

3 What would be good metrics of success for an e-commerce product? (Etsy, Groupon, Birchbox, etc.) A subscription product? (Netflix, Birchbox, Hulu, etc.) Premium subscriptions? (OKCupid, LinkedIn, Spotify, etc.)

4 What would be good metrics of success for a consumer product that relies heavily on engagement and interaction? (Snapchat, Pinterest, Facebook, etc.) A messaging product? (GroupMe, Hangouts, Snapchat, etc.)

5 What would be good metrics of success for a product that offered in-app purchases? (Zynga, Angry Birds, other gaming apps)

6 A certain metric is violating your expectations by going down or up more than you expect. How would you try to identify the cause of the change?

7 Growth for total number of tweets sent has been slow this month. What data would you look at to determine the cause of the problem?

8 You're a restaurant and are approached by Groupon to run a deal. What data would you ask from them in order to determine whether or not to do the deal?

9 You are tasked with improving the efficiency of a subway system. Where would you start?

10 Say you are working on Facebook News Feed. What would be some metrics that you think are important? How would you make the news each person gets more relevant?

PRO TIP

The best choices of engagement metrics are those that benefit both the company and the users while correlating highly with revenue. Pageviews and daily actives would be appropriate for an advertising-driven product, and metrics such as number of purchases or conversion rate would be appropriate for any product that sells services and other products.

PRODUCT METRICS

11 How would you measure the impact that sponsored stories on Facebook News Feed have on user engagement? How would you determine the optimum balance between sponsored stories and organic content on a user's News Feed?

12 You are on the data science team at Uber and you are asked to start thinking about surge pricing. What would be the objectives of such a product and how would you start looking into this?

13 Say that you are Netflix. How would you determine what original series you should invest in and create?

14 What kind of services would find churn (metric that tracks how many customers leave the service) helpful? How would you calculate churn?

15 Let's say that you're are scheduling content for a content provider on television. How would you determine the best times to schedule content?

PRO TIP

Interviewers are looking for candidates who have strong intuition about metrics for success. You should give many possible metrics, each a bit more specific than the previous. The interviewer may stop and ask you to elaborate or describe how you would collect or visualize the data. Prepare to justify why the metric is important, relevant, and measurable.

COMMUNICATION

1. Explain to me a technical concept related to the role that you're interviewing for.
2. Introduce me to something you're passionate about.
3. How would you explain an A/B test to an engineer with no statistics background? A linear regression?
4. How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean?
5. How would you explain to a group of senior executives why data is important?
6. Tell me about a data project that you've done with a team. What did you add to the group?
7. Tell me about a dataset that you've analyzed. What techniques did you find helpful and which ones didn't work?
8. What's your favorite algorithm? Can you explain it to me?
9. How could you help the generate public understanding towards the importance of using data to generate insights?
10. How would you convince a government agency to release their data in a publicly accessible API?
11. I'm a local business owner operating a small restaurant. Convince me to switch my advertising budget from print to internet.

PRO TIP

Interviews are about convincing the interviewer that you know what you're talking about. Naturally, you will gain more ability to do so with a better background in in the topics covered here. Practice teaching a concept, explaining one of your past projects, and discussing your techniques.

DATA SCIENCE HANDBOOK

Knowing and being able to answer these questions will help you succeed in the data science interview. But after landing that job, if you want to learn how to advance in your career as a data scientist, you should check out *The Data Science Handbook* — a curated collection of interviews containing advice and wisdom from some of top data scientists in the world.

You can get it at www.datasciencehandbook.me

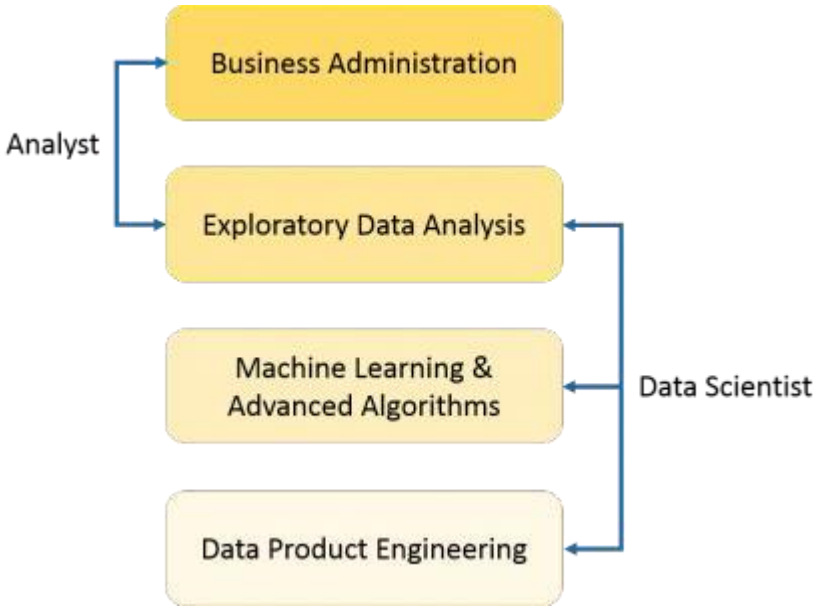
LAYOUT DESIGN BY: BRITTANY CHENG

Data Science Interview Questions

Q1. What is Data Science? List the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years?

The answer lies in the difference between explaining and predicting.



The differences between supervised and unsupervised learning are as follows;

Supervised Learning	Unsupervised Learning
Input data is labelled.	Input data is unlabelled.
Uses a training data set.	Uses the input data set.
Used for prediction.	Used for analysis.
Enables classification and regression.	Enables Classification, Density Estimation, & Dimension Reduction

Q2. What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

The types of selection bias include:

1. **Sampling bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
2. **Time interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.

3. **Data**: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
4. **Attrition**: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

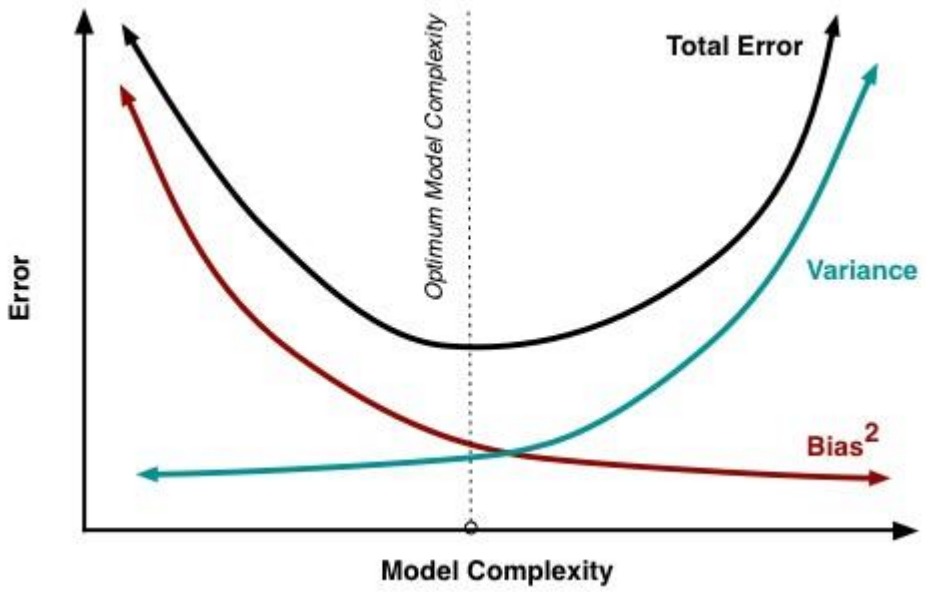
Q3. What is bias-variance trade-off?

Bias: Bias is an error introduced in your model due to oversimplification of the machine learning algorithm. It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.

Low bias machine learning algorithms — Decision Trees, k-NN and SVM High bias machine learning algorithms — Linear Regression, Logistic Regression

Variance: Variance is error introduced in your model due to complex machine learning algorithm, your model learns noise also from the training data set and performs badly on test data set. It can lead to high sensitivity and overfitting.

Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



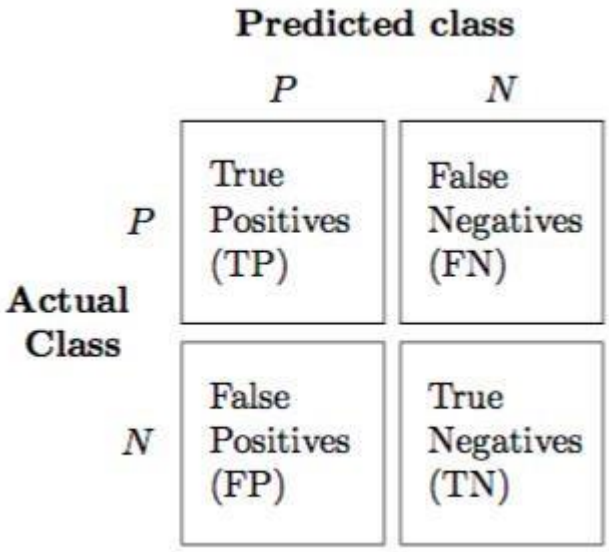
Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbour algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbours that contribute to the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

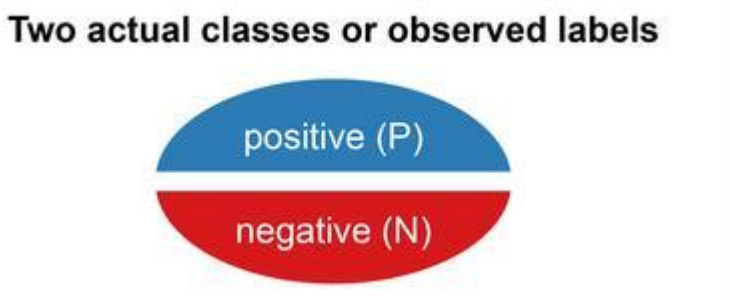
There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

Q4. What is a confusion matrix?

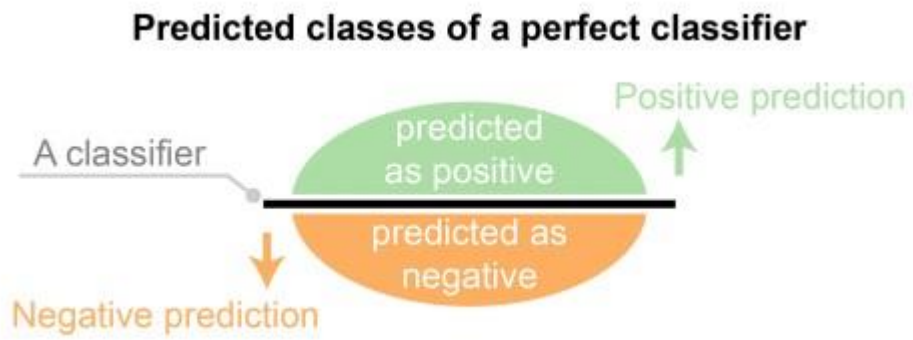
The confusion matrix is a 2X2 table that contains 4 outputs provided by the **binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are derived from it. *Confusion Matrix*



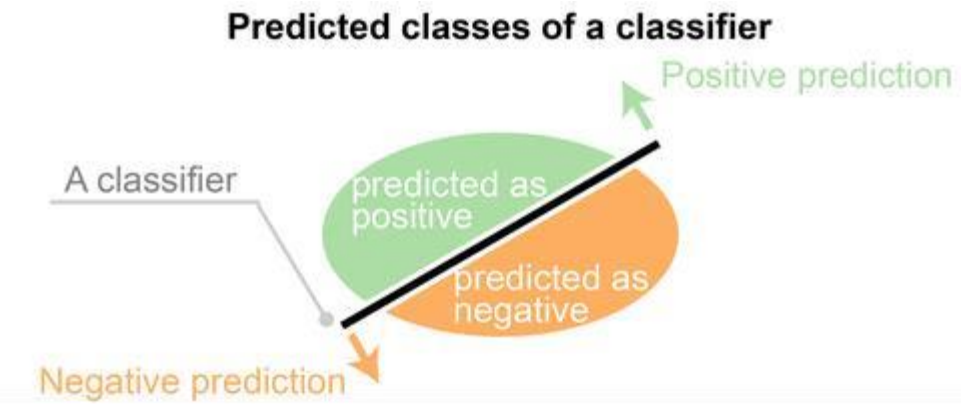
A data set used for performance evaluation is called a **test data set**. It should contain the correct labels and predicted labels.



The predicted labels will exactly the same if the performance of a binary classifier is perfect.

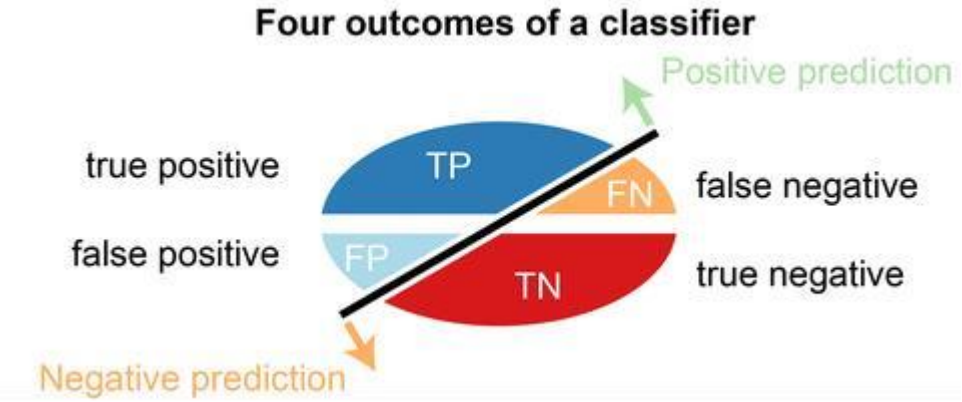


The predicted labels usually match with part of the observed labels in real-world scenarios.



A binary classifier predicts all data instances of a test data set as either positive or negative. This produces four outcomes-

1. True-positive(TP) — Correct positive prediction
2. False-positive(FP) — Incorrect positive prediction
3. True-negative(TN) — Correct negative prediction
4. False-negative(FN) — Incorrect negative prediction



Basic measures derived from the confusion matrix

1. Error Rate = (FP+FN)/(P+N)
2. Accuracy = (TP+TN)/(P+N)
3. Sensitivity(Recall or True positive rate) = TP/P
4. Specificity(True negative rate) = TN/N
5. Precision(Positive predicted value) = TP/(TP+FP)
6. F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/(b^2PREC+REC) where b is commonly 0.5, 1, 2.

STATISTICS INTERVIEW QUESTIONS

Q5. What is the difference between “long” and “wide” format data?

In the **wide-format**, a subject’s repeated responses will be in a single row, and each response is in a separate column. In the **long-format**, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups.

Name	Height	Weight
John	160	67
Christopher	182	78

Figure: *Wide Format*

Name	Attribute	Value
John	Height	160
John	Weight	67
Christopher	Height	182
Christopher	Weight	78

Figure: *Long Format*

Q6. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up.

However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

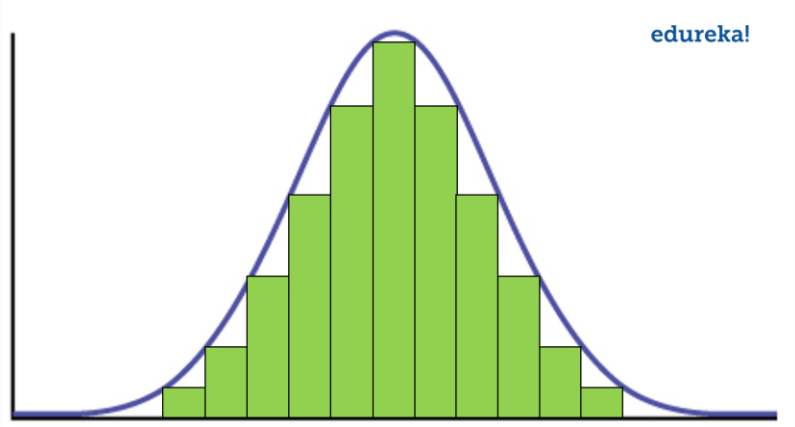


Figure: *Normal distribution in a bell curve*

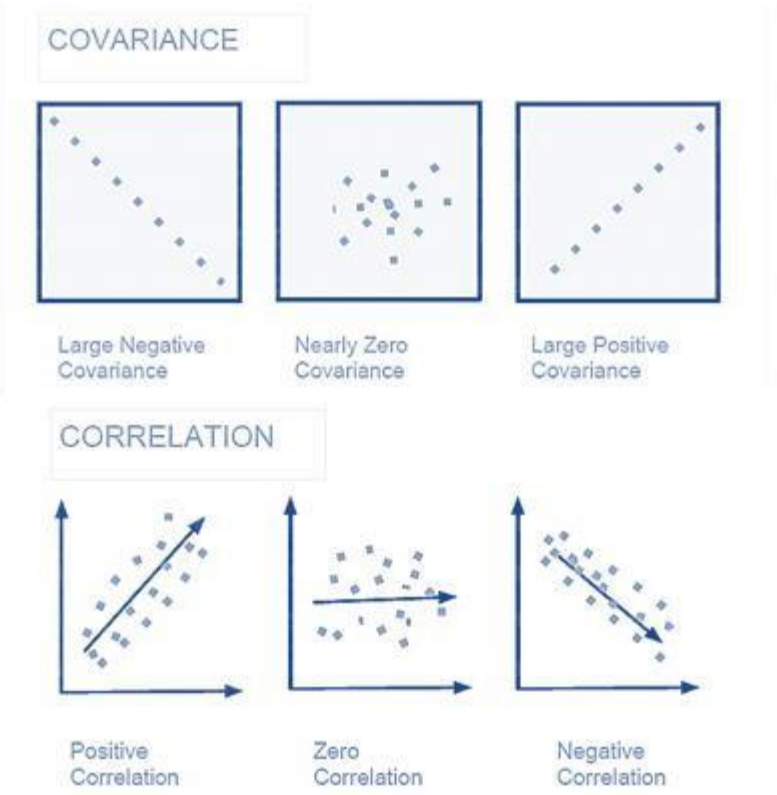
The random variables are distributed in the form of a symmetrical, bell-shaped curve.

Properties of Normal Distribution are as follows;

1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

Q7. What is correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.



Correlation:
Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

Q8. What is the difference between Point Estimates and Confidence Interval?

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by 1 — alpha, where alpha is the level of significance.

Q9. What is the goal of A/B Testing?

It is a hypothesis testing for a randomized experiment with two variables A and B.

The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search ads

An example of this could be identifying the click-through rate for a banner ad.

Q10. What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.

Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

Q11. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

Probability of not seeing any shooting star in 15 minutes is
= 1 – P(Seeing one shooting star)
= 1 – 0.2 = 0.8

Probability of not seeing any shooting star in the period of one hour

= (0.8) ^ 4 = 0.4096

Probability of seeing at least one shooting star in the one hour

= 1 – P(Not seeing any star)
= 1 – 0.4096 = 0.5904

Q12. How can you generate a random number between 1 – 7 with only a die?

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.
- To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.
- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.
- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

Q13. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

In the case of two children, there are 4 equally likely possibilities **BB,**

BG, GB and GG;

where **B** = Boy and **G** = Girl and the first letter denotes the first child.

From the question, we can exclude the first case of BB. Thus from the remaining 3 possibilities of **BG**, **GB** & **BB**, we have to find the probability of the case with two girls.

Thus, P(Having two girls given one girl) = **1 / 3**

Q14. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

Probability of selecting fair coin = 999/1000 = **0.999** Probability of selecting unfair coin = 1/1000 = **0.001**

Selecting 10 heads in a row = Selecting fair coin * Getting 10 heads + Selecting an unfair coin

P (A) = 0.999 * (1/2)^5 = 0.999 * (1/1024) = **0.000976**
P (B) = 0.001 * 1 = **0.001**
P(A / A + B) = 0.000976 / (0.000976 + 0.001) = **0.4939 P(B / A + B) =** 0.001 / 0.001976 = **0.5061**

Probability of selecting another head = P(A/A+B) * 0.5 + P(B/A+B) * 1 = 0.4939 * 0.5 + 0.5061 = **0.7531 Q15. What do you understand by statistical power of sensitivity and how do you calculate it?**

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).

Sensitivity is nothing but “Predicted True events/ Total events”. True events here are the events which were true and model also predicted them as true.

Calculation of seasonality is pretty straightforward.

Seasonality = (**True Positives**) / (**Positives in Actual Dependent Variable**)

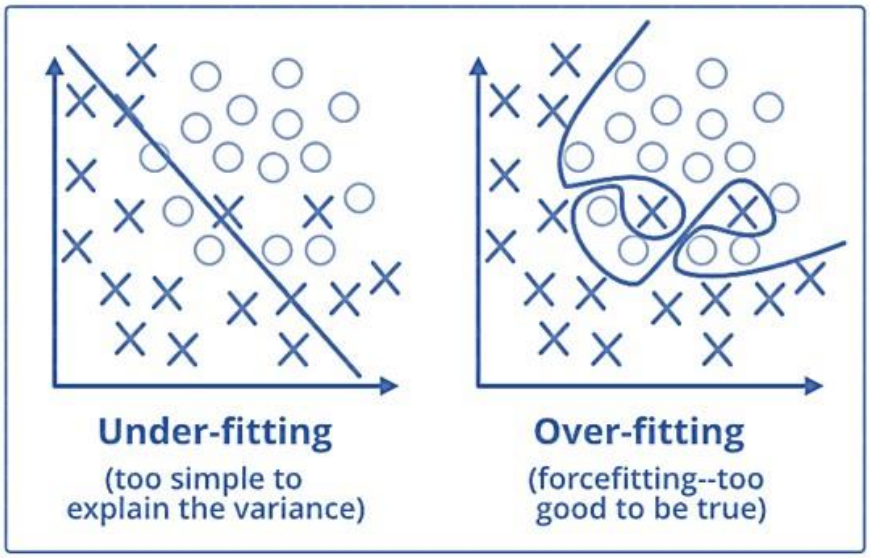
Q16. Why Is Re-sampling Done?

Resampling is done in any of these cases:

- Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
- Substituting labels on data points when performing significance tests
- Validating models by using random subsets (bootstrapping, cross-validation)

Q17. What are the differences between over-fitting and under-fitting?

In statistics and machine learning, one of the most common tasks is to fit a *model* to a set of training data, so as to be able to make reliable predictions on general untrained data.



In **overfitting**, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance. **Q18. How to combat Overfitting and Underfitting?**

To combat overfitting and underfitting, you can resample the data to estimate the model accuracy (k-fold cross-validation) and by having a validation dataset to evaluate the model.

Q19. What is regularisation? Why is it useful?

Regularisation is the process of adding tuning parameter to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant

is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

Q20. What Is the Law of Large Numbers?

It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of **frequency-style** thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate.

Q21. What Are Confounding Variables?

In statistics, a confounder is a variable that influences both the dependent variable and independent variable.

For example, if you are researching whether a lack of exercise leads to weight gain, **lack of exercise = independent variable weight gain = dependent variable.**

A confounding variable here would be any other variable that affects both of these variables, such as the **age of the subject**.

Q22. What Are the Types of Biases That Can Occur During Sampling?

- Selection bias
- Under coverage bias
- Survivorship bias

Q23. What is Survivorship Bias?

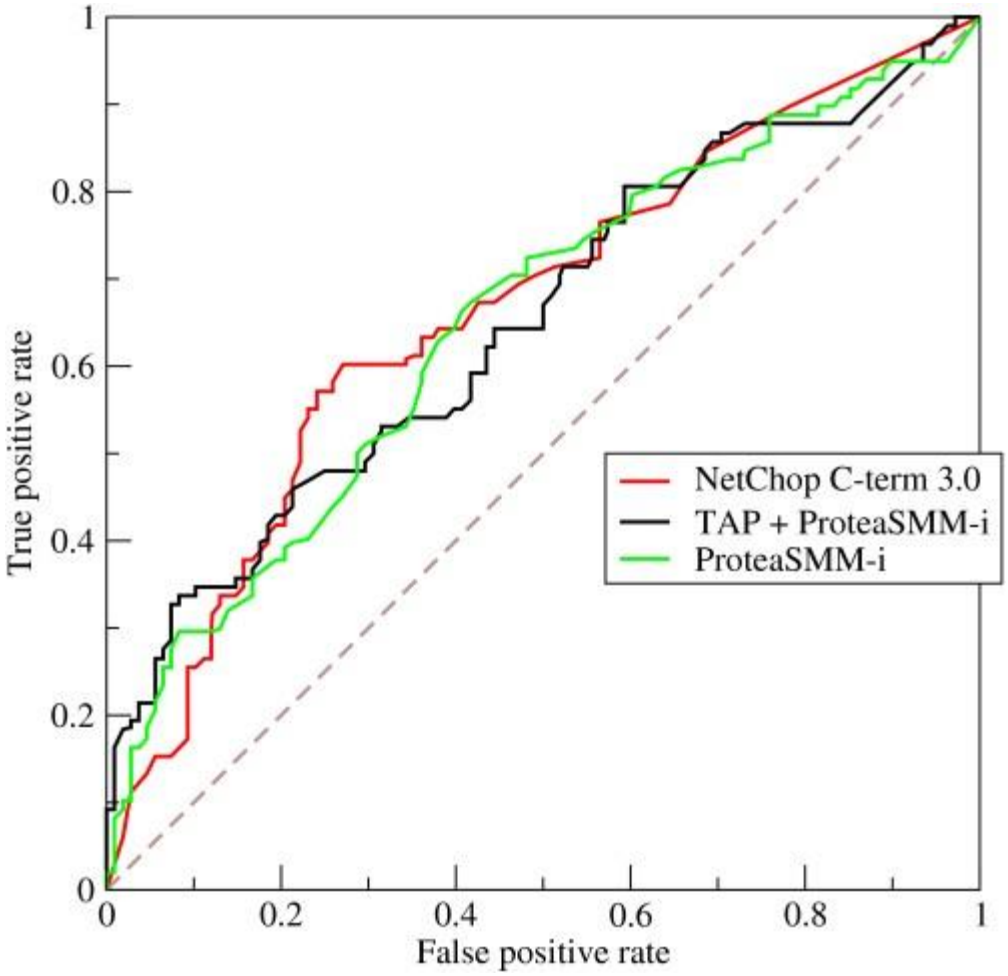
It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not work because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

Q24. What is selection Bias?

Selection bias occurs when the sample obtained is not representative of the population intended to be analysed.

Q25. Explain how a ROC curve works?

The **ROC** curve is a graphical representation of the contrast between true positive rates and false-positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity(true positive rate) and false-positive rate.



Q26. What is TF/IDF vectorization?

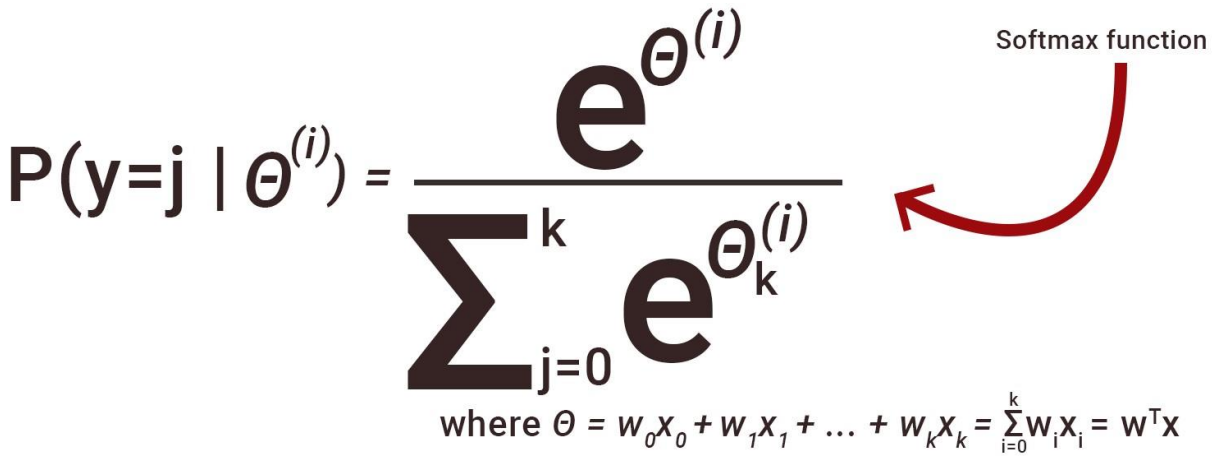
TF–IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

The TF–IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Q27. Why we generally use Softmax non-linearity function as last operation in-network?

It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let x be a vector of real numbers (positive, negative, whatever, there are no constraints).

Then the i'th component of Softmax(*x) is* —



It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.

DATA ANALYSIS INTERVIEW QUESTIONS

Q28. Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:

- ***Python*** would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.
- ***R*** is more suitable for machine learning than just text analysis.
- Python performs faster for all types of text analytics.

Q29. How does data cleaning plays a vital role in the analysis?

Data cleaning can help in analysis because:

- Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
- Data Cleaning helps to increase the accuracy of the model in machine learning.
- It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
- It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

Q30. Differentiate between univariate, bivariate and multivariate analysis.

Univariate *analyses* are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can the analysis can be referred to as univariate analysis.

The ***bivariate*** *analysis* attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

Multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

Q31. Explain Star Schema.

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

Q32. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Let's continue our Data Science Interview Questions blog with some more statistics questions.

Q33. What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

Q34. What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

Q35. Can you cite some examples where a false positive is important than a false negative?

Let us first understand what false positives and false negatives are.

- **False Positives** are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
- **False Negatives** are the cases where you wrongly classify events as non-events, a.k.a Type II error.

Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

Example 2: Let's say an e-commerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchase.

Q36. Can you cite some examples where a false negative important than a false positive?

Example 1**:** Assume there is an airport 'A' which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is being flagged as non-threat by airport model?

Example 2**:** What if Jury or judge decides to make a criminal go free?

Example 3**:** What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

Q37. Can you cite some examples where both false positive and false negatives are equally important?

In the **Banking** industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

Q38. Can you explain the difference between a Validation Set and a Test Set?

A ***Validation set*** can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.

On the other hand, a ***Test Set*** is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

Q39. Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will **generalize** to an **independent dataset**. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

MACHINE LEARNING INTERVIEW QUESTIONS

Q40. What is Machine Learning?

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics. Given below, is an image representing the various domains Machine Learning lends itself to.

Speech Recognition *Face Recognition* *Anti Virus* *Weather Prediction*

Q41. What is Supervised Learning?

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples.

Algorithms: Support Vector Machines, Regression, Naive Bayes, Decision Trees, K-nearest Neighbor Algorithm and Neural Networks

E.g. If you built a fruit classifier, the labels will be "this is an orange, this is an apple and this is a banana", based on showing the classifier examples of apples, oranges and bananas.

Q42. What is Unsupervised learning?

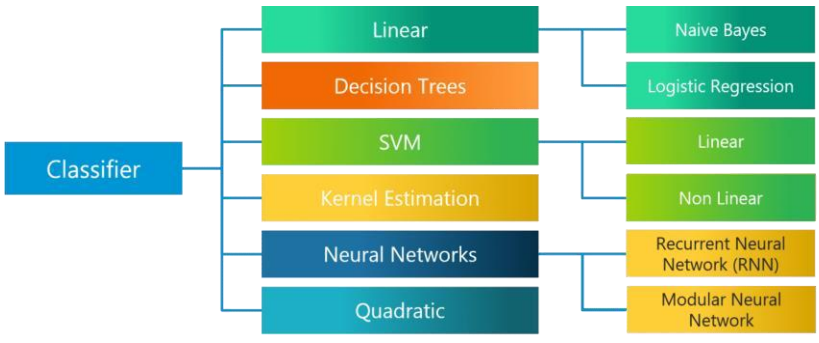
Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses.

Algorithms: Clustering, Anomaly Detection, Neural Networks and Latent Variable Models

E.g. In the same example, a fruit clustering will categorize as "fruits with soft skin and lots of dimples", "fruits with shiny hard skin" and "elongated yellow fruits".

Q43. What are the various classification algorithms?

The diagram lists the most important ***classification algorithms***.



Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

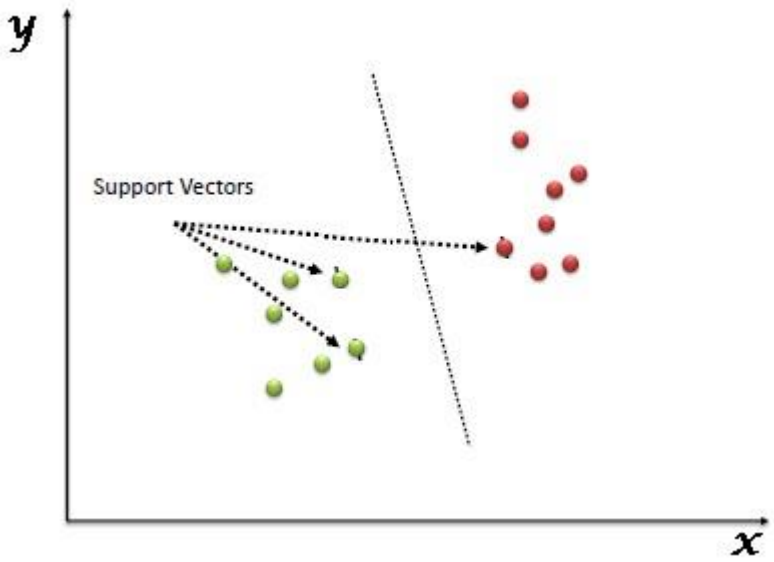
Q44. What is ‘Naive’ in a Naive Bayes?

The ***Naive Bayes Algorithm*** is based on the Bayes Theorem. Bayes’ theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

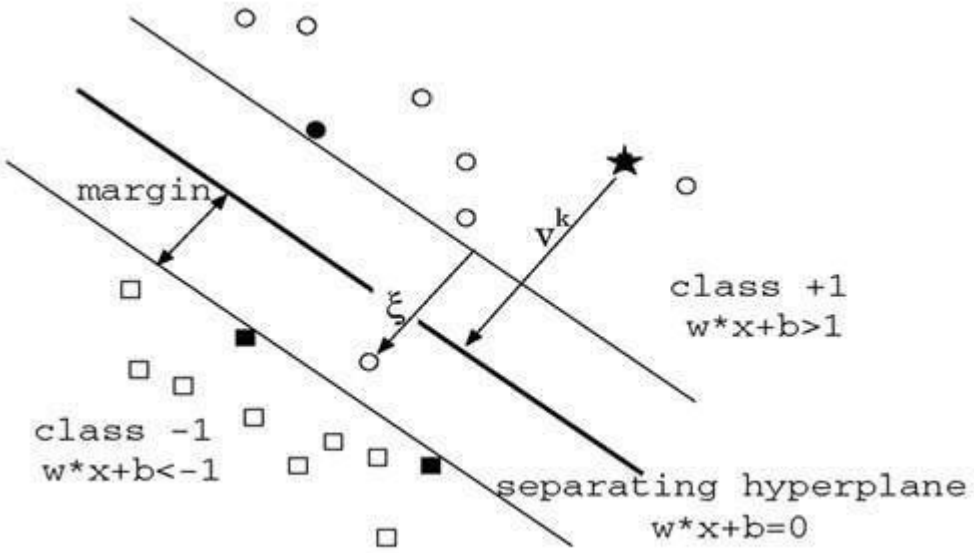
The Algorithm is ‘naive’ because it makes assumptions that may or may not turn out to be correct.

Q45. Explain SVM algorithm in detail.

SVM stands for support vector machine, it is a supervised machine learning algorithm which can be used for both ***Regression*** **and Classification**. If you have n features in your training data set, SVM tries to plot it in n-dimensional space with the value of each feature being the value of a particular coordinate. SVM uses hyperplanes to separate out different classes based on the provided kernel function.



Q46. What are the support vectors in SVM?



In the diagram, we see that the thinner lines mark the distance from the classifier to the closest data points called the support vectors (darkened data points). The distance between the two thin lines is called the margin.

Q47. What are the different kernels in SVM?

There are four types of kernels in SVM.

1. Linear Kernel
2. Polynomial kernel
3. Radial basis kernel
4. Sigmoid kernel

Q48. Explain Decision Tree algorithm in detail.

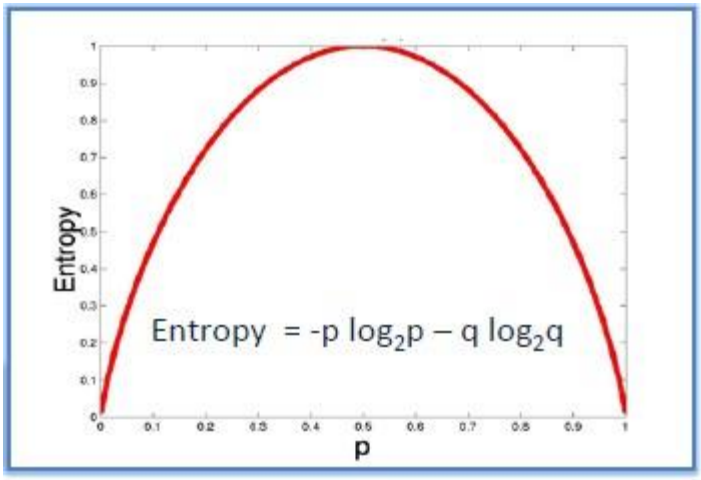
A ***decision tree*** is a supervised machine learning algorithm mainly used for **Regression and Classification**. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree can handle both categorical and numerical data.

Q49. What are Entropy and Information gain in Decision tree algorithm?

The core algorithm for building a decision tree is called **ID3**. **ID3** uses **Entropy** and **Information Gain**

Entropy

A decision tree is built top-down from a root node and involve partitioning of data into homogenious subsets. **ID3** uses enteropy to check the homogeneity of a sample. If the sample is completely homogenious then entropy is zero and if the sample is an equally divided it has entropy of one.



Entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

Information Gain

The **Information Gain** is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attributes that return the highest information gain.

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

Q50. What is pruning in Decision Tree?

Pruning is a technique in machine learning and search algorithms that reduces the size of **decision trees** by removing sections of the **tree** that provide little power to classify instances. So, when we remove sub-nodes of a decision node, this process is called ***pruning*** or opposite process of splitting.

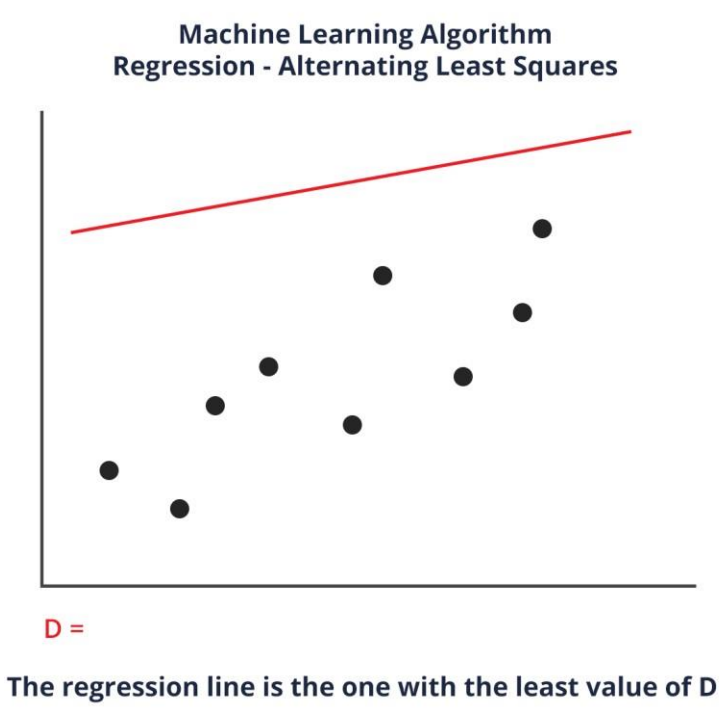
Q51. What is logistic regression? State an example when you have used logistic regression recently.

Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from a linear combination of predictor variables.

For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

Q52. What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.



Follow me for more AI and Data science posts: https://lnkd.in/gZu463X

Q53. What Are the Drawbacks of the Linear Model?

Some drawbacks of the linear model are:

- The assumption of linearity of the errors.
- It can't be used for count outcomes or binary outcomes
- There are overfitting problems that it can't solve

Q54. What is the difference between Regression and classification ML techniques?

Both Regression and classification machine learning techniques come under **Supervised machine learning algorithms**. In Supervised machine learning algorithm, we have to train the model using labelled data set,

While training we have to explicitly provide the correct labels and algorithm tries to learn the pattern from input to output. If our labels are discrete values then it will a classification problem, e.g A,B etc. but if our labels are continuous values then it will be a regression problem, e.g 1.23, 1.333 etc.

Q55. What are Recommender Systems?

Recommender Systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

Examples include movie recommenders in IMDB, Netflix & BookMyShow, product recommenders in ecommerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

Q56. What is Collaborative filtering?

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources and multiple agents.

Movie	Alice	Bob	Carol	Dave
Shutter Island	4	3	5	1
Fight Club	5	4	4	2
Dark Knight	5	3	4	?
21	4	3	?	5
Home Alone	4	4	5	5

Figure: *Predicting the rating of Dave for Dark Knight and Carol for 21 using Collaborative Filtering*

An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others' ratings for all movies. This concept is widely used in recommending movies in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

Q57. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values.

All extreme values are not outlier values. The most common ways to treat outlier values

1. To change the value and bring it within a range.
2. To just remove the value.

Q58. What are the various steps involved in an analytics project?

The following are the various ***steps involved in an analytics project***:

1. Understand the Business problem
2. Explore the data and become familiar with it.
3. Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.

4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.
5. Validate the model using a new data set.
6. Start implementing the model and track the result to analyze the performance of the model over the period of time.

Q59. During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

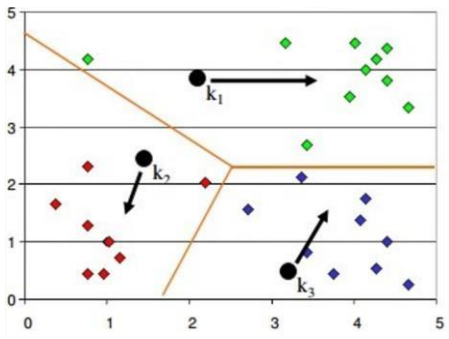
If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

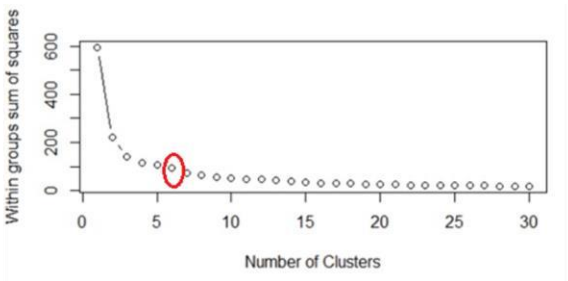
If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

Q60. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question is mostly in reference to ***K-Means clustering*** where “K” defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other. For example, the following image shows three different groups.



Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below.



- The Graph is generally known as ***Elbow Curve***.

- Red circled a point in above graph i.e. ***Number of Cluster =6*** is the point after which you don't see any decrement in WSS.
- This point is known as the **bending** point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendrograms and identify the distinct groups from there.

Q61. What is Ensemble Learning?

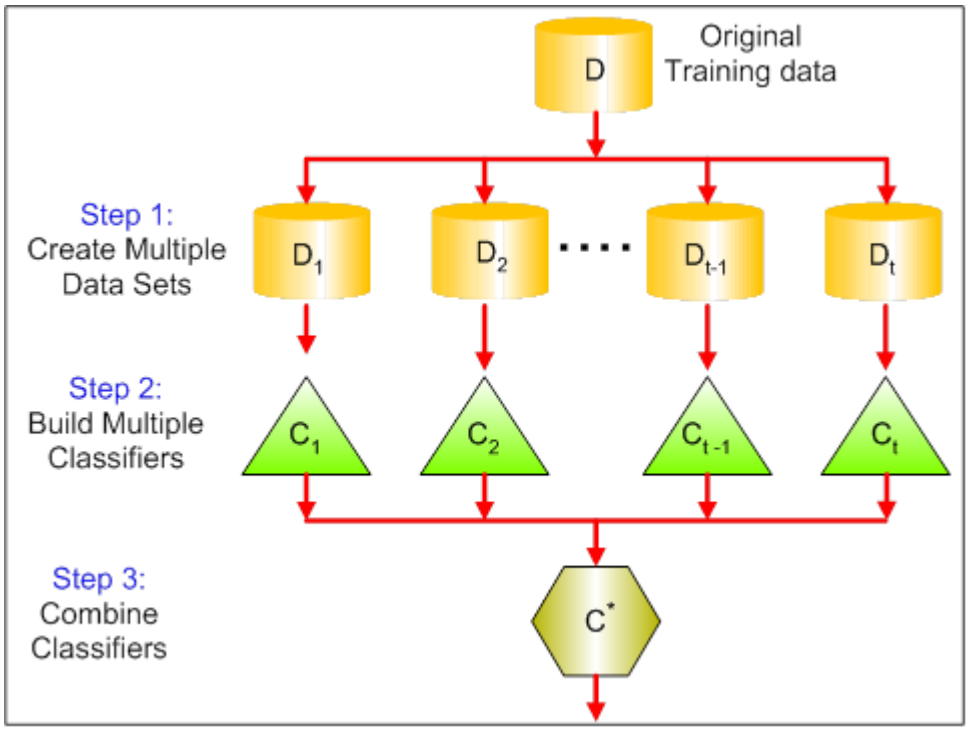
Ensemble Learning is basically combining a diverse set of learners(Individual models) together to improvise on the stability and predictive power of the model.

Q62. Describe in brief any type of Ensemble Learning?

Ensemble learning has many types but two more popular ensemble learning techniques are mentioned below.

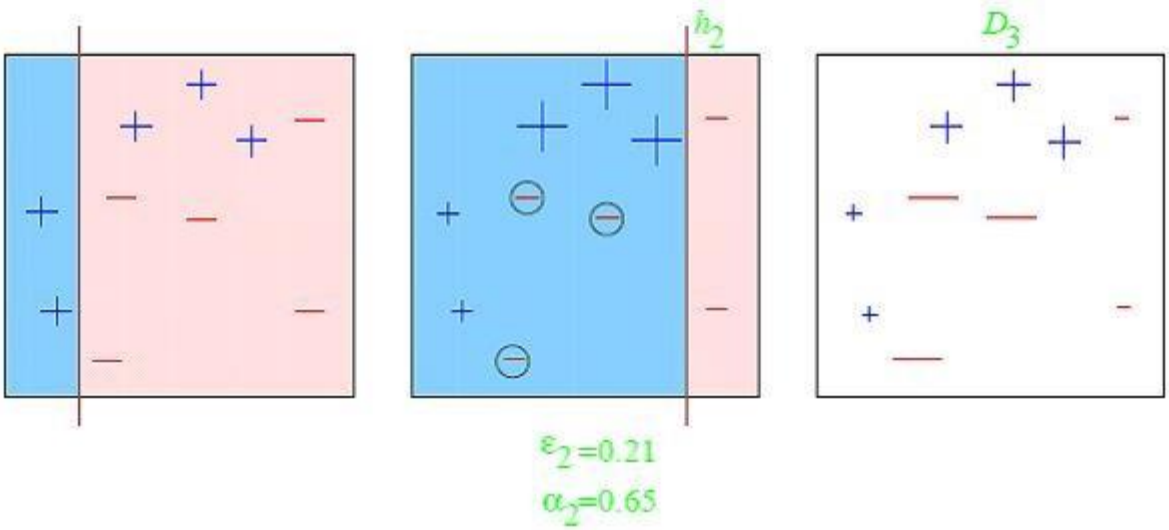
Bagging

Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions. In generalised bagging, you can use different learners on different population. As you expect this helps us to reduce the variance error.



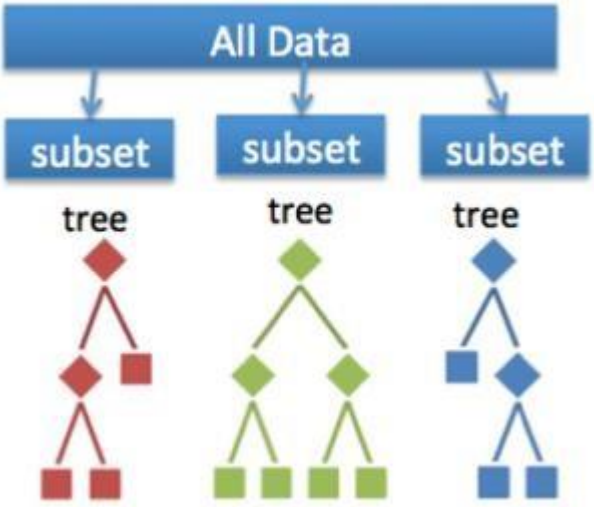
Boosting

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa. Boosting in general decreases the bias error and builds strong predictive models. However, they may over fit on the training data.



Q63. What is a Random Forest? How does it work?

Random forest is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for dimensionality reduction, treats missing values, outlier values. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.



In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most **votes**(Overall the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

Q64. How Do You Work Towards a Random Forest?

The underlying principle of this technique is that several weak learners combined to provide a keen learner. The steps involved are

- Build several decision trees on bootstrapped training samples of data
- On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors • Rule of thumb: At each split m=p√m=p
- Predictions: At the majority rule

Q65. What cross-validation technique would you use on a time series data set?

Instead of using k-fold cross-validation, you should be aware of the fact that a time series is not randomly distributed data — It is inherently ordered by chronological order.

In case of time series data, you should use techniques like forward=chaining — Where you will be model on past data then look at forward-facing data.

fold 1: training[1], test[2] fold

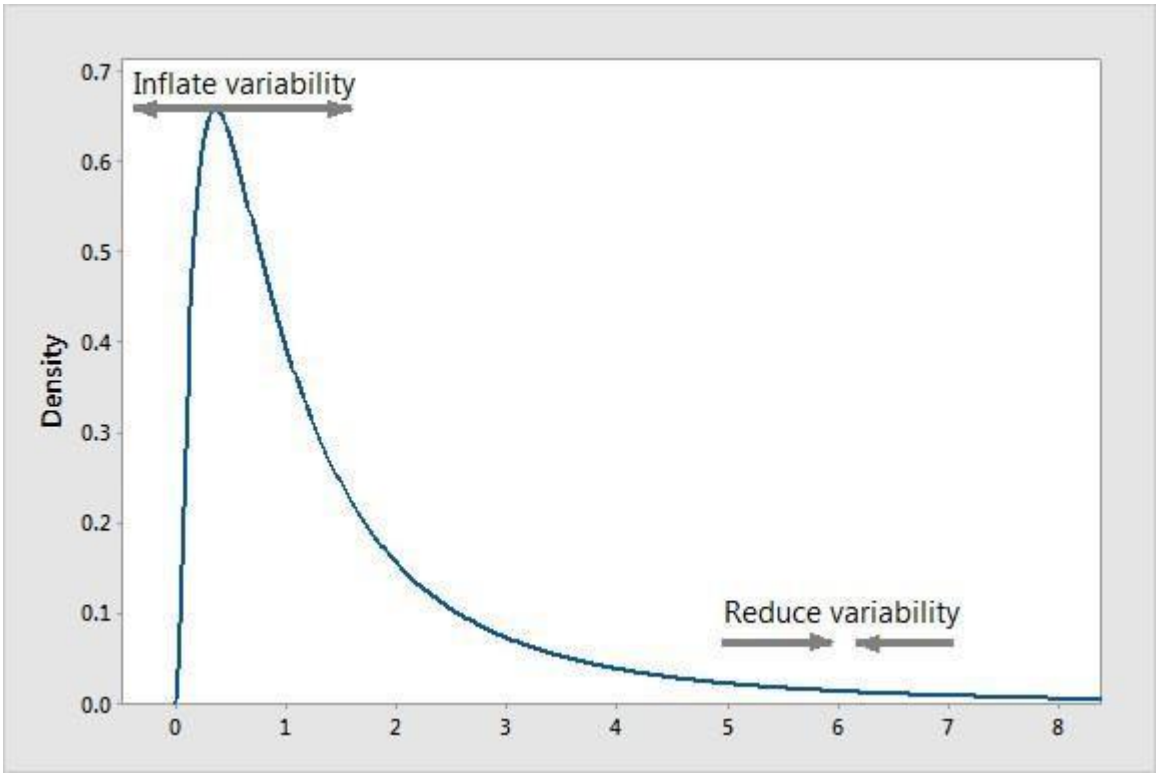
1: training[1 2], test[3] fold 1:

training[1 2 3], test[4] fold 1:

training[1 2 3 4], test[5]

Q66. What is a Box-Cox Transformation?

The dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow the skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.



A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is named after statisticians ***George Box*** and ***Sir David Roxbee Cox*** who collaborated on a 1964 paper and developed the technique.

Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

Q67. How Regularly Must an Algorithm be Updated?

You will want to update an algorithm when:

- You want the model to evolve as data streams through infrastructure
- The underlying data source is changing
- There is a case of non-stationarity
- The algorithm underperforms/ results lack accuracy

Q68. If you are having 4GB RAM in your machine and you want to train your model on 10GB data set. How would you go about this problem? Have you ever faced this kind of problem in your machine learning/data science experience so far?

First of all, you have to ask which ML model you want to train.

For Neural networks: Batch size with Numpy array will work.

Steps:

1. Load the whole data in the Numpy array. Numpy array has a property to create a mapping of the complete data set, it doesn't load complete data set in memory.
2. You can pass an index to Numpy array to get required data.
3. Use this data to pass to the Neural network.

4. Have a small batch size. **For SVM:** Partial fit will work

Steps:

1. Divide one big data set in small size data sets.
2. Use a partial fit method of SVM, it requires a subset of the complete data set.
3. Repeat step 2 for other subsets.

However, you could actually face such an issue in reality. So, you could check out the ***best laptop for Machine Learning*** to prevent that. Having said that, let's move on to some questions on deep learning.

DEEP LEARNING INTERVIEW QUESTIONS

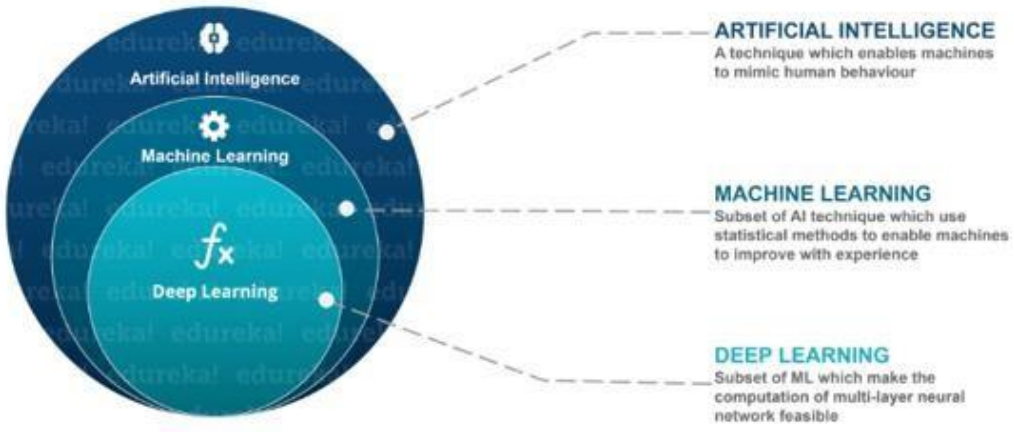
Q69. What do you mean by Deep Learning?

Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in recent years. This is because of the fact that Deep Learning shows a great analogy with the functioning of the human brain.

Q70. What is the difference between machine learning and deep learning?

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning can be categorised in the following three categories.

1. Supervised machine learning,
2. Unsupervised machine learning,
3. Reinforcement learning



Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

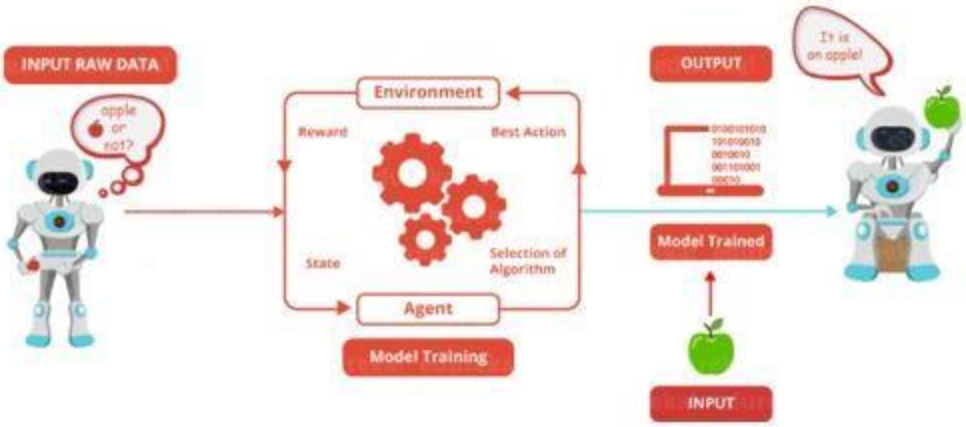
Q71. What, in your opinion, is the reason for the popularity of Deep Learning in recent times?

Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in recent years. This is because of two main reasons:

- The increase in the amount of data generated through various sources
- The growth in hardware resources required to run these models

GPUs are multiple times faster and they help us build bigger and deeper deep learning models in comparatively less time than we required previously.

Q72. What is reinforcement learning?



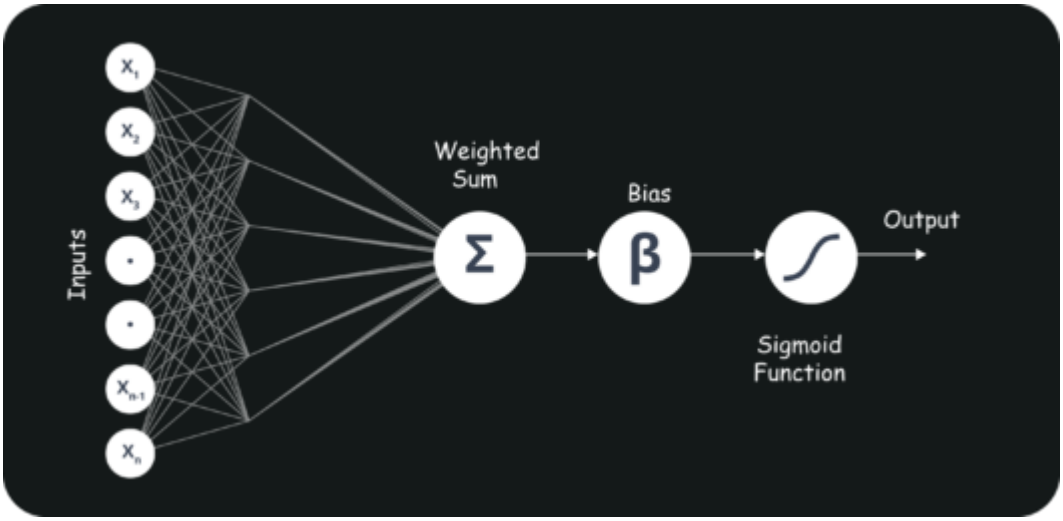
Reinforcement Learning is learning what to do and how to map situations to actions. The end result is to maximise the numerical reward signal. The learner is not told which action to take but instead must discover which action will yield the maximum reward. Reinforcement learning is inspired by the learning of human beings, it is based on the reward/penalty mechanism.

Q73. What are Artificial Neural Networks?

Artificial Neural networks are a specific set of algorithms that have revolutionized machine learning. They are inspired by biological neural networks. **Neural Networks** can adapt to changing the input so the network generates the best possible result without needing to redesign the output criteria.

Q74. Describe the structure of Artificial Neural Networks?

Artificial Neural Networks works on the same principle as a biological Neural Network. It consists of inputs which get processed with weighted sums and Bias, with the help of Activation Functions.



Q75. How Are Weights Initialized in a Network?

There are two methods here: we can either initialize the weights to zero or assign them randomly.

Initializing all weights to 0: This makes your model similar to a linear model. All the neurons and every layer perform the same operation, giving the same output and making the deep net useless.

Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

Initializing all weights randomly: Here, the weights are assigned randomly by initializing them very close to 0. It gives better accuracy to the model since every neuron performs different computations. This is the most commonly used method.

Q76. What Is the Cost Function?

Also referred to as “loss” or “error,” cost function is a measure to evaluate how good your model’s performance is. It’s used to compute the error of the output layer during backpropagation. We push that error backwards through the neural network and use that during the different training functions.

Q77. What Are Hyperparameters?

With neural networks, you’re usually working with ***hyperparameters*** once the data is formatted correctly. A hyperparameter is a parameter whose value is set before the learning process begins. It determines how a network is trained and the structure of the network (such as the number of hidden units, the learning rate, epochs, etc.).

Q78. What Will Happen If the Learning Rate Is Set inaccurately (Too Low or Too High)?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point.

If the learning rate is set too high, this causes undesirable divergent behaviour to the loss function due to drastic updates in weights. It may fail to converge (model can give a good output) or even diverge (data is too chaotic for the network to train).

Q79. What Is the Difference Between Epoch, Batch, and Iteration in Deep Learning?

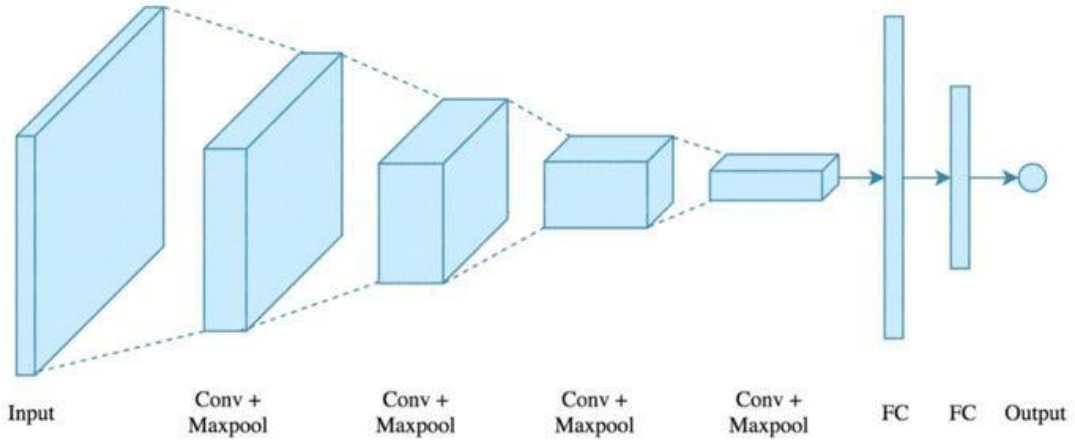
- Epoch – Represents one iteration over the entire dataset (everything put into the training model).
- Batch – Refers to when we cannot pass the entire dataset into the neural network at once, so we divide the dataset into several batches.

- Iteration – if we have 10,000 images as data and a batch size of 200. then an epoch should run 50 iterations (10,000 divided by 50).

Q80. What Are the Different Layers on CNN?

There are four layers in ***CNN***:

1. Convolutional Layer – the layer that performs a convolutional operation, creating several smaller picture windows to go over the data.
2. ReLU Layer – it brings non-linearity to the network and converts all the negative pixels to zero. The output is a rectified feature map.
3. Pooling Layer – pooling is a down-sampling operation that reduces the dimensionality of the feature map.
4. Fully Connected Layer – this layer recognizes and classifies the objects in the image.



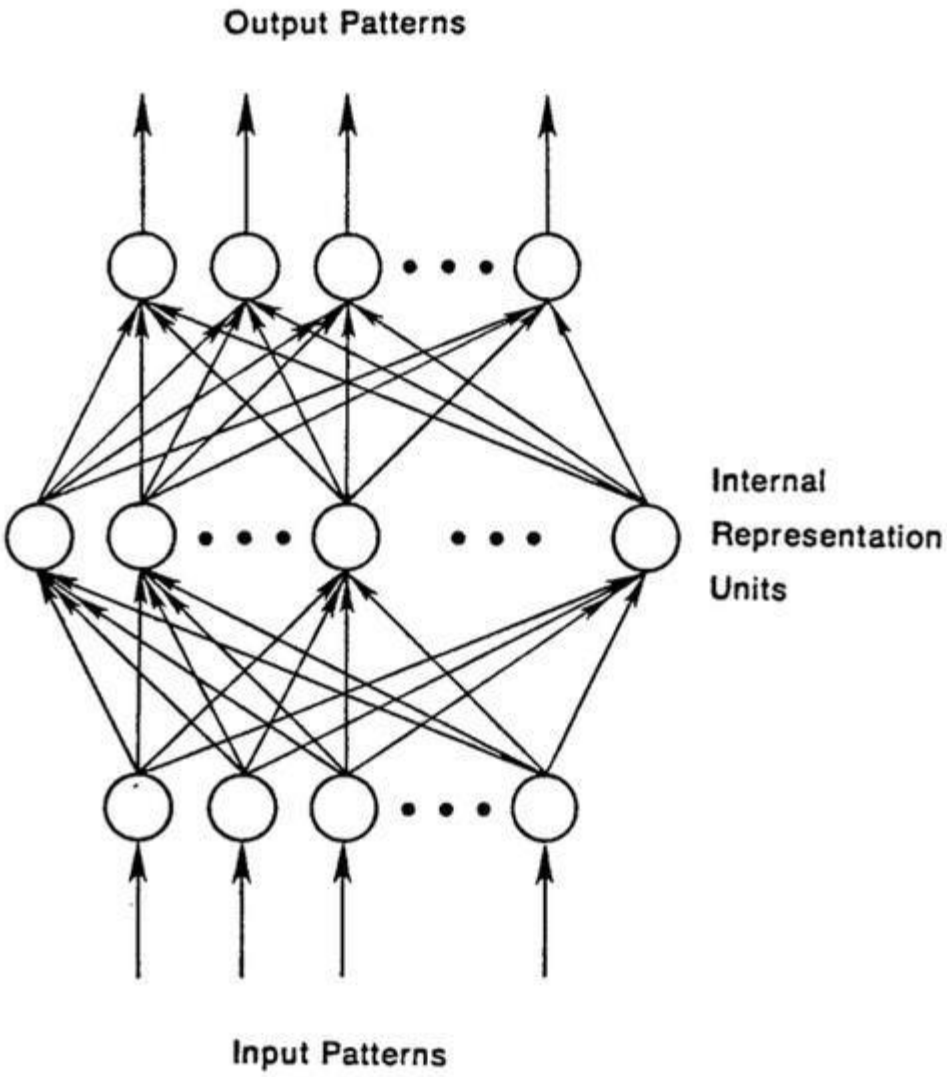
Q81. What Is Pooling on CNN, and How Does It Work?

Pooling is used to reduce the spatial dimensions of a CNN. It performs down-sampling operations to reduce the dimensionality and creates a pooled feature map by sliding a filter matrix over the input matrix.

Q82. What are Recurrent Neural Networks(RNNs)?

RNNs are a type of artificial neural networks designed to recognise the pattern from the sequence of data such as Time series, stock market and government agencies etc. To understand recurrent nets, first, you have to understand the basics of feedforward nets.

Both these networks RNN and feed-forward named after the way they channel information through a series of mathematical orations performed at the nodes of the network. One feeds information through straight(never touching the same node twice), while the other cycles it through a loop, and the latter are called recurrent.



Recurrent networks, on the other hand, take as their input, not just the current input example they see, but also the what they have perceived previously in time.

The decision a recurrent neural network reached at time t-1 affects the decision that it will reach one moment later at time t. So recurrent networks have two sources of input, the present and the recent past, which combine to determine how they respond to new data, much as we do in life.

The error they generate will return via backpropagation and be used to adjust their weights until error can't go any lower. Remember, the purpose of recurrent nets is to accurately classify sequential input. We rely on the backpropagation of error and gradient descent to do so.

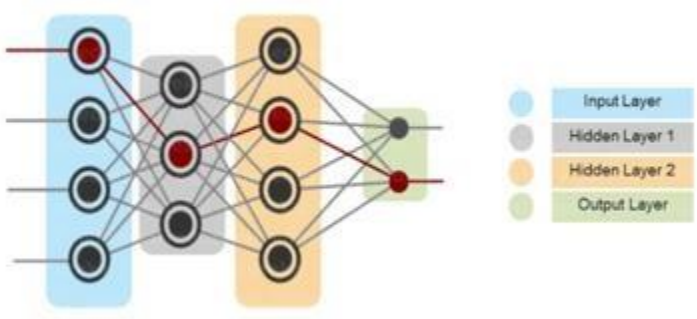
Q83. How Does an LSTM Network Work?

Long-Short-Term Memory (LSTM) is a special kind of recurrent neural network capable of learning longterm dependencies, remembering information for long periods as its default behaviour. There are three steps in an LSTM network:

- **Step 1:** The network decides what to forget and what to remember.
- **Step 2:** It selectively updates cell state values.
- **Step 3:** The network decides what part of the current state makes it to the output.

Q84. What Is a Multi-layer Perceptron(MLP)?

As in ***Neural Networks, MLPs*** have an input layer, a hidden layer, and an output layer. It has the same structure as a single layer ***perceptron*** with one or more hidden layers. A single layer perceptron can classify only linear separable classes with binary output (0,1), but MLP can classify nonlinear classes.



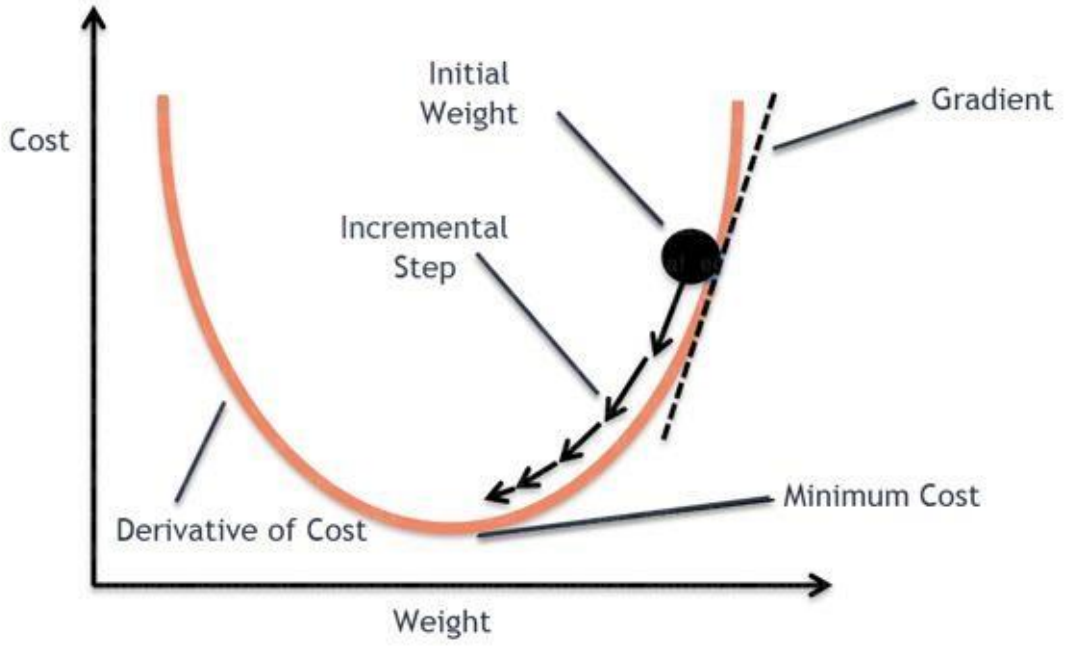
Except for the input layer, each node in the other layers uses a nonlinear activation function. This means the input layers, the data coming in, and the activation function is based upon all nodes and weights being added together, producing the output. MLP uses a supervised learning method called “backpropagation.” In ***backpropagation***, the neural network calculates the error with the help of cost function. It propagates this error backward from where it came (adjusts the weights to train the model more accurately).

Q85. Explain Gradient Descent.

To Understand Gradient Descent, Let’s understand what is a ***Gradient*** first.

A **gradient** measures how much the output of a function changes if you change the inputs a little bit. It simply measures the change in all weights with regard to the change in error. You can also think of a gradient as the slope of a function.

Gradient Descent can be thought of climbing down to the bottom of a valley, instead of climbing up a hill. This is because it is a minimization algorithm that minimizes a given function (**Activation Function**).



Q86. What is exploding gradients?

While training an RNN, if you see **exponentially growing (very large) error gradients** which accumulate and result in very large updates to neural network model weights during training, they're known as exploding gradients. At an extreme, the values of weights can become so large as to overflow and result in NaN values.

This has the effect of your model is unstable and unable to learn from your training data.

Q87. What is vanishing gradients?

While training an RNN, your slope can become either too small; this makes the training difficult. When the slope is too small, the problem is known as a Vanishing Gradient. It leads to long training times, poor performance, and low accuracy.

Q89. What is Back Propagation and Explain it's Working.

Backpropagation is a training algorithm used for multilayer neural network. In this method, we move the error from an end of the network to all weights inside the network and thus allowing efficient computation of the gradient.

It has the following steps:

See Batch Details

- Forward Propagation of Training Data
- Derivatives are computed using output and target
- Back Propagate for computing derivative of error wrt output activation
- Using previously calculated derivatives for output
- Update the Weights

Q90. What are the variants of Back Propagation?

- **Stochastic Gradient Descent:** We use only a single training example for calculation of gradient and update parameters.
- **Batch Gradient Descent:** We calculate the gradient for the whole dataset and perform the update at each iteration. • **Mini-batch Gradient Descent**: It's one of the most popular optimization algorithms. It's a variant of Stochastic Gradient Descent and here instead of single training example, mini-batch of samples is used.

Q91. What are the different *Deep Learning Frameworks*?

- ***Pytorch***
- ***TensorFlow***
- Microsoft Cognitive Toolkit

- Keras
- Caffe
- Chainer

Q92. What is the role of the Activation Function?

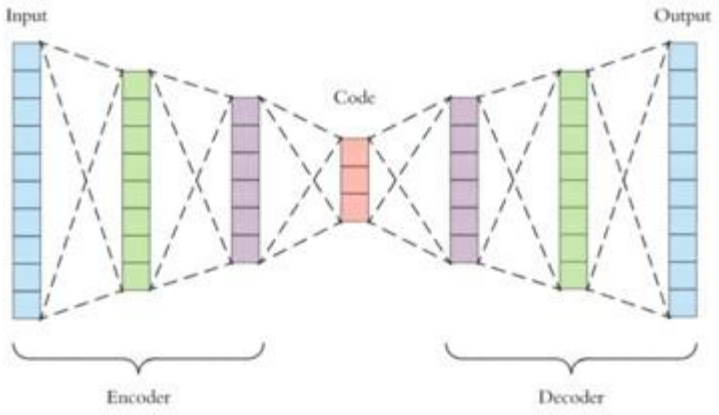
The ***Activation function*** is used to introduce non-linearity into the neural network helping it to learn more complex function. Without which the neural network would be only able to learn linear function which is a linear combination of its input data. An activation function is a function in an artificial neuron that delivers an output based on inputs.

Q93. Name a few *Machine Learning libraries* for various purposes.

Purpose	Libraries
Scientific Computation	Numpy
Tabular Data	Pandas
Data Modelling & Preprocessing	Scikit Learn
Time-Series Analysis	Statsmodels
Text processing	Regular Expressions, NLTK
Deep Learning	Tensorflow, Pytorch

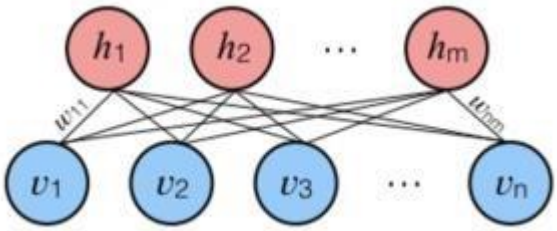
Q94. What is an Auto-Encoder?

Auto-encoders are simple learning networks that aim to transform inputs into outputs with the minimum possible error. This means that we want the output to be as close to input as possible. We add a couple of layers between the input and the output, and the sizes of these layers are smaller than the input layer. The auto-encoder receives unlabelled input which is then encoded to reconstruct the input.



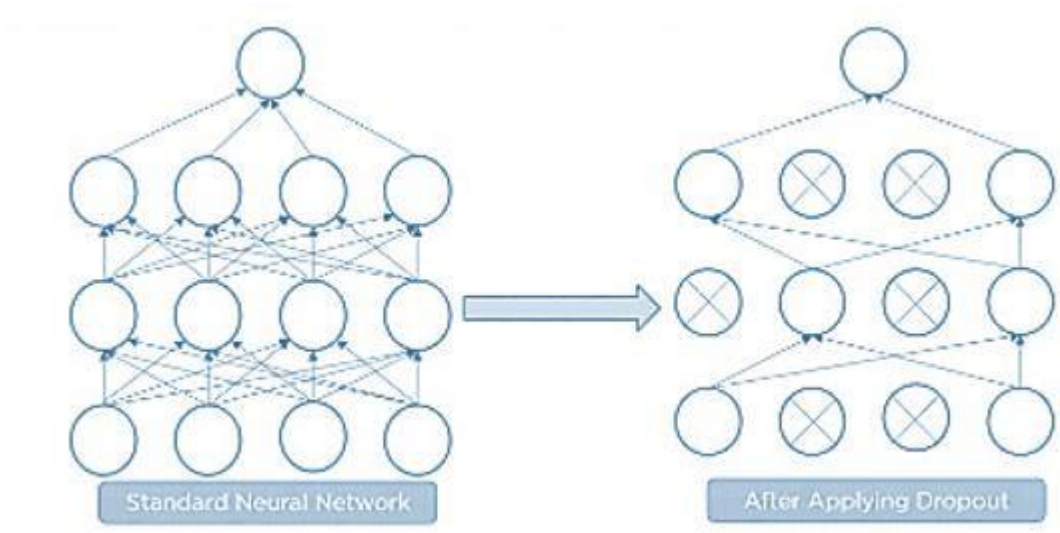
Q95. What is a Boltzmann Machine?

Boltzmann machines have a simple learning algorithm that allows them to discover interesting features that represent complex regularities in the training data. The Boltzmann machine is basically used to optimise the weights and the quantity for the given problem. The learning algorithm is very slow in networks with many layers of feature detectors. "**Restricted Boltzmann Machines**" algorithm has a single layer of feature detectors which makes it faster than the rest.



Q96. What Is Dropout and Batch Normalization?

Dropout is a technique of dropping out hidden and visible units of a network randomly to prevent overfitting of data (typically dropping 20 per cent of the nodes). It doubles the number of iterations needed to converge the network.



Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

Batch normalization is the technique to improve the performance and stability of neural networks by normalizing the inputs in every layer so that they have mean output activation of zero and standard deviation of one.

Q97. What Is the Difference Between Batch Gradient Descent and Stochastic Gradient Descent?

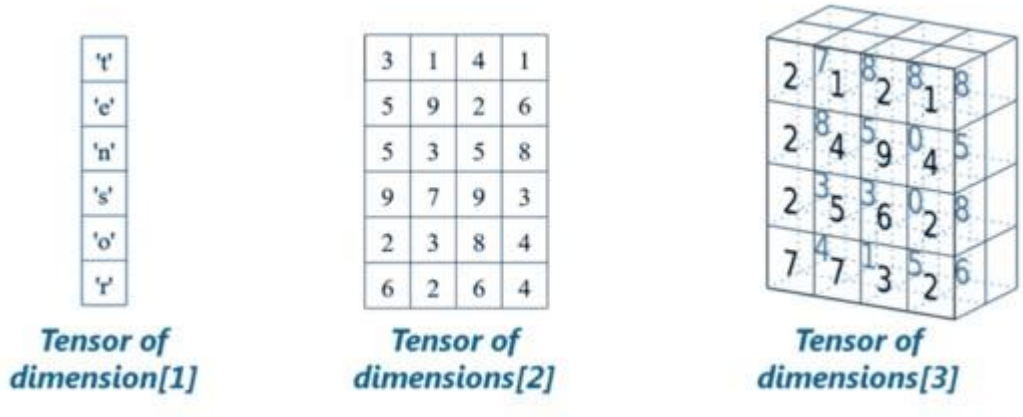
Batch Gradient Descent	Stochastic Gradient Descent
The batch gradient computes the gradient using the entire dataset.	The stochastic gradient computes the gradient using a single sample.
It takes time to converge because the volume of data is huge, and weights update slowly.	It converges much faster than the batch gradient because it updates weight more frequently.

Q98. Why Is Tensorflow the Most Preferred Library in Deep Learning?

Tensorflow provides both C++ and Python APIs, making it easier to work on and has a faster compilation time compared to other Deep Learning libraries like Keras and Torch. Tensorflow supports both CPU and GPU computing devices.

Q99. What Do You Mean by Tensor in Tensorflow?

A tensor is a mathematical object represented as arrays of higher dimensions. These arrays of data with different dimensions and ranks fed as input to the neural network are called "***Tensors***."



Q100. What is the Computational Graph?

Everything in a tensorflow is based on creating a computational graph. It has a network of nodes where each node operates, Nodes represent mathematical operations, and edges represent tensors. Since data flows in the form of a graph, it is also called a "DataFlow Graph."

Q101. What are the differences between supervised and unsupervised learning?

Supervised Learning	Unsupervised Learning

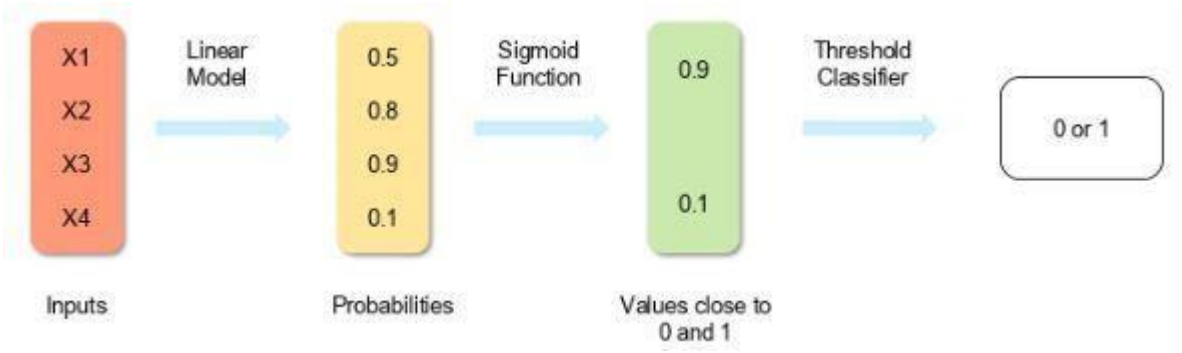
Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

• Uses known and labeled data as input • Supervised learning has a feedback mechanism • Most commonly used supervised learning algorithms are decision trees, logistic regression, and support vector machine	• Uses unlabeled data as input • Unsupervised learning has no feedback mechanism • Most commonly used unsupervised learning algorithms are k-means clustering, hierarchical clustering, and apriori algorithm

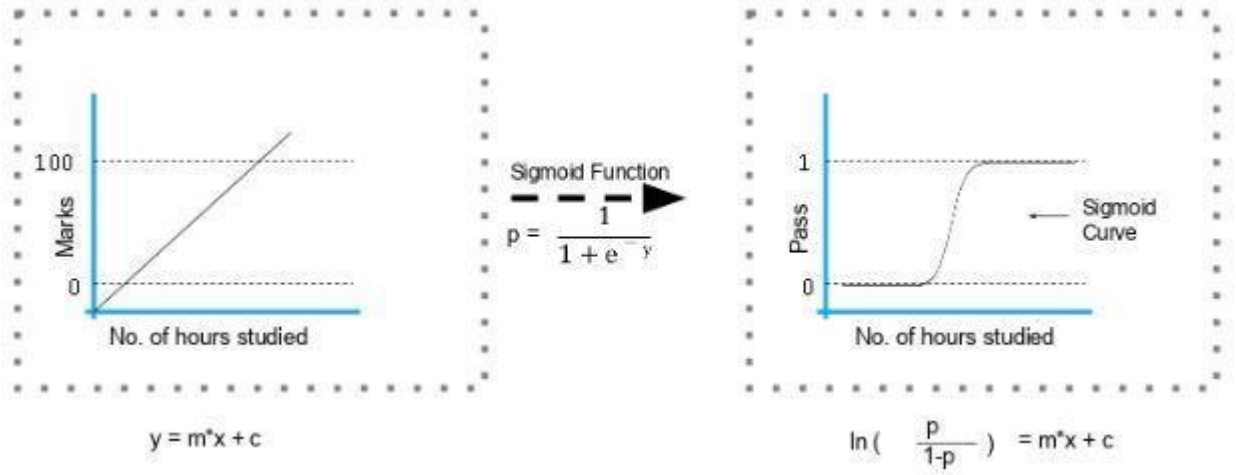
102. How is logistic regression done?

Logistic regression measures the relationship between the dependent variable (our label of what we want to predict) and one or more independent variables (our features) by estimating probability using its underlying logistic function (sigmoid).

The image shown below depicts how logistic regression works:



The formula and graph for the sigmoid function are as shown:



Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

103. Explain the steps in making a decision tree.

1. Take the entire data set as input
2. Calculate entropy of the target variable, as well as the predictor attributes
3. Calculate your information gain of all attributes (we gain information on sorting different objects from each other)
4. Choose the attribute with the highest information gain as the root node
5. Repeat the same procedure on every branch until the decision node of each branch is finalized

For example, let's say you want to build a decision tree to decide whether you should accept or decline a job offer. The decision tree for this case is as shown:

It is clear from the decision tree that an offer is accepted if:

- Salary is greater than $50,000
- The commute is less than an hour
- Incentives are offered

104. How do you build a random forest model?

A random forest is built up of a number of decision trees. If you split the data into different packages and make a decision tree in each of the different groups of data, the random forest brings all those trees together.

Steps to build a random forest model:

1. Randomly select 'k' features from a total of'm' features where k << m
2. Among the 'k' features, calculate the node D using the best split point
3. Split the node into daughter nodes using the best split

4. Repeat steps two and three until leaf nodes are finalized
5. Build forest by repeating steps one to four for 'n' times to create 'n' number of trees

105. How can you avoid the overfitting your model?

Overfitting refers to a model that is only set for a very small amount of data and ignores the bigger picture. There are three main methods to avoid overfitting:

1. Keep the model simple—take fewer variables into account, thereby removing some of the noise in the training data
2. Use cross-validation techniques, such as k folds cross-validation
3. Use regularization techniques, such as LASSO, that penalize certain model parameters if they're likely to cause overfitting

106. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate

Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it. Example: height of students

Height (in cm)
164
167.3
170
174.2
178

180

The patterns can be studied by drawing conclusions using mean, median, mode, dispersion or range, minimum, maximum, etc.

Bivariate

Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.

Example: temperature and ice cream sales in the summer season

Temperature (in Celcius)	Sales
20	2,000
25	2,100
26	2,300
28	2,400
30	2,600
36	3,100

Here, the relationship is visible from the table that temperature and sales are directly proportional to each other. The hotter the temperature, the better the sales.

Multivariate

Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate, but contains more than one dependent variable.

Example: data for house price prediction

No. of rooms	Floors	Area (sq ft)	Price
2	0	900	$4000,00
3	2	1,100	$600,000
3.5	5	1,500	$900,000
4	3	2,100	$1,200,000

The patterns can be studied by drawing conclusions using mean, median, and mode, dispersion or range, minimum, maximum, etc. You can start describing the data and using it to guess what the price of the house will be.

107. What are the feature selection methods used to select the right variables?

There are two main methods for feature selection:

Filter Methods This

involves:

- Linear discrimination analysis

- ANOVA
- Chi-Square

The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting the features, it's all about cleaning up the data coming in.

Wrapper Methods This

involves:

- Forward Selection: We test one feature at a time and keep adding them until we get a good fit
- Backward Selection: We test all the features and start removing them to see what works better
- Recursive Feature Elimination: Recursively looks through all the different features and how they pair together

Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is performed with the wrapper method.

108. In your choice of language, write a program that prints the numbers ranging from one to 50.

But for multiples of three, print "Fizz" instead of the number and for the multiples of five, print "Buzz." For numbers which are multiples of both three and five, print "FizzBuzz" The code is shown below:

```
for fizzbuzz in range(51):
    if fizzbuzz % 3 == 0 and fizzbuzz % 5 == 0:
        print("fizzbuzz")
        continue
    elif fizzbuzz % 3 == 0:
        print("fizz")
        continue
    elif fizzbuzz % 5 == 0:
        print("buzz")
        continue
    print(fizzbuzz)
```

Note that the range mentioned is 51, which means zero to 50. However, the range asked in the question is one to 50. Therefore, in the above code, you can include the range as (1,51).

The output of the above code is as shown:

109. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

The following are ways to handle missing data values:

If the data set is large, we can just simply remove the rows with missing data values. It is the quickest way; we use the rest of the data to predict the values.

For smaller data sets, we can substitute missing values with the mean or average of the rest of the data using pandas data frame in python. There are different ways to do so, such as df.mean(), df.fillna(mean).

110. For the given points, how will you calculate the Euclidean distance in Python?

plot1 = [1,3] plot2

= [2,5]

The Euclidean distance can be calculated as follows:

euclidean_distance = sqrt((plot1[0]-plot2[0])**2 + (plot1[1]-plot2[1])**2)

111. What are dimensionality reduction and its benefits?

Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.

This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

112. How will you calculate eigenvalues and eigenvectors of the following 3x3 matrix?

-2	-4	2
-2	1	2
4	2	5

The characteristic equation is as shown:

Expanding determinant:

(-2 – λ) [(1-λ) (5-λ)-2x2] + 4[(-2) x (5-λ) -4x2] + 2[(-2) x 2-4(1-λ)] =0

- λ3 + 4λ2 + 27λ – 90 = 0,

λ3 - 4 λ2 -27 λ + 90 = 0

Here we have an algebraic equation built from the eigenvectors.

By hit and trial:

33 – 4 x 32 - 27 x 3 +90 = 0 Hence,

(λ - 3) is a factor:

λ3 - 4 λ2 - 27 λ +90 = (λ – 3) (λ2 – λ – 30) Eigenvalues

are 3,-5,6:

(λ – 3) (λ2 – λ – 30) = (λ – 3) (λ+5) (λ-6),

Calculate eigenvector for λ = 3

For X = 1,

-5 - 4Y + 2Z =0,

-2 - 2Y + 2Z =0

Subtracting the two equations:

3 + 2Y = 0,

Subtracting back into second equation:

Y = -(3/2)

Z = -(1/2)

Similarly, we can calculate the eigenvectors for -5 and 6.

113. How should you maintain a deployed model?

The steps to maintain a deployed model are:

Monitor

Constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it's doing what it's supposed to do.

Evaluate

Evaluation metrics of the current model are calculated to determine if a new algorithm is needed.

Compare

The new models are compared to each other to determine which model performs the best.

Rebuild

The best performing model is re-built on the current state of data.

114. What are recommender systems?

A recommender system predicts what a user would rate a specific product based on their preferences. It can be split into two different areas:

Collaborative filtering

As an example, Last.fm recommends tracks that other users with similar interests play often. This is also commonly seen on Amazon after making a purchase; customers may notice the following message accompanied by product recommendations: "Users who bought this also bought…" Content-based filtering

As an example: Pandora uses the properties of a song to recommend music with similar properties. Here, we look at content, instead of looking at who else is listening to music.

115. How do you find RMSE and MSE in a linear regression model?

RMSE and MSE are two of the most common measures of accuracy for a linear regression model.

RMSE indicates the Root Mean Square Error.

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

MSE indicates the Mean Square Error.

$$MSE=\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}$$

116. How can you select k for k-means?

We use the elbow method to select k for k-means clustering. The idea of the elbow method is to run kmeans clustering on the data set where 'k' is the number of clusters.

Within the sum of squares (WSS), it is defined as the sum of the squared distance between each member of the cluster and its centroid.

117. What is the significance of p-value?

p-value typically ≤ 0.05

This indicates strong evidence against the null hypothesis; so you reject the null hypothesis.

p-value typically > 0.05

This indicates weak evidence against the null hypothesis, so you accept the null hypothesis. p-value

at cutoff 0.05

This is considered to be marginal, meaning it could go either way.

118. How can outlier values be treated?

You can drop outliers only if it is a garbage value.

Example: height of an adult = abc ft. This cannot be true, as the height cannot be a string value. In this case, outliers can be removed.

If the outliers have extreme values, they can be removed. For example, if all the data points are clustered between zero to 10, but one point lies at 100, then we can remove this point.

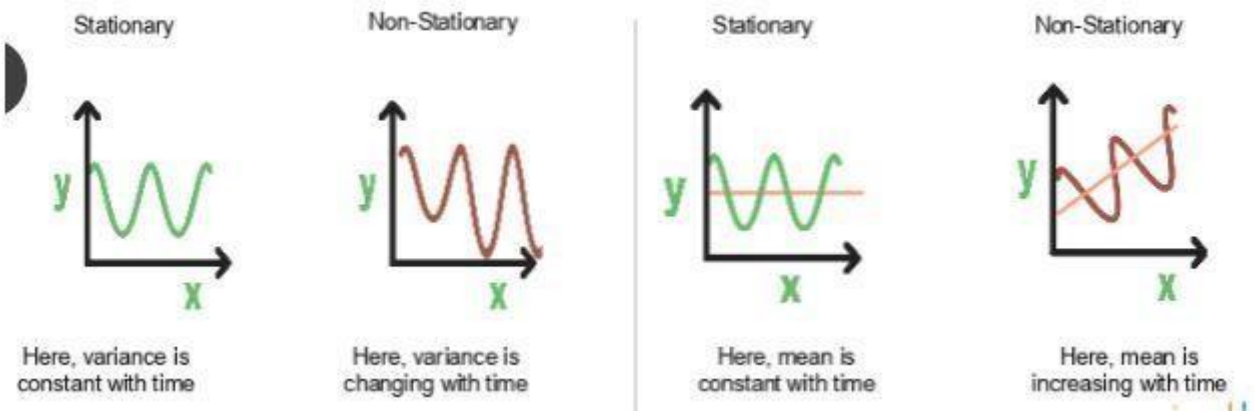
If you cannot drop outliers, you can try the following:

- Try a different model. Data detected as outliers by linear models can be fit by nonlinear models. Therefore, be sure you are choosing the correct model.
- Try normalizing the data. This way, the extreme data points are pulled to a similar range.
- You can use algorithms that are less affected by outliers; an example would be random forests.

119. How can a time-series data be declared as stationery?

It is stationary when the variance and mean of the series are constant with time.

Here is a visual example:



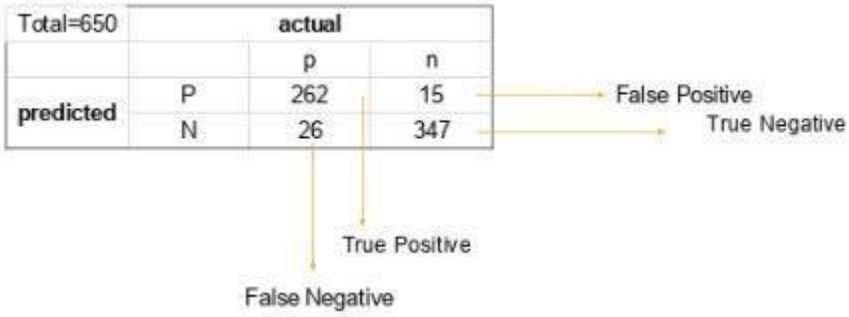
Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

In the first graph, the variance is constant with time. Here, X is the time factor and Y is the variable. The value of Y goes through the same points all the time; in other words, it is stationary.

In the second graph, the waves get bigger, which means it is non-stationary and the variance is changing with time.

120. How can you calculate accuracy using a confusion matrix?

Consider this confusion matrix:



You can see the values for total data, actual values, and predicted values.

The formula for accuracy is:

Accuracy = (True Positive + True Negative) / Total Observations

= (262 + 347) / 650

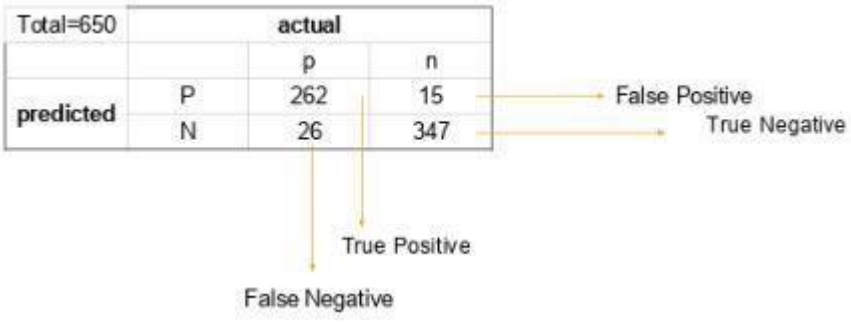
= 609 / 650

= 0.93

As a result, we get an accuracy of 93 percent.

121. Write the equation and calculate the precision and recall rate.

Consider the same confusion matrix used in the previous question.



Follow Steve Nouri for more AI and Data science posts: https://lnkd.in/gZu463X

Precision = (True positive) / (True Positive + False Positive)

= 262 / 277

= 0.94

Recall Rate = (True Positive) / (Total Positive + False Negative)

= 262 / 288

= 0.90

122. 'People who bought this also bought…' recommendations seen on Amazon are a result of which algorithm?

The recommendation engine is accomplished with collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.

The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.

For example, a sales page shows that a certain number of people buy a new phone and also buy tempered glass at the same time. Next time, when a person buys a phone, he or she may see a recommendation to buy tempered glass as well.

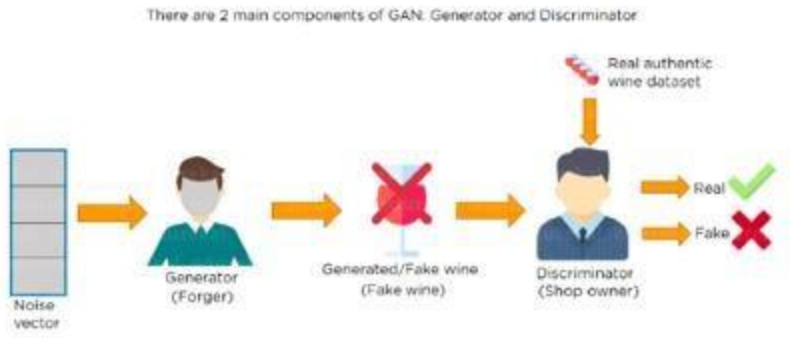
123. What is a Generative Adversarial Network?

Suppose there is a wine shop purchasing wine from dealers, which they resell later. But some dealers sell fake wine. In this case, the shop owner should be able to distinguish between fake and authentic wine.

The forger will try different techniques to sell fake wine and make sure specific techniques go past the shop owner's check. The shop owner would probably get some feedback from wine experts that some of the wine is not original. The owner would have to improve how he determines whether a wine is fake or authentic.

The forger's goal is to create wines that are indistinguishable from the authentic ones while the shop owner intends to tell if the wine is real or not accurately

Let us understand this example with the help of an image.



There is a noise vector coming into the forger who is generating fake wine.

Here the forger acts as a Generator.

The shop owner acts as a Discriminator.

The Discriminator gets two inputs; one is the fake wine, while the other is the real authentic wine. The shop owner has to figure out whether it is real or fake.

So, there are two primary components of Generative Adversarial Network (GAN) named:

1. Generator
2. Discriminator

The generator is a CNN that keeps keys producing images and is closer in appearance to the real images while the discriminator tries to determine the difference between real and fake images The ultimate aim is to make the discriminator learn to identify real and fake images.

Apart from the very technical questions, your interviewer could even hit you up with a few simple ones to check your overall confidence, in the likes of the following.

124. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why shouldn't you be happy with your model performance? What can you do about it?

Cancer detection results in imbalanced data. In an imbalanced dataset, accuracy should not be based as a measure of performance. It is important to focus on the remaining four percent, which represents the patients who were wrongly diagnosed. Early diagnosis is crucial when it comes to cancer detection, and can greatly improve a patient's prognosis.

Hence, to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine the class wise performance of the classifier.

125. Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables?

- K-means clustering

- Linear regression
- K-NN (k-nearest neighbor)
- Decision trees

The K nearest neighbor algorithm can be used because it can compute the nearest neighbor and if it doesn't have a value, it just computes the nearest neighbor based on all the other features.

When you're dealing with K-means clustering or linear regression, you need to do that in your preprocessing, otherwise, they'll crash. Decision trees also have the same problem, although there is some variance.

126. Below are the eight actual values of the target variable in the train file. What is the entropy of the target variable?

[0, 0, 0, 1, 1, 1, 1, 1]

Choose the correct answer.

1. -(5/8 log(5/8) + 3/8 log(3/8))
2. 5/8 log(5/8) + 3/8 log(3/8)
3. 3/8 log(5/8) + 5/8 log(3/8)
4. 5/8 log(3/8) – 3/8 log(5/8)

The target variable, in this case, is 1.

The formula for calculating the entropy is:

Putting p=5 and n=8, we get

Entropy = A = -(5/8 log(5/8) + 3/8 log(3/8))

127. We want to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case?

Choose the correct option:

1. Logistic Regression
2. Linear Regression
3. K-means clustering
4. Apriori algorithm

The most appropriate algorithm for this case is A, logistic regression.

128. After studying the behavior of a population, you have identified four specific individual types that are valuable to your study. You would like to find all users who are most similar to each individual type. Which algorithm is most appropriate for this study?

Choose the correct option:

1. K-means clustering
2. Linear regression
3. Association rules
4. Decision trees

As we are looking for grouping people together specifically by four different similarities, it indicates the value of k. Therefore, K-means clustering (answer A) is the most appropriate algorithm for this study.

129. You have run the association rules algorithm on your dataset, and the two rules {banana, apple} => {grape} and {apple, orange} => {grape} have been found to be relevant. What else must be true?

Choose the right answer:

1. {banana, apple, grape, orange} must be a frequent itemset
2. {banana, apple} => {orange} must be a relevant rule
3. {grape} => {banana, apple} must be a relevant rule
4. {grape, apple} must be a frequent itemset

The answer is A: {grape, apple} must be a frequent itemset

130. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should you use?

1. One-way ANOVA
2. K-means clustering
3. Association rules
4. Student's t-test

The answer is A: One-way ANOVA

131. What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an object in a mathematical way that's easy to analyze.

132. What are the steps in making a decision tree?

1. Take the entire data set as input.
2. Look for a split that maximizes the separation of the classes. A split is any test that divides the data into two sets.
3. Apply the split to the input data (divide step).
4. Re-apply steps one and two to the divided data.
5. Stop when you meet any stopping criteria.
6. This step is called pruning. Clean up the tree if you went too far doing splits.

133. What is root cause analysis?

Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring.

134. What is logistic regression?

Logistic regression is also known as the logit model. It is a technique used to forecast the binary outcome from a linear combination of predictor variables.

135. What are recommender systems?

Recommender systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product.

136. Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. It is mainly used in backgrounds where the objective is to forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) to limit problems like overfitting and gain insight into how the model will generalize to an independent data set.

137. What is collaborative filtering?

Most recommender systems use this filtering process to find patterns and information by collaborating perspectives, numerous data sources, and several agents.

138. Do gradient descent methods always converge to similar points?

They do not, because in some cases, they reach a local minima or a local optima point. You would not reach the global optima point. This is governed by the data and the starting conditions.

139. What is the goal of A/B Testing?

This is statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B testing is to detect any changes to a web page to maximize or increase the outcome of a strategy.

140. What are the drawbacks of the linear model?

- The assumption of linearity of the errors
- It can't be used for count outcomes or binary outcomes
- There are overfitting problems that it can't solve

141. What is the law of large numbers?

It is a theorem that describes the result of performing the same experiment very frequently. This theorem forms the basis of frequency-style thinking. It states that the sample mean, sample variance and sample standard deviation converge to what they are trying to estimate.

142. What are the confounding variables?

These are extraneous variables in a statistical model that correlates directly or inversely with both the dependent and the independent variable. The estimate fails to account for the confounding factor.

143. What is star schema?

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes, star schemas involve several layers of summarization to recover information faster.

144. How regularly must an algorithm be updated?

You will want to update an algorithm when:

- You want the model to evolve as data streams through infrastructure
- The underlying data source is changing
- There is a case of non-stationarity

145. What are eigenvalue and eigenvector?

Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing, or stretching.

Eigenvectors are for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix.

146. Why is resampling done?

Resampling is done in any of these cases:

- Estimating the accuracy of sample statistics by using subsets of accessible data, or drawing randomly with replacement from a set of data points
- Substituting labels on data points when performing significance tests • Validating models

  by using random subsets (bootstrapping, cross-validation)

147. What is selection bias?

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample.

148. What are the types of biases that can occur during sampling?

1. Selection bias
2. Undercoverage bias
3. Survivorship bias

149. What is survivorship bias?

Survivorship bias is the logical error of focusing on aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.

150. How do you work towards a random forest?

The underlying principle of this technique is that several weak learners combine to provide a strong learner. The steps involved are:

1. Build several decision trees on bootstrapped training samples of data
2. On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates out of all pp predictors
3. Rule of thumb: At each split m=p√m=p
4. Predictions: At the majority rule

151. What are the important skills to have in Python with regard to data analysis?

The following are some of the important skills to possess which will come handy when performing data analysis using Python.

- Good understanding of the built-in data types especially lists, dictionaries, tuples, and sets.
- Mastery of N-dimensional NumPy Arrays.
- Mastery of Pandas dataframes.
- Ability to perform element-wise vector and matrix operations on NumPy arrays.
- Knowing that you should use the Anaconda distribution and the conda package manager.
- Familiarity with Scikit-learn. ****Scikit-Learn Cheat Sheet****
- Ability to write efficient list comprehensions instead of traditional for loops.
- Ability to write small, clean functions (important for any developer), preferably pure functions that don't alter objects.
- Knowing how to profile the performance of a Python script and how to optimize bottlenecks.

Credit: kdnuggets, Simplilearn, Edureka, Guru99, Hackernoon, Datacamp, Nitin Panwar, Michael Rundell

Part 1

MACHINE LEARNING INTERVIEW QUESTIONS

SWIPE →

Q1: What is the difference between supervised and unsupervised machine learning?

Supervised learning requires training labelled data.

For example, in order to do classification (a supervised learning task), you'll need to first label the data you'll use to train the model to classify data into your labelled groups.

Unsupervised learning, in contrast, does not require labelling data explicitly.

Q2: What is regularization? Can you give some examples of regularization techniques?

Regularization is any technique that aims to improve the validation score, sometimes at the cost of reducing the training score.

Some regularization techniques:

- L1 tries to minimize the absolute value of the parameters of the model. It produces sparse parameters.
- L2 tries to minimize the square value of the parameters of the model. It produces parameters with small values.

Q3: What is stratified cross-validation and when should we use it?

Cross-validation is a technique for dividing data between training and validation sets. On typical cross-validation, this split is done randomly. But in stratified cross-validation, the split preserves the ratio of the categories on both the training and validation datasets.

Stratified cross-validation may be applied in the following scenarios:

- On a dataset with multiple categories
- On a dataset with data of different distributions

Q4: What is deep learning, and how does it contrast with other machine learning algorithms?

Deep learning is a subset of machine learning that is concerned with neural networks: how to use backpropagation and certain principles from neuroscience to more accurately model large sets of unlabelled or semi-structured data.

In that sense, deep learning represents an unsupervised learning algorithm that learns representations of data through the use of neural nets.

Q5: How would you handle an imbalanced dataset?

An imbalanced dataset is when you have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if you have no predictive power on the other category of data! Here are a few tactics to get over the hump:

- Collect more data to even the imbalances in the dataset.
- Resample the dataset to correct for imbalances.
- Try a different algorithm altogether on your dataset.

Q6: Why do ensembles typically have higher scores than individual models?

An ensemble is the combination of multiple models to create a single prediction. The key idea for making better predictions is that the models should make different errors. That way the errors of one model will be compensated by the right guesses of the other models and thus the score of the ensemble will be higher.

Many winning solutions to data science competitions are ensembles. However, in real-life machine learning projects, engineers need to find a balance between execution time and accuracy.

Q7: What evaluation approaches would you work to gauge the effectiveness of a ML model?

You would first split the dataset into training and test sets, or perhaps use cross-validation techniques to further segment the dataset into composite sets of training and test sets within the data.

You should then implement a choice selection of performance metrics. You could use measures such as the F1 score, the accuracy, and the confusion matrix. What's important here is to demonstrate that you understand the nuances of how a model is measured and how to choose the right performance measures for the right situations.

Q8: What's the "kernel trick" and how is it useful?

The Kernel trick involves kernel functions that can enable in higher-dimension spaces without explicitly calculating the coordinates of points within that dimension: instead, kernel functions compute the inner products between the images of all pairs of data in a feature space.

This allows them the very useful attribute of calculating the coordinates of higher dimensions while being computationally cheaper than the explicit calculation of said coordinates. Many algorithms can be expressed in terms of inner products. Using the kernel trick enables us effectively run algorithms in a high-dimensional space with lower-dimensional data.

Data-Driven Science

FOUND IT HELPFUL?

Let us know in the comments

At DDS, we will help you build a career in Data Science

LINK IN BIO
