MACHINE LEARNING

(PROGRAMMING & COMPANY/INDUSTRY SPECIFIC)

Interview Questions

BY – VIKAS MAURYA

YOUTUBE

1. CODE WITH VIKAS

2. VIKAS MAURYA ACADEMY

Q1: What's the trade-off between bias and variance?

Answer: Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you're using. This can lead to the model underfitting your data, making it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set.

Variance is error due to too much complexity in the learning algorithm you're using. This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead your model to overfit the data. You'll be carrying too much noise from your training data for your model to be very useful for your test data.

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance. You don't want either high bias or high variance in your model.

Q2: What is the difference between supervised and unsupervised machine learning?

Answer: Supervised learning requires training labeled 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 labeled groups. Unsupervised learning, in contrast, does not require labeling data explicitly.

Q3: How is KNN different from k-means clustering?

Answer: K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this really means is that in order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled point into (thus the nearest neighbor part). K-means clustering requires only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between different points.

The critical difference here is that KNN needs labeled points and is thus supervised learning, while k-means doesn't—and is thus unsupervised learning.

Q4: Explain how a ROC curve works.

Answer: The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).

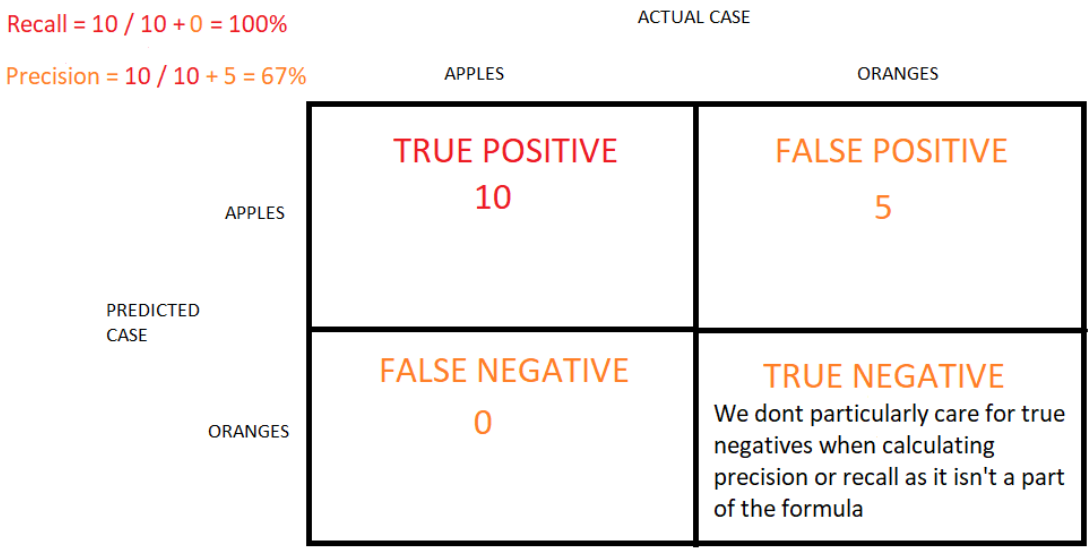
Q5: Define precision and recall.

Answer: Recall is also known as the true positive rate: the amount of positives your model claims compared to the actual number of positives there are throughout the data. Precision is also known as the positive predictive value, and it is a measure of the amount of accurate positives your model claims compared to the number of positives it actually claims. It can be easier to think of recall and precision in the context of a case where you've predicted that there were 10 apples and 5 oranges in a case of 10 apples. You'd have perfect recall (there are actually 10 apples, and you predicted there would be 10) but 66.7% precision because out of the 15 events you predicted, only 10 (the apples) are correct.

Explanation: Out of a sample size of 15 (10 apples + 5 oranges), you have identified 10 apples as apples BUT you have also incorrectly predicted 5 oranges as apples. This implies that the true positive figure is 10 (10 correctly identified apples), whereas the false positive figure is 5 (5 oranges incorrectly tagged as apples).

As per the formula of Precision = True Positive / (True Positive + False Positive), therefore the precision rate is 67%.

As per the Recall formula = True Positive / (True Positive + False Negative), hence the recall rate is 100%. This is because not a single apple was incorrectly predicted as an orange.



Q6: What is Bayes' Theorem? How is it useful in a machine learning context?

Answer: Bayes' Theorem gives you the posterior probability of an event given what is known as prior knowledge.

Mathematically, it's expressed as the true positive rate of a condition sample divided by the sum of the false positive rate of the population and the true positive rate of a condition. Say you had a 60% chance of actually having the flu after a flu test, but out of people who had the flu, the test will be false 50% of the time, and the overall population only has a 5% chance of having the flu. Would you actually have a 60% chance of having the flu after having a positive test?

Bayes' Theorem says no. It says that you have a (.6 * 0.05) (True Positive Rate of a Condition Sample) / (.6*0.05)(True Positive Rate of a Condition Sample) + (.5*0.95) (False Positive Rate of a Population) = 0.0594 or 5.94% chance of getting a flu.

Bayes' Theorem is the basis behind a branch of machine learning that most notably includes the Naive Bayes classifier. That's something important to consider when you're faced with machine learning interview questions.

Q7: Why is "Naive" Bayes naive?

Answer: Despite its practical applications, especially in text mining, Naive Bayes is considered "Naive" because it makes an assumption that is virtually impossible to see in real-life data: the conditional probability is

calculated as the pure product of the individual probabilities of components. This implies the absolute independence of features — a condition probably never met in real life.

As a Quora commenter put it whimsically, a Naive Bayes classifier that figured out that you liked pickles and ice cream would probably naively recommend you a pickle ice cream.

Q8: Explain the difference between L1 and L2 regularization.

Answer: L2 regularization tends to spread error among all the terms, while L1 is more binary/sparse, with many variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a Laplacean prior on the terms, while L2 corresponds to a Gaussian prior.

Q9: What's your favorite algorithm, and can you explain it to me in less than a minute?

Answer: Interviewers ask such machine learning interview questions to test your understanding of how to communicate complex and technical nuances with poise and the ability to summarize quickly and efficiently. While answering such questions, make sure you have a choice and ensure you can explain different algorithms so simply and effectively that a five-year-old could grasp the basics!

Q10: What's the difference between Type I and Type II error?

Answer: Don't think that this is a trick question! Many machine learning interview questions will be an attempt to lob basic questions at you just to make sure you're on top of your game and you've prepared all of your bases.

Type I error is a false positive, while Type II error is a false negative. Briefly stated, Type I error means claiming something has happened when it hasn't, while Type II error means that you claim nothing is happening when in fact something is.

A clever way to think about this is to think of Type I error as telling a man he is pregnant, while Type II error means you tell a pregnant woman she isn't carrying a baby.

Q11: What's a Fourier transform?

Answer: A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. Or as this more intuitive tutorial puts it, given a smoothie, it's how we find the recipe. The Fourier transform finds the set of cycle speeds, amplitudes, and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain—it's a very common way to extract features from audio signals or other time series such as sensor data.

Q11: What's a Fourier transform?

Answer: A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. Or as this more intuitive tutorial puts it, given a smoothie, it's how we find the recipe. The Fourier transform finds the set of cycle speeds, amplitudes, and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain—it's a very common way to extract features from audio signals or other time series such as sensor data.

Q14: What's the difference between a generative and discriminative model?

Answer: A generative model will learn categories of data while a discriminative model will simply learn the distinction between different categories of data. Discriminative models will generally outperform generative models on classification tasks.

Q15: What cross-validation technique would you use on a time series dataset?

Answer: Instead of using standard k-folds cross-validation, you have to pay attention to the fact that a time series is not randomly distributed data—it is inherently ordered by chronological order. If a pattern emerges in later time periods, for example, your model may still pick up on it even if that effect doesn't hold in earlier years!

You'll want to do something like forward chaining where you'll be able to model on past data then look at forward-facing data.

- Fold 1 : training [1], test [2]
- Fold 2 : training [1 2], test [3]
- Fold 3 : training [1 2 3], test [4]
- Fold 4 : training [1 2 3 4], test [5]
- Fold 5 : training [1 2 3 4 5], test [6]

Q16: How is a decision tree pruned?

Answer: Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning.

Reduced error pruning is perhaps the simplest version: replace each node. If it doesn't decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

Q17: Which is more important to you: model accuracy or model performance?

Answer: Such machine learning interview questions tests your grasp of the nuances of machine learning model performance! Machine learning interview questions often look towards the details. There are models with higher accuracy that can perform worse in predictive power—how does that make sense?

Well, it has everything to do with how model accuracy is only a subset of model performance, and at that, a sometimes misleading one. For example, if you wanted to detect fraud in a massive dataset with a sample of millions, a more accurate model would most likely predict no fraud at all if only a vast minority of cases were fraud. However, this would be useless for a predictive model—a model designed to find fraud that asserted there was no fraud at all! Questions like this help you demonstrate that you understand model accuracy isn't the be-all and end-all of model performance.

Q18: What's the F1 score? How would you use it?

Answer: The F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst. You would use it in classification tests where true negatives don't matter much.

Q19: How would you handle an imbalanced dataset?

Answer: 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:

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

What's important here is that you have a keen sense for what damage an unbalanced dataset can cause, and how to balance that.

Q20: When should you use classification over regression?

Answer: Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points. You would use classification over regression if you wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)

Q21: Name an example where ensemble techniques might be useful.

Answer: Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data).

You could list some examples of ensemble methods (bagging, boosting, the "bucket of models" method) and demonstrate how they could increase predictive power.

Q22: How do you ensure you're not overfitting with a model?

Answer: This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.

There are three main methods to avoid overfitting:

1. Keep the model simpler: reduce variance by taking into account fewer variables and parameters, 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.

Q23: What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

Answer: 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: here is a fairly comprehensive list. 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.

Q24: How would you evaluate a logistic regression model?

Answer: A subsection of the question above. You have to demonstrate an understanding of what the typical goals of a logistic regression are (classification, prediction, etc.) and bring up a few examples and use cases.

Q25: What's the "kernel trick" and how is it useful?

Answer: 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.

Machine Learning Interview Questions: Programming

Q26: How do you handle missing or corrupted data in a dataset?

Answer: You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.

In Pandas, there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

Q27: Do you have experience with Spark or big data tools for machine learning?

Answer: You'll want to get familiar with the meaning of big data for different companies and the different tools they'll want. Spark is the big data tool most in demand now, able to handle immense datasets with speed. Be honest if you don't have experience with the tools demanded, but also take a look at job descriptions and see what tools pop up: you'll want to invest in familiarizing yourself with them.

Q28: Pick an algorithm. Write the pseudo-code for a parallel implementation.

Answer: This kind of question demonstrates your ability to think in parallelism and how you could handle concurrency in programming implementations dealing with big data. Take a look at pseudocode frameworks such as Peril-L and visualization tools such as Web Sequence Diagrams to help you demonstrate your ability to write code that reflects parallelism.

Q29: What are some differences between a linked list and an array?

Answer: An array is an ordered collection of objects. A linked list is a series of objects with pointers that direct how to process them sequentially. An array assumes that every element has the same size, unlike the linked list. A linked list can more easily grow organically: an array has to be pre-defined or re-defined for organic growth. Shuffling a linked list involves changing which points direct where—meanwhile, shuffling an array is more complex and takes more memory.

Q30: Describe a hash table.

Answer: A hash table is a data structure that produces an associative array. A key is mapped to certain values through the use of a hash function. They are often used for tasks such as database indexing.

Q31: Which data visualization libraries do you use? What are your thoughts on the best data visualization tools?

Answer: What's important here is to define your views on how to properly visualize data and your personal preferences when it comes to tools. Popular tools include R's ggplot, Python's seaborn and matplotlib, and tools such as Plot.ly and Tableau.

Q32: Given two strings, A and B, of the same length n, find whether it is possible to cut both strings at a common point such that the first part of A and the second part of B form a palindrome.

Answer: You'll often get standard algorithms and data structures questions as part of your interview process as a machine learning engineer that might feel akin to a software engineering interview. In this case, this comes from Google's interview process. There are multiple ways to check for palindromes—one way of doing so if you're using a programming language such as Python is to reverse the string and check to see if it still equals the original string, for example. The thing to look out for here is the category of questions you can expect, which will be akin to software engineering questions that drill down to your knowledge of algorithms and data structures. Make sure that you're totally comfortable with the language of your choice to express that logic.

Q33: How are primary and foreign keys related in SQL?

Answer: Most machine learning engineers are going to have to be conversant with a lot of different data formats. SQL is still one of the key ones used. Your ability to understand how to manipulate SQL databases will be something you'll most likely need to demonstrate. In this example, you can talk about how foreign keys allow you to match up and join tables together on the primary key of the corresponding table—but just as useful is to talk through how you would think about setting up SQL tables and querying them.

Q34: How does XML and CSVs compare in terms of size?

Answer: In practice, XML is much more verbose than CSVs are and takes up a lot more space. CSVs use some separators to categorize and organize data into neat columns. XML uses tags to delineate a tree-like structure for key-value pairs. You'll often get XML back as a way to semi-structure data from APIs or HTTP responses. In practice, you'll want to ingest XML data and try to process it into a usable CSV. This sort of question tests your familiarity with data wrangling sometimes messy data formats.

Q35: What are the data types supported by JSON?

Answer: This tests your knowledge of JSON, another popular file format that wraps with JavaScript. There are six basic JSON datatypes you can manipulate: strings, numbers, objects, arrays, booleans, and null values.

Q36: How would you build a data pipeline?

Answer: Data pipelines are the bread and butter of machine learning engineers, who take data science models and find ways to automate and scale them. Make sure you're familiar with the tools to build data pipelines (such as Apache Airflow) and the platforms where you can host models and pipelines (such as Google Cloud or AWS or Azure). Explain the steps required in a functioning data pipeline and talk through your actual experience building and scaling them in production.

Machine Learning Interview Questions: Company/Industry Specific

Q37: What do you think is the most valuable data in our business?

Answer: This question or questions like it really try to test you on two dimensions. The first is your knowledge of the business and the industry itself, as well as your understanding of the business model. The second is whether you can pick how correlated data is to business outcomes in general, and then how you apply that thinking to your context about the company. You'll want to research the business model and ask good questions to your recruiter—and start thinking about what business problems they probably want to solve most with their data.

Q38: How would you implement a recommendation system for our company's users?

Answer: A lot of machine learning interview questions of this type will involve the implementation of machine learning models to a company's problems. You'll have to research the company and its industry in-depth, especially the revenue drivers the company has, and the types of users the company takes on in the context of the industry it's in.

Q39: How can we use your machine learning skills to generate revenue?

Answer: This is a tricky question. The ideal answer would demonstrate knowledge of what drives the business and how your skills could relate. For example, if you were interviewing for music-streaming startup Spotify, you could remark that your skills at developing a better recommendation model would increase user retention, which would then increase revenue in the long run.

The startup metrics Slideshare linked above will help you understand exactly what performance indicators are important for startups and tech companies as they think about revenue and growth.

Q40: What do you think of our current data process?

Answer: This kind of question requires you to listen carefully and impart feedback in a manner that is constructive and insightful. Your interviewer is trying to gauge if you'd be a valuable member of their team and whether you grasp the nuances of why certain things are set the way they are in the company's data process based on company or industry-specific conditions. They're trying to see if you can be an intellectual peer. Act accordingly.

Q41: What are the last machine learning papers you've read?

Answer: Keeping up with the latest scientific literature on machine learning is a must if you want to demonstrate an interest in a machine learning position. This overview of deep learning in Nature by the scions

of deep learning themselves (from Hinton to Bengio to LeCun) can be a good reference paper and an overview of what's happening in deep learning — and the kind of paper you might want to cite.

Q42: Do you have research experience in machine learning?

Answer: Related to the last point, most organizations hiring for machine learning positions will look for your formal experience in the field. Research papers, co-authored or supervised by leaders in the field, can make the difference between you being hired and not. Make sure you have a summary of your research experience and papers ready—and an explanation for your background and lack of formal research experience if you don't.

Q43: What are your favorite use cases of machine learning models?

Answer: The Quora thread below contains some examples, such as decision trees that categorize people into different tiers of intelligence based on IQ scores. Make sure that you have a few examples in mind and describe what resonated with you. It's important that you demonstrate an interest in how machine learning is implemented.

Q44: How would you approach the "Netflix Prize" competition?

Answer: The Netflix Prize was a famed competition where Netflix offered $1,000,000 for a better collaborative filtering algorithm. The team that won called BellKor had a 10% improvement and used an ensemble of different methods to win. Some familiarity with the case and its solution will help demonstrate you've paid attention to machine learning for a while.

Q45: Where do you usually source datasets?

Answer: Machine learning interview questions like these try to get at the heart of your machine learning interest. Somebody who is truly passionate about machine learning will have gone off and done side projects on their own, and have a good idea of what great datasets are out there. If you're missing any, check out Quandl for economic and financial data, and Kaggle's Datasets collection for another great list.

Q46: How do you think Google is training data for self-driving cars?

Answer: Machine learning interview questions like this one really test your knowledge of different machine learning methods, and your inventiveness if you don't know the answer. Google is currently using recaptcha to source labeled data on storefronts and traffic signs. They are also building on training data collected by Sebastian Thrun at GoogleX—some of which was obtained by his grad students driving buggies on desert dunes!

Q47: How would you simulate the approach AlphaGo took to beat Lee Sedol at Go?

Answer: AlphaGo beating Lee Sedol, the best human player at Go, in a best-of-five series was a truly seminal event in the history of machine learning and deep learning. The Nature paper above describes how this was accomplished with "Monte-Carlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play."

Q48: What are your thoughts on GPT-3 and OpenAI's model?

Answer: GPT-3 is a new language generation model developed by OpenAI. It was marked as exciting because with very little change in architecture, and a ton more data, GPT-3 could generate what seemed to be human-like conversational pieces, up to and including novel-size works and the ability to create code from natural

language. There are many perspectives on GPT-3 throughout the Internet — if it comes up in an interview setting, be prepared to address this topic (and trending topics like it) intelligently to demonstrate that you follow the latest advances in machine learning.

Q49: What models do you train for fun, and what GPU/hardware do you use?

Answer: Such machine learning interview questions tests whether you've worked on machine learning projects outside of a corporate role and whether you understand the basics of how to resource projects and allocate GPU-time efficiently. Expect questions like this to come from hiring managers that are interested in getting a greater sense behind your portfolio, and what you've done independently.

Q50: What are some of your favorite APIs to explore?

Answer: If you've worked with external data sources, it's likely you'll have a few favorite APIs that you've gone through. You can be thoughtful here about the kinds of experiments and pipelines you've run in the past, along with how you think about the APIs you've used before.

Q51: How do you think quantum computing will affect machine learning?

Answer: With the recent announcement of more breakthroughs in quantum computing, the question of how this new format and way of thinking through hardware serves as a useful proxy to explain classical computing and machine learning, and some of the hardware nuances that might make some algorithms much easier to do on a quantum machine. Demonstrating some knowledge in this area helps show that you're interested in machine learning at a much higher level than just implementation details.