Top 100 Machine Learning Questions & Answers

Steve Nouri

Q1 Explain the difference between supervised and unsupervised machine learning?

In supervised machine learning algorithms, we have to provide labeled data, for example, prediction of stock market prices, whereas in unsupervised we need not have labeled data, for example, classification of emails into spam and non-spam.

Q2 What are the parametric models? Give an example.

Parametric models are those with a finite number of parameters. To predict new data, you only need to know the parameters of the model. Examples include linear regression, logistic regression, and linear SVMs.

Non-parametric models are those with an unbounded number of parameters, allowing for more flexibility. To predict new data, you need to know the parameters of the model and the state of the data that has been observed. Examples include decision trees, k-nearest neighbors, and topic models using latent Dirichlet analysis.

Q3 What is the difference between classification and regression?

Classification is used to produce discrete results, classification is used to classify data into some specific categories. For example, classifying emails into spam and non-spam categories.

Whereas, We use regression analysis when we are dealing with continuous data, for example predicting stock prices at a certain point in time.

Q4 What Is Overfitting, and How Can You Avoid It?

Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data.

When a model is given the training data, it shows 100 percent accuracy—technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting.

There are multiple ways of avoiding overfitting, such as:

- Regularization. It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and parameters, the variance can be reduced
- Cross-validation methods like k-folds can also be used
- If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

Q5 What is meant by 'Training set' and 'Test Set'?

We split the given data set into two different sections namely,'Training set' and 'Test Set'.
'Training set' is the portion of the dataset used to train the model.
'Testing set' is the portion of the dataset used to test the trained model.

Q6 How Do You Handle Missing or Corrupted Data in a Dataset?

One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or replace them entirely with some other value.
There are two useful methods in Pandas:

- IsNull() and dropna() will help to find the columns/rows with missing data and drop them
- Fillna() will replace the wrong values with a placeholder value

Q7 Explain Ensemble learning.

In ensemble learning, many base models like classifiers and regressors are generated and combined together so that they give better results. It is used when we build component classifiers that are accurate and independent. There are sequential as well as parallel ensemble methods.

Q8 Explain the Bias-Variance Tradeoff.

Predictive models have a tradeoff between bias (how well the model fits the data) and variance (how much the model changes based on changes in the inputs).
Simpler models are stable (low variance) but they don't get close to the truth (high bias).
More complex models are more prone to overfitting (high variance) but they are expressive enough to get close to the truth (low bias). The best model for a given problem usually lies somewhere in the middle.

Q9 What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

Both algorithms are methods for finding a set of parameters that minimize a loss function by evaluating parameters against data and then making adjustments.
In standard gradient descent, you'll evaluate all training samples for each set of parameters. This is akin to taking big, slow steps toward the solution.
In stochastic gradient descent, you'll evaluate only 1 training sample for the set of parameters before updating them. This is akin to taking small, quick steps toward the solution.

Q10 How Can You Choose a Classifier Based on a Training Set Data Size?

When the training set is small, a model that has a right bias and low variance seems to work better because they are less likely to overfit.

For example, Naive Bayes works best when the training set is large. Models with low bias and high variance tend to perform better as they work fine with complex relationships.

Q11 What are 3 data preprocessing techniques to handle outliers?

1. Winsorize (cap at threshold).
2. Transform to reduce skew (using Box-Cox or similar).
3. Remove outliers if you're certain they are anomalies or measurement errors.

Q12 How much data should you allocate for your training, validation, and test sets?

You have to find a balance, and there's no right answer for every problem.
If your test set is too small, you'll have an unreliable estimation of model performance (performance statistic will have high variance). If your training set is too small, your actual model parameters will have a high variance.
A good rule of thumb is to use an 80/20 train/test split. Then, your train set can be further split into train/validation or into partitions for cross-validation.

Q13 What Is a False Positive and False Negative and How Are They Significant?

False positives are those cases which wrongly get classified as True but are False.
False negatives are those cases which wrongly get classified as False but are True.
In the term 'False Positive,' the word 'Positive' refers to the 'Yes' row of the predicted value in the confusion matrix. The complete term indicates that the system has predicted it as a positive, but the actual value is negative.

Q14 Explain the difference between L1 and L2 regularization.

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 to the terms, while L2 corresponds to a Gaussian prior.

Q15 What's a Fourier transform?

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.

Q16 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.

Q17 What's the difference between a generative and discriminative model?

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.

Q18 What Are the Applications of Supervised Machine Learning in Modern Businesses?

Applications of supervised machine learning include:

- **Email Spam Detection**
  Here we train the model using historical data that consists of emails categorized as spam or not spam. This labeled information is fed as input to the model.
- **Healthcare Diagnosis**
  By providing images regarding a disease, a model can be trained to detect if a person is suffering from the disease or not.
- **Sentiment Analysis**
  This refers to the process of using algorithms to mine documents and determine whether they're positive, neutral, or negative in sentiment.
- **Fraud Detection**
  Training the model to identify suspicious patterns, we can detect instances of possible fraud.

Q19 What Is Semi-supervised Machine Learning?

Supervised learning uses data that is completely labeled, whereas unsupervised learning uses no training data.
In the case of semi-supervised learning, the training data contains a small amount of labeled data and a large amount of unlabeled data.

Q20. What Are Unsupervised Machine Learning Techniques?

There are two techniques used in unsupervised learning: clustering and association.

Clustering

- Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.

Association

- In an association problem, we identify patterns of associations between different variables or items.
- For example, an eCommerce website can suggest other items for you to buy, based on the prior purchases that you have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.

Q21 What Is 'naive' in the Naive Bayes Classifier?

The classifier is called 'naive' because it makes assumptions that may or may not turn out to be correct.

The algorithm assumes that the presence of one feature of a class is not related to the presence of any other feature (absolute independence of features), given the class variable.

For instance, a fruit may be considered to be a cherry if it is red in color and round in shape, regardless of other features. This assumption may or may not be right (as an apple also matches the description).

Q22 Explain Latent Dirichlet Allocation (LDA).

Latent Dirichlet Allocation (LDA) is a common method of topic modeling, or classifying documents by subject matter.

LDA is a generative model that represents documents as a mixture of topics that each have their own probability distribution of possible words.

The "Dirichlet" distribution is simply a distribution of distributions. In LDA, documents are distributions of topics that are distributions of words.

Q23 Explain Principle Component Analysis (PCA).

PCA is a method for transforming features in a dataset by combining them into uncorrelated linear combinations.

These new features, or principal components, sequentially maximize the variance represented (i.e. the first principal component has the most variance, the second principal component has the second most, and so on).

As a result, PCA is useful for dimensionality reduction because you can set an arbitrary variance cutoff.

Q24 What's the F1 score? How would you use it?

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.

Q25 When should you use classification over regression?

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.)

Q26 How do you ensure you're not overfitting with a model?

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.

Q27 How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:

- If accuracy is a concern, test different algorithms and cross-validate them
- If the training dataset is small, use models that have low variance and high bias
- If the training dataset is large, use models that have high variance and little bias

Q28 How Do You Design an Email Spam Filter?

Building a spam filter involves the following process:

- The email spam filter will be fed with thousands of emails
- Each of these emails already has a label: 'spam' or 'not spam.'
- The supervised machine learning algorithm will then determine which type of emails are being marked as spam based on spam words like the lottery, free offer, no money, full refund, etc.
- The next time an email is about to hit your inbox, the spam filter will use statistical analysis and algorithms like Decision Trees and SVM to determine how likely the email is spam
- If the likelihood is high, it will label it as spam, and the email won't hit your inbox

- Based on the accuracy of each model, we will use the algorithm with the highest accuracy after testing all the models

Q29 What evaluation approaches would you work to gauge the effectiveness of a machine learning 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: 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.

Q30 How would you implement a recommendation system for our company's users?

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.

Q31 Explain bagging.

Bagging, or Bootstrap Aggregating, is an ensemble method in which the dataset is first divided into multiple subsets through resampling.
Then, each subset is used to train a model, and the final predictions are made through voting or averaging the component models.
Bagging is performed in parallel.

Q32 What is the ROC Curve and what is AUC (a.k.a. AUROC)?

The ROC (receiver operating characteristic) the performance plot for binary classifiers of True Positive Rate (y-axis) vs. False Positive Rate (x-
axis).
AUC is the area under the ROC curve, and it's a common performance metric for evaluating binary classification models.
It's equivalent to the expected probability that a uniformly drawn random positive is ranked before a uniformly drawn random negative.

Q33 Why is Area Under ROC Curve (AUROC) better than raw accuracy as an out-of-sample evaluation metric?

AUROC is robust to class imbalance, unlike raw accuracy.
For example, if you want to detect a type of cancer that's prevalent in only 1% of the population, you can build a model that achieves 99% accuracy by simply classifying everyone has cancer-free.

Q34 What are the advantages and disadvantages of neural networks?

Advantages: Neural networks (specifically deep NNs) have led to performance breakthroughs for unstructured datasets such as images, audio, and video. Their incredible flexibility allows them to learn patterns that no other ML algorithm can learn.
Disadvantages: However, they require a large amount of training data to converge. It's also difficult to pick the right architecture, and the internal "hidden" layers are incomprehensible.

Q35 Define Precision and Recall.

Precision

- Precision is the ratio of several events you can correctly recall to the total number of events you recall (mix of correct and wrong recalls).
- Precision = (True Positive) / (True Positive + False Positive)

Recall

- A recall is the ratio of a number of events you can recall the number of total events.
- Recall = (True Positive) / (True Positive + False Negative)

Q36 What Is Decision Tree Classification?

A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.

Q37 What Is Pruning in Decision Trees, and How Is It Done?

Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.
Pruning can occur in:

- Top-down fashion. It will traverse nodes and trim subtrees starting at the root
- Bottom-up fashion. It will begin at the leaf nodes

There is a popular pruning algorithm called reduced error pruning, in which:

- Starting at the leaves, each node is replaced with its most popular class
- If the prediction accuracy is not affected, the change is kept
- There is an advantage of simplicity and speed

Q38 What Is a Recommendation System?

Anyone who has used Spotify or shopped at Amazon will recognize a recommendation system: It's an information filtering system that predicts what a user might want to hear or see based on choice patterns provided by the user.

Q39 What Is Kernel SVM?

Kernel SVM is the abbreviated version of the kernel support vector machine. Kernel methods are a class of algorithms for pattern analysis, and the most common one is the kernel SVM.

Q40 What Are Some Methods of Reducing Dimensionality?

You can reduce dimensionality by combining features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction.

Now that you have gone through these machine learning interview questions, you must have got an idea of your strengths and weaknesses in this domain.

Q41 What Are the Three Stages of Building a Model in Machine Learning?

The three stages of building a machine learning model are:

- **Model Building** Choose a suitable algorithm for the model and train it according to the requirement
- **Model Testing** Check the accuracy of the model through the test data
- **Applying the Mode** Make the required changes after testing and use the final model for real-time projects. Here, it's important to remember that once in a while, the model needs to be checked to make sure it's working correctly. It should be modified to make sure that it is up-to-date.

Q42 How is KNN different from k-means clustering?

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.

Q43 Mention the difference between Data Mining and Machine learning?

Machine learning relates to the study, design, and development of the algorithms that give computers the capability to learn without being explicitly programmed. While data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns. During this processing machine, learning algorithms are used.

Q44 What are the different Algorithm techniques in Machine Learning?

The different types of techniques in Machine Learning are

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning
- Transduction
- Learning to Learn

Q45 You are given a data set. The data set has missing values that spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

This question has enough hints for you to start thinking! Since the data is spread across the median, let's assume it's a normal distribution. We know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

Q46 What are PCA, KPCA, and ICA used for?

PCA (Principal Components Analysis), KPCA (Kernel-based Principal Component Analysis) and ICA (Independent Component Analysis) are important feature extraction techniques used for dimensionality reduction.

Q47 What are support vector machines?

Support vector machines are supervised learning algorithms used for classification and regression analysis.

Q48 What is batch statistical learning?

Statistical learning techniques allow learning a function or predictor from a set of observed data that can make predictions about unseen or future data. These techniques provide guarantees on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

Q49 What is the bias-variance decomposition of classification error in the ensemble method?

The expected error of a learning algorithm can be decomposed into bias and variance. A bias term measures how closely the average classifier produced by the learning algorithm matches the target function. The variance term measures how much the learning algorithm's prediction fluctuates for different training sets.

Q50 When is Ridge regression favorable over Lasso regression?

You can quote ISLR's authors Hastie, Tibshirani who asserted that, in the presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small/medium-sized effects, use ridge regression.

Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In the presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.

Q51 You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?

The model has overfitted. Training error 0.00 means the classifier has mimicked the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on an unseen sample, it couldn't find those patterns and returned predictions with higher error. In a random forest, it happens when we use a larger number of trees than necessary. Hence, to avoid this situation, we should tune the number of trees using cross-validation.

Q50 What is a convex hull?

In the case of linearly separable data, the convex hull represents the outer boundaries of the two groups of data points. Once the convex hull is created, we get maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to create the greatest separation between two groups.

Q51 What do you understand by Type I vs Type II error?

Type I error is committed when the null hypothesis is true and we reject it, also known as a 'False Positive'. Type II error is committed when the null hypothesis is false and we accept it, also known as 'False Negative'.

In the context of the confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).

Q52. In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance?

We don't use manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, the euclidean metric can be used in any space to calculate distance. Since the data points can be present in any dimension, euclidean distance is a more viable option.

Example: Think of a chessboard, the movement made by a bishop or a rook is calculated by manhattan distance because of their respective vertical & horizontal movements.

Q53 Do you suggest that treating a categorical variable as a continuous variable would result in a better predictive model?

For better predictions, the categorical variable can be considered as a continuous variable only when the variable is ordinal in nature.

Q54 OLS is to linear regression. The maximum likelihood is logistic regression. Explain the statement.

OLS and Maximum likelihood are the methods used by the respective regression methods to approximate the unknown parameter (coefficient) value. In simple words,
Ordinary least square(OLS) is a method used in linear regression which approximates the parameters resulting in minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.

Q55 When does regularization becomes necessary in Machine Learning?

Regularization becomes necessary when the model begins to overfit/underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce the cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

Q56 What is Linear Regression?

Linear Regression is a supervised Machine Learning algorithm. It is used to find the linear relationship between the dependent and the independent variables for predictive analysis.

Q57 What is the Variance Inflation Factor?

Variance Inflation Factor (VIF) is the estimate of the volume of multicollinearity in a collection of many regression variables.
VIF = Variance of the model / Variance of the model with a single independent variable
We have to calculate this ratio for every independent variable. If VIF is high, then it shows the high collinearity of the independent variables.

Q58 We know that one hot encoding increases the dimensionality of a dataset, but label encoding doesn't. How?

When we use **one-hot encoding**, there is an increase in the dimensionality of a dataset. The reason for the increase in dimensionality is that, for every class in the categorical variables, it forms a different variable.

Q59 What is a Decision Tree?

A decision tree is used to explain the sequence of actions that must be performed to get the desired output. It is a hierarchical diagram that shows the actions.

Q60 What is the Binarizing of data? How to Binarize?

In most of the Machine Learning Interviews, apart from theoretical questions, interviewers focus on the implementation part. So, this ML Interview Questions focused on the implementation of the theoretical concepts.

Converting data into binary values on the basis of threshold values is known as the binarizing of data. The values that are less than the threshold are set to 0 and the values that are greater than the threshold are set to 1. This process is useful when we have to perform feature engineering, and we can also use it for adding unique features.

Q61 What is cross-validation?

Cross-validation is essentially a technique used to assess how well a model performs on a new independent dataset. The simplest example of cross-validation is when you split your data into two groups: training data and testing data, where you use the training data to build the model and the testing data to test the model.

Q62 When would you use random forests Vs SVM and why?

There are a couple of reasons why a random forest is a better choice of the model than a support vector machine:

- Random forests allow you to determine the feature importance. SVM's can't do this.
- Random forests are much quicker and simpler to build than an SVM.
- For multi-class classification problems, SVMs require a one-vs-rest method, which is less scalable and more memory intensive.

Q63 What are the drawbacks of a linear model?

There are a couple of drawbacks of a linear model:

- A linear model holds some strong assumptions that may not be true in the application. It assumes a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation, and homoscedasticity
- A linear model can't be used for discrete or binary outcomes.
- You can't vary the model flexibility of a linear model.

Q64 Do you think 50 small decision trees are better than a large one? Why?

Another way of asking this question is “Is a random forest a better model than a decision tree?” And the answer is yes because a random forest is an ensemble method that takes many weak decision trees to make a strong learner. Random forests are more accurate, more robust, and less prone to overfitting.

Q65 What is a kernel? Explain the kernel trick

A kernel is a way of computing the dot product of two vectors $\mathbf{x}x$ and $\mathbf{y}y$ in some (possibly very high dimensional) feature space, which is why kernel functions are sometimes called “generalized dot product”
The kernel trick is a method of using a linear classifier to solve a non-linear problem by transforming linearly inseparable data to linearly separable ones in a higher dimension.

Q66 State the differences between causality and correlation?

Causality applies to situations where one action, say X, causes an outcome, say Y, whereas Correlation is just relating one action (X) to another action(Y) but X does not necessarily cause Y.

Q67 What is the exploding gradient problem while using the backpropagation technique?

When large error gradients accumulate and result in large changes in the neural network weights during training, it is called the exploding gradient problem. The values of weights can become so large as to overflow and result in NaN values. This makes the model unstable and the learning of the model to stall just like the vanishing gradient problem.

Q68 What do you mean by Associative Rule Mining (ARM)?

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features (dimensions) which are correlated.

Q69 What is Marginalisation? Explain the process.

Marginalizationarginalisation is summing the probability of a random variable X given the joint probability distribution of X with other variables. It is an application of the law of total probability.

Q70 Why is the rotation of components so important in Principle Component Analysis(PCA)?

Rotation in PCA is very important as it maximizes the separation within the variance obtained by all the components because of which interpretation of components would become easier. If the components are not rotated, then we need extended components to describe the variance of the components.

Q71 What is the difference between regularization and normalisation?

Normalisation adjusts the data; regularisation adjusts the prediction function. If your data is on very different scales (especially low to high), you would want to normalise the data. Alter each column to have compatible basic statistics. This can be helpful to make sure there is no loss of accuracy. One of the goals of model training is to identify the signal and ignore the noise if the model is given free rein to minimize error, there is a possibility of suffering from overfitting. Regularization imposes some control on this by providing simpler fitting functions over complex ones.

Q72 When does the linear regression line stop rotating or finds an optimal spot where it is fitted on data?

A place where the highest RSquared value is found, is the place where the line comes to rest. RSquared represents the amount of variance captured by the virtual linear regression line with respect to the total variance captured by the dataset.

Q73 How does the SVM algorithm deal with self-learning?

SVM has a learning rate and expansion rate which takes care of this. The learning rate compensates or penalises the hyperplanes for making all the wrong moves and expansion rate deals with finding the maximum separation area between classes.

Q74 How do you handle outliers in the data?

Outlier is an observation in the data set that is far away from other observations in the data set. We can discover outliers using tools and functions like box plot, scatter plot, Z-Score, IQR score etc. and then handle them based on the visualization we have got. To handle outliers, we can cap at some threshold, use transformations to reduce skewness of the data and remove outliers if they are anomalies or errors.

Q75 Name and define techniques used to find similarities in the recommendation system.

Pearson correlation and Cosine correlation are techniques used to find similarities in recommendation systems.

Q76 Why would you Prune your tree?

In the context of data science or AIML, pruning refers to the process of reducing redundant branches of a decision tree. Decision Trees are prone to overfitting, pruning the tree helps to reduce the size and minimizes the chances of overfitting. Pruning involves turning branches of a decision tree into leaf nodes and removing the leaf nodes from the original branch. It serves as a tool to perform the tradeoff.

Q77 Mention some of the EDA Techniques?

Exploratory Data Analysis (EDA) helps analysts to understand the data better and forms the foundation of better models.

Visualization

- Univariate visualization
- Bivariate visualization
- Multivariate visualization

Missing Value Treatment – Replace missing values with Either Mean/Median

Outlier Detection – Use Boxplot to identify the distribution of Outliers, then Apply IQR to set the boundary for IQR

Q78 What is data augmentation? Can you give some examples?

Data augmentation is a technique for synthesizing new data by modifying existing data in such a way that the target is not changed, or it is changed in a known way.

CV is one of the fields where data augmentation is very useful. There are many modifications that we can do to images:

- Resize
- Horizontal or vertical flip
- Rotate
- Add noise
- Deform
- Modify colors

Each problem needs a customized data augmentation pipeline. For example, on OCR, doing flips will change the text and won't be beneficial; however, resizes and small rotations may help.

Q79 What is Inductive Logic Programming in Machine Learning (ILP)?

Inductive Logic Programming (ILP) is a subfield of machine learning which uses logic programming representing background knowledge and examples.

Q80 What is the difference between inductive machine learning and deductive machine learning?

The difference between inductive machine learning and deductive machine learning are as follows: machine-learning where the model learns by examples from a set of observed instances to draw a generalized conclusion whereas in deductive learning the model first draws the conclusion and then the conclusion is drawn.

Q81 Difference between machine learning and deep learning

Machine learning is a branch of computer science and a method to implement artificial intelligence. This technique provides the ability to automatically learn and improve from experiences without being explicitly programmed.

Deep learning can be said as a subset of machine learning. It is mainly based on the artificial neural network where data is taken as an input and the technique makes intuitive decisions using the artificial neural network.

Q82 What Are The Steps Involved In Machine Learning Project?

As you plan for doing a machine learning project. There are several important steps you must follow to achieve a good working model and they are data collection, data preparation, choosing a machine learning model, training the model, model evaluation, parameter tuning and lastly prediction.

Q83 Differences between Artificial Intelligence and Machine Learning?

Artificial intelligence is a broader prospect than machine learning. Artificial intelligence mimics the cognitive functions of the human brain. The purpose of AI is to carry out a task in an intelligent manner based on algorithms. On the other hand, machine learning is a subclass of artificial intelligence. To develop an autonomous machine in such a way so that it can learn without being explicitly programmed is the goal of machine learning.

Q84 Steps Needed to Choose the Appropriate Machine Learning Algorithm for your Classification problem.

Firstly, you need to have a clear picture of your data, your constraints, and your problems before heading towards different machine learning algorithms. Secondly, you have to understand which type and kind of data you have because it plays a primary role in deciding which algorithm you have to use.

Following this step is the data categorization step, which is a two-step process – categorization by input and categorization by output. The next step is to understand your constraints; that is, what is your data storage capacity? How fast the prediction has to be? etc.

Finally, find the available machine learning algorithms and implement them wisely. Along with that, also try to optimize the hyperparameters which can be done in three ways – grid search, random search, and Bayesian optimization.

Q85 Explain Backpropagation in Machine Learning.

A very important question for your machine learning interview. **Backpropagation** is the algorithm for computing artificial neural networks (ANN). It is used by the gradient descent optimization that exploits the chain rule. By calculating the gradient of the loss function, the weight of the neurons is adjusted to a certain value. To train a multi-layered neural network is the prime motivation of backpropagation so that it can learn the appropriate internal demonstrations. This will help them learn to map any input to its respective output arbitrarily.

Q86 What is the Convex Function?

This question is very often asked in machine learning interviews. A convex function is a continuous function, and the value of the midpoint at every interval in its given domain is less than the numerical mean of the values at the two ends of the interval.

Q87 What's the Relationship between True Positive Rate and Recall?

The True positive rate in machine learning is the percentage of the positives that have been properly acknowledged, and recall is just the count of the results that have been correctly identified and are relevant. Therefore, they are the same things, just having different names. It is also known as sensitivity.

Q88 List some Tools for Parallelizing Machine Learning Algorithms.

Although this question may seem very easy, make sure not to skip this one because it is also very closely related to artificial intelligence and thereby, AI interview questions. Almost all machine learning algorithms are easy to serialize. Some of the basic tools for parallelizing are Matlab, Weka, R, Octave, or the Python-based sci-kit learn.

Q89 What do you mean by Genetic Programming?

Genetic Programming (GP) is almost similar to an Evolutionary Algorithm, a subset of machine learning. Genetic programming software systems implement an algorithm that uses random mutation, a fitness function, crossover, and multiple generations of evolution to resolve a user-defined task. The genetic programming model is based on testing and choosing the best option among a set of results.

Q90 What do you know about Bayesian Networks?

Bayesian Networks also referred to as 'belief networks' or 'casual networks', are used to represent the graphical model for probability relationship among a set of variables.
For example, a Bayesian network can be used to represent the probabilistic relationships between diseases and symptoms. As per the symptoms, the network can also compute the probabilities of the presence of various diseases.

Efficient algorithms can perform inference or learning in Bayesian networks. Bayesian networks which relate the variables (e.g., speech signals or protein sequences) are called dynamic Bayesian networks.

Q91 Which are the two components of the Bayesian logic program?

A Bayesian logic program consists of two components:

- **Logical** It contains a set of Bayesian Clauses, which capture the qualitative structure of the domain.
- **Quantitative** It is used to encode quantitative information about the domain.

Q92 How is machine learning used in day-to-day life?

Most of the people are already using machine learning in their everyday life. Assume that you are engaging with the internet, you are actually expressing your preferences, likes, dislikes through your searches. All these things are picked up by cookies coming on your computer, from this, the behavior of a user is evaluated. It helps to increase the progress of a user through the internet and provide similar suggestions.

The navigation system can also be considered as one of the examples where we are using machine learning to calculate a distance between two places using optimization techniques. Surely, people are going to more engage with machine learning in the near future

Q93 Define Sampling. Why do we need it?

Answer: Sampling is a process of choosing a subset from a target population that would serve as its representative. We use the data from the sample to understand the pattern in the community as a whole. Sampling is necessary because often, we can not gather or process the complete data within a reasonable time.

Q94 What does the term decision boundary mean?

Answer: A decision boundary or a decision surface is a hypersurface which divides the underlying feature space into two subspaces, one for each class. If the decision boundary is a hyperplane, then the classes are linearly separable.

Q95 Define entropy?

Answer: Entropy is the measure of uncertainty associated with random variable Y. It is the expected number of bits required to communicate the value of the variable.

Q96 Indicate the top intents of machine learning?

Answer: The top intents of machine learning are stated below,

- The system gets information from the already established computations to give well-founded decisions and outputs.
- It locates certain patterns in the data and then makes certain predictions on it to provide answers on matters.

Q97 Highlight the differences between the Generative model and the Discriminative model?

The aim of the Generative model is to generate new samples from the same distribution and new data instances, Whereas, the Discriminative model highlights the differences between different kinds of data instances. It tries to learn directly from the data and then classifies the data.

Q98 Identify the most important aptitudes of a machine learning engineer?

Machine learning allows the computer to learn itself without being decidedly programmed. It helps the system to learn from experience and then improve from its mistakes. The intelligence system, which is based on machine learning, can learn from recorded data and past incidents. In-depth knowledge of statistics, probability, data modelling, programming language, as well as CS, Application of ML Libraries and algorithms, and software design is required to become a successful machine learning engineer.

Q99 What is feature engineering? How do you apply it in the process of modelling?

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Q100 How can learning curves help create a better model?

Learning curves give the indication of the presence of overfitting or underfitting.
In a learning curve, the training error and cross-validating error are plotted against the number of training data points.