**Machine Learning Interview Questions and Answers**

**Core Machine Learning Questions**

1. **How would you explain machine learning to a school-going kid?**
   * **Machine learning involves teaching a machine to recognize patterns and make decisions. For example, at a party with strangers, you might classify people based on their appearance or behavior, which is like unsupervised learning. If you have prior knowledge about some guests, it resembles supervised learning.**
2. **What are the types of machine learning?**
   * **Supervised Learning: Learning with labeled data, e.g., identifying apples and bananas with labels.**
   * **Unsupervised Learning: Discovering hidden patterns in unlabeled data, e.g., clustering fruits into groups based on similarity.**
   * **Reinforcement Learning: Learning through rewards and penalties, like playing a video game and improving based on feedback.**
3. **Explain supervised learning with an example.**
   * **Example: Labeling images of apples and bananas to teach a model to differentiate between them. The model uses this knowledge to classify new images.**
4. **What is unsupervised learning?**
   * **Learning without labels. For example, grouping similar fruits (apples, bananas, mangoes) into clusters without knowing their names.**
5. **What is reinforcement learning?**
   * **Learning through interaction with the environment. Example: Playing a game where actions like moving, jumping, or firing are rewarded or penalized.**

**Scenario-Based Questions**

1. **What would you do if given a dataset with 30% missing values?**
   * **Options include:**
     + **Creating a binary variable to indicate missing values.**
     + **Removing the rows/columns with missing data if it doesn't affect the analysis.**
     + **Imputing missing values based on clustering, distributions, or business rules.**
2. **How would you handle outliers in a dataset?**
   * **Use box plots, statistical methods, or clustering algorithms to detect outliers.**
   * **Options for handling:**
     + **Remove outliers if data volume is sufficient.**
     + **Cap extreme values using percentiles.**
     + **Impute values based on patterns or domain knowledge.**

**Practical Machine Learning Questions**

1. **What is overfitting and how can it be avoided?**
   * **Overfitting: The model performs well on training data but poorly on unseen data.**
   * **Solutions:**
     + **Collect more data.**
     + **Use simpler models.**
     + **Apply regularization (L1 or L2).**
     + **Use ensemble methods like bagging or boosting.**
2. **Explain bagging and boosting.**
   * **Bagging: Combines predictions from multiple models trained on random subsets of data. Example: Random Forest.**
   * **Boosting: Sequentially trains models to correct the errors of previous ones. Example: XGBoost.**

**Statistical and Conceptual Questions**

1. **What is the difference between Gini Impurity and Entropy in Decision Trees?**
   * **Both measure data impurity:**
     + **Gini Impurity: Measures the likelihood of incorrect classification.**
     + **Entropy: Measures the randomness in data. More computationally intensive than Gini.**
2. **What is the difference between precision and recall?**
   * **Precision: True positives divided by total predicted positives.**
   * **Recall: True positives divided by actual positives.**
3. **Explain the ROC Curve.**
   * **Plots True Positive Rate (TPR) vs. False Positive Rate (FPR).**
   * **Measures model performance. A higher area under the curve (AUC) indicates better performance.**

**Practical Python Questions**

1. **Name Python libraries used for machine learning.**
   * **NumPy: Numerical operations.**
   * **Pandas: Data manipulation.**
   * **Scikit-learn: Machine learning models.**
   * **Matplotlib/Seaborn: Data visualization.**
2. **How would you handle duplicate values in a dataset?**
   * **Use Pandas' .duplicated() to find duplicates and .drop\_duplicates() to remove them.**
3. **How would you write a basic machine learning program to check model accuracy?**
   * **Steps:**
     1. **Import necessary libraries (e.g., Scikit-learn).**
     2. **Split data into training and testing sets.**
     3. **Train the model on the training set.**
     4. **Evaluate accuracy using metrics like precision, recall, or F1-score.**

**If you'd like, I can provide more detailed answers or further refine specific sections. Let me know how you'd like to proceed!**

**You said:**

**provide all the questions with answer**

**3/3**

**ChatGPT said:**

**ChatGPT**

**Here is a comprehensive list of questions from the transcript along with their answers:**

**Core Machine Learning Questions**

**1. How would you explain machine learning to a school-going kid?**

* **Machine learning involves teaching a machine to learn patterns and make decisions. For example, if you meet strangers at a party, you might group them by their clothes or behavior (unsupervised learning). If you already know some people, you classify them based on prior knowledge (supervised learning).**

**2. What are the types of machine learning?**

* **Supervised Learning: Learning with labeled data (e.g., apples and bananas labeled in a dataset).**
* **Unsupervised Learning: Finding patterns in unlabeled data (e.g., clustering similar objects).**
* **Reinforcement Learning: Learning through rewards and penalties (e.g., playing a game and improving after each attempt).**

**3. Explain supervised learning with an example.**

* **Example: Training a model with labeled images of cats and dogs. The model learns features of each class and predicts if a new image is a cat or a dog.**

**4. What is unsupervised learning?**

* **Learning without labeled data. For example, clustering images of animals into groups without labeling them.**

**5. What is reinforcement learning?**

* **Learning by interacting with an environment. Example: Playing a Mario game where the agent (Mario) gets rewarded for progressing and penalized for failing.**

**Scenario-Based Questions**

**6. How would you handle a dataset with 30% missing values?**

* **Options:**
  + **Create a new binary variable to indicate missing values.**
  + **Remove rows/columns with missing data if it doesn't impact analysis.**
  + **Impute missing values using clustering, statistical distributions, or domain knowledge.**

**7. How would you handle outliers in a dataset?**

* **Detection Methods:**
  + **Box plots or interquartile range (IQR).**
  + **Statistical methods based on distributions.**
  + **Proximity-based clustering.**
* **Handling Options:**
  + **Remove outliers if data is abundant.**
  + **Cap extreme values using percentiles (e.g., 99th percentile).**
  + **Impute outliers with averages or medians.**

**8. You are given a dataset with variables having high correlation. Would you remove correlated variables before applying PCA?**

* **PCA can handle correlated variables by transforming them into uncorrelated components, so you don’t need to remove them beforehand.**

**Practical Machine Learning Questions**

**9. What is overfitting, and how can it be avoided?**

* **Overfitting: The model learns the noise in training data and performs poorly on unseen data.**
* **Prevention Methods:**
  + **Collect more data.**
  + **Use simpler models.**
  + **Apply regularization (L1 or L2).**
  + **Use ensemble methods like bagging or boosting.**

**10. Explain bagging and boosting.**

* **Bagging: Combines predictions from multiple models trained on random data subsets (e.g., Random Forest).**
* **Boosting: Trains models sequentially to correct errors from previous models (e.g., XGBoost).**

**11. What is bias-variance trade-off?**

* **Bias: Error from overly simplistic models (underfitting).**
* **Variance: Error from overly complex models (overfitting).**
* **The goal is to balance bias and variance for optimal performance.**

**Statistical and Conceptual Questions**

**12. What is the difference between Gini Impurity and Entropy in Decision Trees?**

* **Gini Impurity: Measures how often a randomly chosen element would be incorrectly classified.**
* **Entropy: Measures the randomness or uncertainty in data. It's computationally more expensive than Gini.**

**13. What is the difference between precision and recall?**

* **Precision: The ratio of correctly predicted positive instances to total predicted positives.**
* **Recall: The ratio of correctly predicted positive instances to total actual positives.**

**14. What is an ROC curve, and how do you interpret it?**

* **ROC Curve plots True Positive Rate (TPR) vs. False Positive Rate (FPR).**
* **A higher area under the curve (AUC) indicates better model performance.**

**Python and Practical Coding Questions**

**15. Name Python libraries used for machine learning.**

* **NumPy: Numerical operations.**
* **Pandas: Data manipulation.**
* **Scikit-learn: Machine learning algorithms.**
* **Matplotlib/Seaborn: Data visualization.**
* **TensorFlow/PyTorch: Deep learning frameworks.**

**16. How would you handle duplicate values in a dataset?**

* **Use Pandas functions:**
  + **.duplicated() to identify duplicates.**
  + **.drop\_duplicates() to remove duplicates.**

**17. How would you write a basic machine learning program?**

* **Steps:**
  1. **Import required libraries (e.g., Scikit-learn).**
  2. **Load and preprocess the dataset.**
  3. **Split data into training and testing sets.**
  4. **Train a model (e.g., Logistic Regression).**
  5. **Evaluate performance using metrics like accuracy, precision, or recall.**

**18. What is the difference between a Pandas DataFrame and a Series?**

* **Series: One-dimensional labeled array (like a single column).**
* **DataFrame: Two-dimensional labeled structure (like a table with multiple columns).**

**Advanced Scenario-Based Questions**

**19. You’re building a classification model for cancer detection with 96% accuracy. Why might this not be a good model?**

* **If the dataset is imbalanced (e.g., very few cancer cases), accuracy alone isn’t sufficient. Use precision, recall, or F1-score for better evaluation.**

**20. How would you evaluate a recommendation system?**

* **Use metrics like precision, recall, and Mean Average Precision (MAP).**
* **For collaborative filtering:**
  + **Precision measures how relevant recommendations are.**
  + **Recall evaluates the completeness of recommendations.**

**Miscellaneous Questions**

**21. What is A/B Testing?**

* **A statistical hypothesis test to compare two models or strategies.**
* **Example: Comparing the click-through rates of two recommendation algorithms.**

**22. What is the difference between Type I and Type II errors?**

* **Type I Error: Rejecting a true null hypothesis (false positive).**
* **Type II Error: Failing to reject a false null hypothesis (false negative).**

**23. What is collaborative filtering in recommendation systems?**

* **Uses user-item interactions to recommend items.**
* **Example: Amazon suggests products based on similar users' purchases.**

***Q1 Explain different types of machine learning with their pros and cons?***

They are classified into 4 categories : Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning.

***1.Supervised Learning****:* Deals with “Labelled data” as input and based on that generates an corresponding output.

Pros : Have an exact idea about classes in training data, Simple process and easy to interpret.

Cons : In the case of classification, if we give an input that is not from any of the classes in the training data, then the output may be a wrong class label, training needs a lot of computation time, so do the classification, especially if the data set is very large.

***2.Unsupervised Learning :*** Deal with “Un-Labelled data” so the main aim here is to **group or categories the unsorted dataset according to the similarities, patterns, and differences.** Machines are instructed to find the hidden patterns from the input dataset.

Pros : Unlabeled data can be easily made available, lesser complexity compared to the supervised learning task, Identify hidden patterns which humans is difficult to visualize.

Cons : Most of times having low accuracy, It is costlier because it requires human intervention to correlated data with domain knowledge.

***3.Semi-Supervised Learning :*** Training data comprises of both labelled and unlabeled data. Steps of working are below :

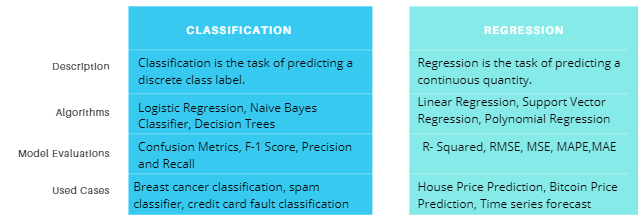
* Firstly, it trains the model with less amount of training data similar to the supervised learning models. The training continues until the model gives accurate results.
* The algorithms use the unlabeled dataset with pseudo labels in the next step, and now the result may not be accurate.
* Now, the labels from labeled training data and pseudo labels data are linked together.
* The input data in labeled training data and unlabeled training data are also linked.
* In the end, again train the model with the new combined input as did in the first step. It will reduce errors and improve the accuracy of the model.

**Assumptions:**

* ***Continuity Assumption:***The objects near each other tend to share the same group or label. This assumption is also used in supervised learning, and the datasets are separated by the decision boundaries. But in semi-supervised, the decision boundaries are added with the smoothness assumption in low-density boundaries.
* ***Cluster assumptions:*** Data is divided into different discrete clusters. Further, the points in the same cluster share the output label.
* ***Manifold assumptions:*** This assumption helps to use distances and densities, and this data lie on a manifold of fewer dimensions than input space.
* The dimensional data are created by a process that has less degree of freedom and may be hard to model directly. (***This assumption becomes practical if high****)*.

***4. Reinforcement Learning:***Here agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty. There is no labelled data so, agent is bound to learn from experience only.

***Q2 What is difference between Regression and Classification ?***



Key Difference Between Regression & Classification

***Q3 What is difference between Structure & Unstructured Data ?***

**Structured Data :**Structured data stand for the information that is highly factual, organized and to the point. eg — SQL Data base, Spreadsheets, OLTP system, Online Forms.

**Unstructured Data :**Unstructured data does not have any pre-defined structure to it and comes in all its diversity of forms eg — Images, Videos, Audio, Reports.

***Q4. What are assumption of Linear Regression ?***

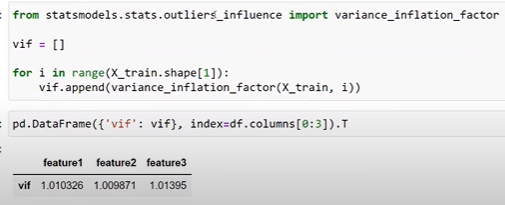
* Linear Relationship between input and output
* No Multicollinearity
* Normality of residual
* Homoscedasticity
* No auto-correlation of errors

***Q5. What is multicollinearity and why it is a problem in linear regression ?***

Multicollinearity exist when an independent variable is highly correlated with another independent variable in multiple regression equation.

This can be problematic because it undermines the statistical significance of an independent variable.

You can use the ***Variation Inflation Factors(VIF)***to determine if there is any multicollinearity between independent variables — if VIF is greater than 5 then multicollinearity exists.



Variation Inflation Factor (VIF)

Eg : y = m1x1 + m2x2 + m3x3 +c

Ideally x1,x2,x3 should be independent but if we increase values of x1 then value of x2 also increase/decreases which concludes multicollinearity exists. As per linear regression, m1 represents if we make changes in m1 then what changes can be observed in y and other will be constant. However, this will not work when multicollinearity exists. So, it violates linear regression.

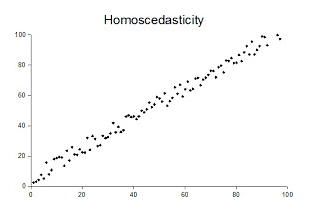
***Q6. In linear regression, what is the value of the sum of the residuals for a given dataset?***

1. ***Normality assumption:*** It is assumed that the error terms, ε(i), are normally distributed.
2. ***Zero mean assumption:*** It is assumed that the residuals have a mean value of zero.
3. ***Constant variance assumption:*** It is assumed that the residual terms have the same (but unknown) variance, σ2 This assumption is also known as the assumption of homogeneity or homoscedasticity.
4. ***Independent error assumption:*** It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

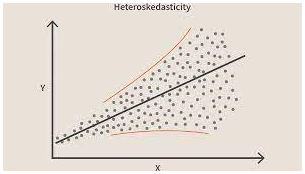
We can plot KDE plot and check whether it is normal bell shapes curve or not. Also, we can plot Q-Q plot and check what line is crossing max points or not.

***Q7. What is homoscedasticity & Heteroscedasticity ?***

**Homoscedasticity** means to be of “The same Variance” or “same scatter”. In other words, Linear Regression assumes that for all the instances, the error terms will be the same and of very little variance.



Homoscedasticity



Heteroscedasticity

**Heteroscedasticity**refers to data for which the [variance](https://deepai.org/machine-learning-glossary-and-terms/variance) of the dependent variable is unequal across the range of independent variables. A regression model assumes a consistent variance, or homoscedasticity, across the data.

Heteroscedasticity in the data results in regression providing accurate outputs on one end of the data range but highly inaccurate outputs on the other end of the data. An easy way to visualize these concepts is to create a *scatter plot* of the data.

A heteroscedastic data set will exhibit a conical shape across the range of independent variables. The wider the cone, the more heteroscedastic the data is and the less friendly for regression analysis.

It is important to understand that a regression analysis on the data set is still possible but the results will prove unreliable outside of a specific range.

***Q8. What are the reasons and effect of Heteroscedasticity on model ?***

*Reasons for Heteroscedasticity :*

* When you are fitting the wrong model. If you fit a linear regression model to a data which is non-linear, it will lead to Heteroscedasticity.
* When the is large variance in machine learning model.
* When the scale of values in a variable is not the same.
* When a wrong transformation on data is used for regression.
* When there is left/right skewness present in the data.

*Effect of Heteroscedasticity :*

* Presence of Heteroscedasticity makes the coefficients less precise and hence the correct coefficients are further away from the population value.
* Heteroscedasticity is also likely to produce p-values smaller than the actual values. This is due to the fact that the variance of coefficient estimates has increased but the standard OLS (*Ordinary Least Squares*) model did not detect it. Therefore the OLS model calculates *p-values* using an underestimated variance. This can lead us to incorrectly make a conclusion that the regression coefficients are significant when they are actually not significant.
* The *standard errors* produced will also be biased. Standard errors are crucial in calculating significant tests and confidence intervals. If the Standard errors are biased, it will mean that the tests are incorrect and the regression coefficient estimates will be incorrect.

***Q9. How to handle problem of Heteroscedasticity on model ?***

**Manipulating the variables**

We can make some modifications to the variables/features we have to reduce the impact of this large variance on the model predictions.

**Weighted Regression**

Weighted regression is a modification of normal regression where the data points are assigned certain weights according to their variance. The ones with large variance are given small weights and the ones with less variance are given larger weights.

So when these weights are squared, the square of small weights underestimates the effect of high variance.

When correct weights are used, Heteroscedasticity is replaced by Homoscedasticity. But how to find correct weights? One quick way is to use the inverse of that variable as the weight.

**Transformation**

Use Box-Cox transformations and log transformations.