Carrier.com interview questions

1.What is normalization\*\*\*vs standardization

Normalization

1. Minimum and max value of features are used for scaling
2. It is used when features are n different scales.
3. Scales values between [0,1] or[-1,1]
4. It is affected by outliers.
5. Scikit learn provides transformer called minmax scaler for normalization
6. It is often called as Scaling Normalization and useful when we don’t know about the distribution

Standardization

1.Mean and standard deviation is used for scaling.

2.It is used when we want to ensure zero mean and unit standard deviation.

3.It is not bounded to a certain range

4.It is less affected by the outliers

5.Scikitlearn provides a transformer called Standard scaler

6.It is often called as Z-score normalization and useful when feature disturbution is normal or gaussio

2.what is dimensionality reduction techniques explain its working:\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

* Dimensionality reduction is the process of reducing the number of features in a dataset while retaining as much information as possible.  
  This can be done to reduce the complexity of a model, improve the performance of a learning algorithm, or make it easier to visualize the data.

**Methods of Dimensionality Reduction:**

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)
* Generalized Discriminant Analysis (GDA)

As the number of features or dimensions in a dataset increases, the amount of data required to obtain a statistically significant result increases exponentially. This can lead to issues such as overfitting, increased computation time, and reduced accuracy of machine learning models this is known as the curse of dimensionality problems that arise while working with high-dimensional data

It involves:

* Construct the covariance matrix of the data.
* Compute the eigenvectors of this matrix.
* Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large fraction of variance of the original data.

**Step-By-Step Explanation of PCA (Principal Component Analysis)**

**Step 1: Standardization**

**Step2: Covariance Matrix Computation**

### Step3: Step 3: Compute Eigenvalues and Eigenvectors of Covariance Matrix to Identify Principal Components

### https://github.com/priyankabanda2202/interviewprep/blob/main/Dimensionality\_reduction\_techinique.ipynb

**3.what is dot product:**

The Dot product is a way to multiply two equal-length vectors together

1.It is also called as **Scalar product, Inner product, Projection product**

np.dot(A, B)

a.dot(b)

* [cosine similarity](https://www.learndatasci.com/glossary/cosine-similarity/) is one of the most important similarity metrics and relies on the dot product.
* Neural networks use dot products to compute weighted sums efficiently.
* Calculations of [orthogonality](https://www.learndatasci.com/glossary/orthogonal-and-orthonormal-vectors/)

**4.how is dimntionality reduction in imags works:**

In computer vision, images and videos are often represented as high-dimensional pixel arrays. Dimensionality reduction techniques allow us to extract essential features and patterns from these images, aiding tasks such as object recognition, facial expression analysis, and image clustering. PCA and Autoencoders are commonly used to reduce the dimensionality of image

data, making it easier to train models and recognize objects efficiently.

**Auto Encoders:**

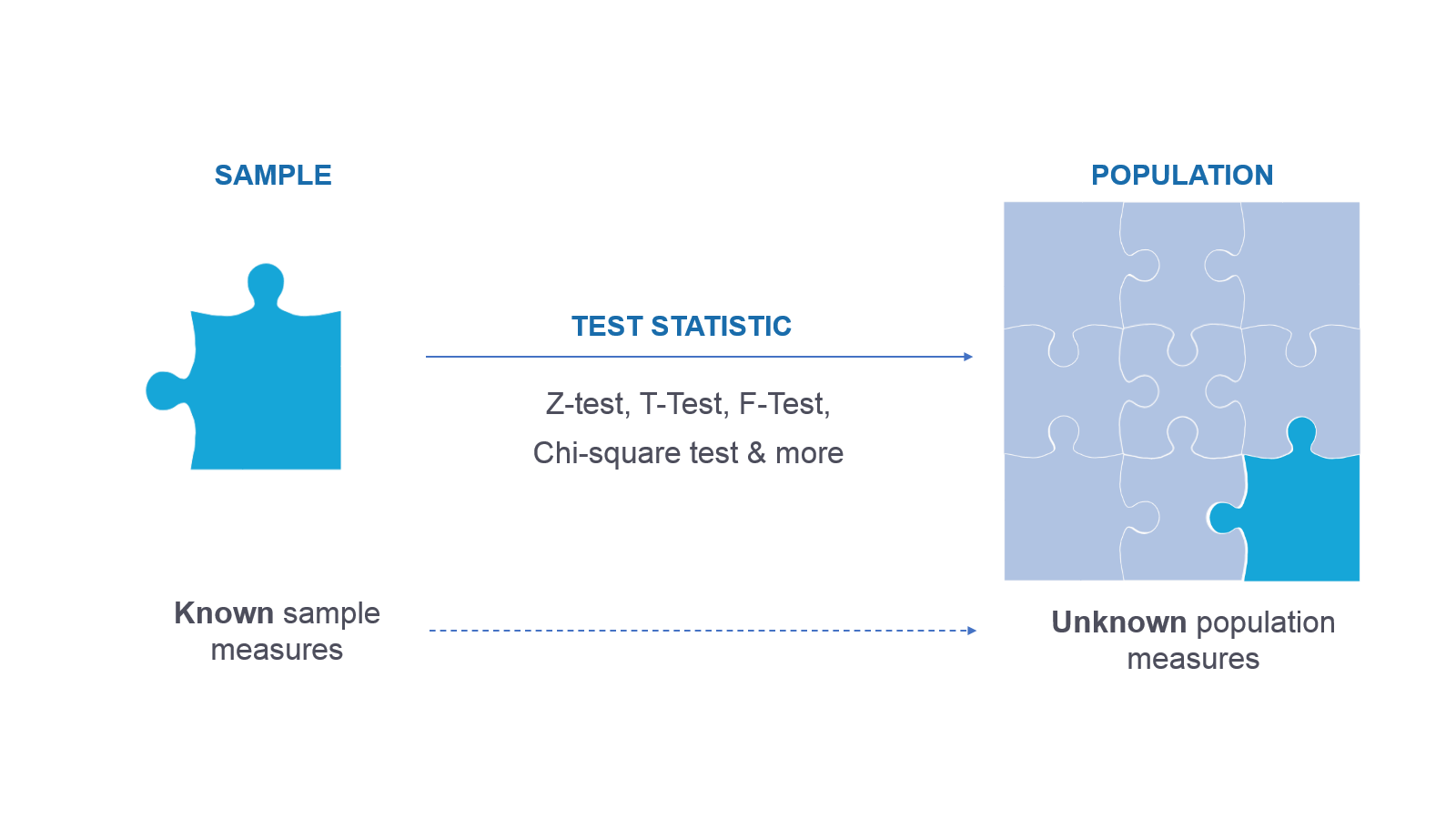
**The Architecture:**

1. **Encoder:** The encoder maps the input data to a lower-dimensional latent space representation.
2. **Bottleneck Layer:** The bottleneck layer is a crucial part of the encoder that creates the compressed representation.
3. **Decoder:** The decoder reconstructs the data from the compressed representation.

Autoencoders are versatile tools for dimensionality reduction and feature learning. They can capture complex relationships in data and are often used for denoising data, generating novel samples, and reducing dimensionality.

## What is Hypothesis Testing?\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Any data science project starts with exploring the data. When we perform an analysis on a sample through exploratory data analysis and inferential statistics we get information about the sample. Now, we want to use this information to predict values for the entire population.



**Hypothesis testing** is done to confirm our observation about the population using sample data, within the desired error level. Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude if the hypothesis about the population is true or not.

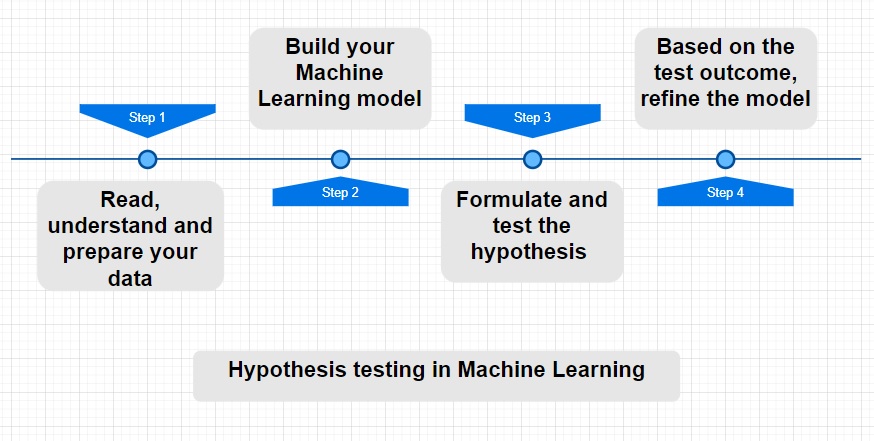
Central Limit Theorem Statement The central limit theorem relies on the concept of sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population. The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial or any other distribution, the sampling distribution of mean will be normal.

Quartiles Quartiles are a set of three values that divide a dataset into four equal parts, each containing 25% of the data points. These three values are: ● Q1 (First Quartile): The 25th percentile. This value is the middle number between the smallest number (minimum) and the median of the dataset. 25% of the data points are less than Q1. ● Q2 (Second Quartile): The 50th percentile (median). This value is the middle number of the dataset, with 50% of the data points below it and 50% above it. 21 ● Q3 (Third Quartile): The 75th percentile. This value is the middle number between the median and the largest number (maximum) of the dataset. 75% of the data points are less than Q3. Interpretation of Quartiles: Quartiles, along with the median, provide valuable insights into the distribution of a dataset. They can be used to: ● Identify the spread of the data: The difference between Q1 and Q3 is known as the interquartile range (IQR). A large IQR indicates a wide spread of data, while a small IQR indicates a narrow spread. ● Identify outliers: Data points that are significantly lower than Q1 or higher than Q3 can be considered outliers. ● Compare different datasets: Quartiles allow you to compare the distributions of different datasets, even if they have different units or scales.

Box Plot A box plot, also known as a whisker plot, is a graphical representation of a five-number summary of a dataset. The five-number summary consists of the following statistics: 1. Minimum: The smallest number in the dataset. 2. First Quartile (Q1): The middle number between the minimum and the median (25th percentile). 3. Median (Q2): The middle number of the dataset (50th percentile). 4. Third Quartile (Q3): The middle value between the median and the maximum (75th percentile). 5. Maximum: The largest number in the dataset. Here’s how these five numbers are represented in a box plot: 25 ● The box in the box plot represents the interquartile range (IQR), which is the range between Q1 and Q3. The length of the box is therefore IQR = Q3 - Q1. ● The line inside the box represents the median of the dataset. ● The whiskers (lines extending from the box) represent the range of the data within 1.5 times the IQR from the box. Any data point outside this range is considered an outlier and is represented as a dot (or a different marker). ● The ends of the whiskers represent the minimum and maximum data values excluding outliers. This way, a box plot provides a visual summary of the distribution of a dataset, including its central tendency (median), variability (IQR), and outliers. It’s a useful tool for comparing distributions across different categories

## How to perform hypothesis testing in machine learning?

To trust your model and make predictions, we utilize hypothesis testing. When we will use sample data to train our model, we make assumptions about our population. By performing hypothesis testing, we validate these assumptions for a desired significance level.



Key steps to perform hypothesis test are as follows:

1. Formulate a Hypothesis
2. Determine the significance level
3. Determine the type of test
4. Calculate the Test Statistic values and the p values
5. Make Decision
6. **Formulating the hypothesis**
7. One of the key steps to do this is to formulate the below two hypotheses:
8. **The null hypothesis**represented as H₀ is the initial claim that is based on the prevailing belief about the population.  
   **The alternate hypothesis**represented as H₁ is the challenge to the null hypothesis. It is the claim which we would like to prove as True
9. One of the main points which we should consider while formulating the null and alternative hypothesis is that the null hypothesis always looks at confirming the existing notion. Hence, it has sign >= or , < and ≠

## Select the type of Hypothesis test

We choose the type of test statistic based on the predictor variable – quantitative or categorical. Below are a few of the commonly used test statistics for quantitative data

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of predictor variable** | **Distribution type** | **Desired Test** | **Attributes** |
| Quantitative | Normal Distribution | Z – Test | * Large sample size * Population standard deviation known |
| Quantitative | T Distribution | T-Test | * Sample size less than 30 * Population standard deviation unknown |
| Quantitative | Positively skewed distribution | F – Test | * When you want to compare 3 or more variables |
| Quantitative | Negatively skewed distribution | NA | * Requires feature transformation to perform a hypothesis test |
| Categorical | NA | Chi-Square test | * Test of independence * Goodness of fit |
|  |  |  |  |

#### Chi-Square Test

For categorical variables, we would be performing a chi-Square test.

Following are the two types of chi-squared tests:

1. Chi-squared test of independence – We use the Chi-Square test to determine whether or not there is a significant relationship between two categorical variables.
2. Chi-squared Goodness of fit helps us determine if the sample data correctly represents the population.

## The decision about your model

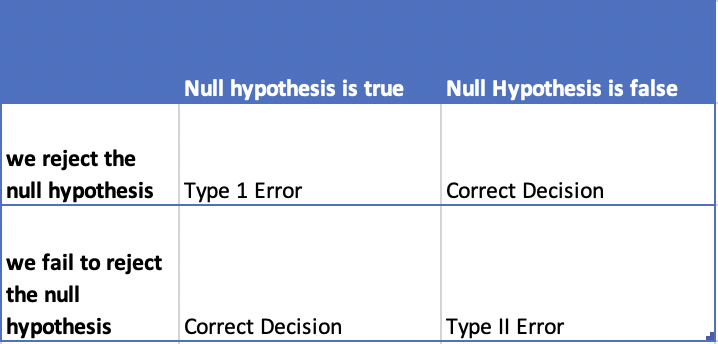
Test Statistic is then used to calculate P-Value. A P-value measures the strength of evidence in support of a null hypothesis. If the P-value is less than the significance level, we reject the null hypothesis.

if the **p-value < α**, then we have statistically significant evidence against the null hypothesis, so we reject the null hypothesis and accept the alternate hypothesis

if the p-value > α then we do not have statistically significant evidence against the null hypothesis, so we fail to reject the null hypothesis.

As we make decisions, it is important to understand the errors that can happen while testing.

There are two possible types of error we could commit while performing hypothesis testing.



1) **Type1 Error**– This occurs when the null hypothesis is true but we reject it.The probability of type I error is denoted by alpha (α). Type 1 error is also known as the level of significance of the hypothesis test

2) **Type 2 Error** – This occurs when the null hypothesis is false but we fail to reject it. The probability of type II error is denoted by beta (β)

* P-value is given in the column P>|t| – As mentioned above, for a good model, we want this value to be less than the significance level.

**What is cosine similarity?**

* Cosine similarity is the cosine of the angle between the vectors; that is, it is the dot product of the vectors divided by the product of their lengths. It follows that the cosine similarity does not depend on the magnitudes of the vectors, but only on their angle.

Cosine similarity is a metric used to measure how similar the documents are irrespective of their size.