ML Interview Questions — 1.1

Dishita Neve · Follow

7 min read · Jul 6, 2022

296 3

There are multiple resources of ML interviews questions. Often times we missed sum of key points of certain topics. That's why I am planning to write series of interview questions and nailed it in you next interviews. I'll try my best to explain concepts in easiest possible manner.

Today I am starting with very basic questions..

> *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:

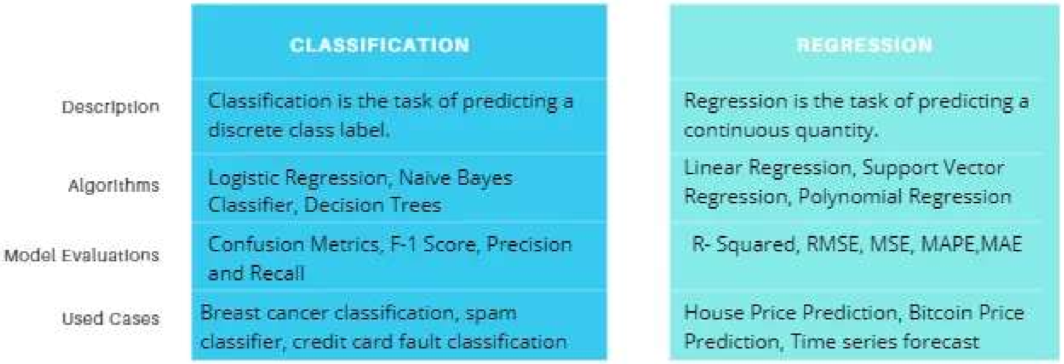
- ***Continuity Assumption:*** The objects near each other tend to share the same group or label. This assumption is also used in supervised learning, and the datasets are separated by the decision boundaries. But in semi-supervised, the decision boundaries are added with the smoothness assumption in 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 ?*



	CLASSIFICATION	REGRESSION
Description	Classification is the task of predicting a discrete class label.	Regression is the task of predicting a continuous quantity.
Algorithms	Logistic Regression, Naive Bayes Classifier, Decision Trees	Linear Regression, Support Vector Regression, Polynomial Regression
Model Evaluations	Confusion Metrics, F-1 Score, Precision and Recall	R- Squared, RMSE, MSE, MAPE,MAE
Used Cases	Breast cancer classification, spam classifier, credit card fault classification	House Price Prediction, Bitcoin Price Prediction, Time series forecast

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 ?*

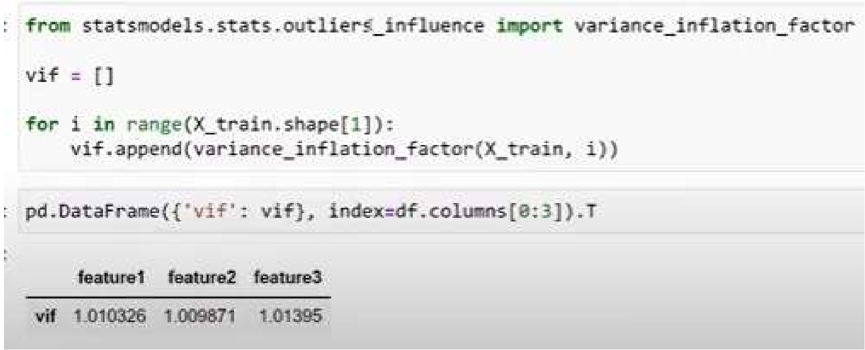
- Linear Relationship between input and output
- No Multicollinearity
- Normality of residual
- Homoscedasticity
- No auto-correlation of errors

> *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.



```
from statsmodels.stats.outliers_influence import variance_inflation_factor

vif = []

for i in range(X_train.shape[1]):
    vif.append(variance_inflation_factor(X_train, i))
```

```
pd.DataFrame({'vif': vif}, index=df.columns[0:3]).T
```

```
        feature1  feature2  feature3
vif     1.010326  1.009871   1.01395
```

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.

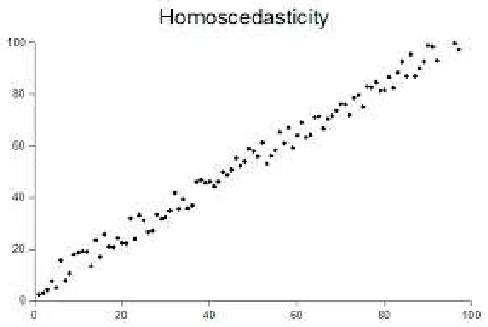
3. ***Constant variance assumption:*** It is assumed that the residual terms have the same (but unknown) variance, σ2 This assumption is also known as the assumption of homogeneity or homoscedasticity.

4. ***Independent error assumption:*** It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

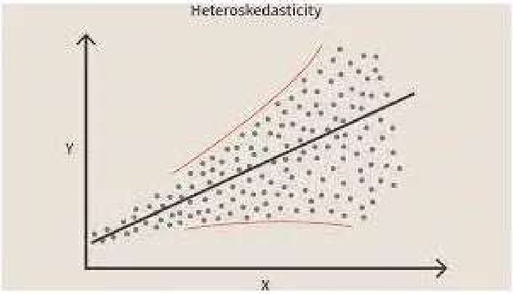
We can plot KDE plot and check whether it is normal bell shapes curve or not. Also, we can plot Q-Q plot and check what line is crossing max points or not.

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 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.

Here, you no longer know you no longer can easily explain what the feature is showing.

> *In next part I will cover topics related to hypothesis testing, statistical test and outlier detection technique. If you have any interview questions and struggling to find the answer you can mention in chats or reach out to me on linkedin*

Machine Learning · Interview Questions · Data Science · Regression · Interview Preparation

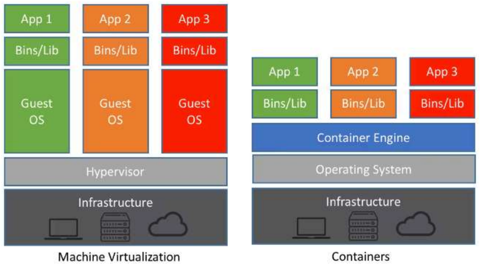
Written by Dishita Neve

107 Followers · 10 Following

Follow

AI Consultant | Generative AI | LLM | Data Engineering

More from Dishita Neve

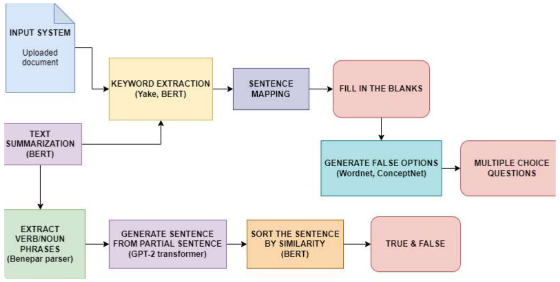


Dishita Neve

Docker Basics Part 1: A Practical Understanding of Its Theory and…

In today's era, urge to learn Docker has become essential for efficient…

Apr 23, 2023 20


INPUT SYSTEM
Uploaded document
KEYWORD EXTRACTION (Yake, BERT)
SENTENCE MAPPING
FILL IN THE BLANKS
TEXT SUMMARIZATION (BERT)
GENERATE FALSE OPTIONS (Wordnet, ConceptNet)
MULTIPLE CHOICE QUESTIONS
EXTRACT VERB/NOUN PHRASES (Benepar parser)
GENERATE SENTENCE FROM PARTIAL SENTENCE (GPT-2 transformer)
SORT THE SENTENCE BY SIMILARITY (BERT)
TRUE & FALSE

Dishita Neve

Automated PDF Question Generator: Enhancing Learning…

This blog focuses on projects, and within this article, I have detailed the steps involved in…

Jul 24, 2023 50

Dishita Neve

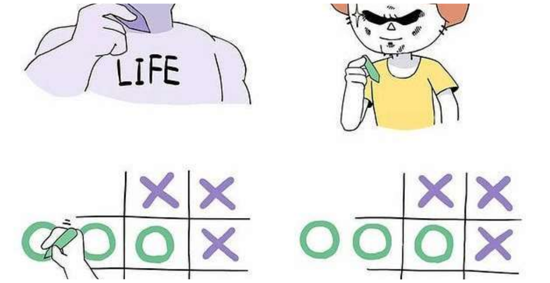
LinkedIn Insight Plus: Revolutionizing Hiring and...

Introduction

Jul 24, 2023 👏 2

See all from Dishita Neve

Recommended from Medium



Amit Yadav

Top 11 Linear Regression Interview Questions (With Answers)

If you think you need to spend $2,000 on a 120-day program to become a data scientist…

Jul 21 5

Vikash Singh

Top Interview Questions and Answers on Bagging Algorithms…

If you're preparing for a data science interview, understanding ensemble method…

Sep 3 17

Lists

Predictive Modeling w/ Python

20 stories · 1695 saves

Practical Guides to Machine Learning

10 stories · 2062 saves

Natural Language Processing

1839 stories · 1463 saves

data science and AI

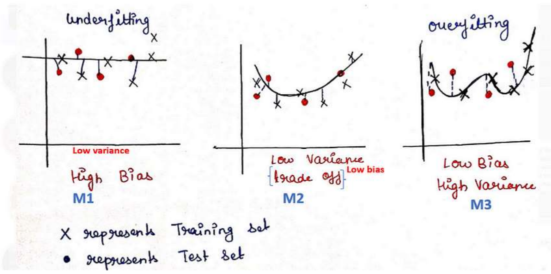
40 stories · 295 saves

tds In Towards Data Science by Samuel Flender

How I Cracked the Meta Machine Learning Engineering Interview

Practical tips for the coding, design, and behavior rounds

★ Oct 25, 2022 👏 2K 💬 12



Akanksha Verma, MSc Data Science

Bias-Variance Tradeoff in Machine Learning

Machine Learning models are powerful tools for making predictions, but achieving the...

Oct 18 👏 71 💬 1

Mikel

Data Science Internship at Amazon with these 25 interview questions.

Interview Prep, Process, and Perks.

★ Sep 3 👏 6

Tilak Mudgal

Machine Learning Concepts Part 1

Scenario-based questions for an ML engineer interview:

Nov 10 👏 1

See more recommendations