**Sheza Nauman**

**Assignment 4**

**Artificial Intelligence using Machine Learning and Deep Learning**

1. **Hypothesis:-**

A hypothesis in machine learning (ML) and deep learning (DL) usually refers to a model's approximate representation of the real underlying link between input data and output predictions. In supervised learning, when the objective is to learn a function that maps input data to output labels or values, the hypothesis is the fundamental building block.

**Types of Hypotheses in Machine Learning**

#### Null Hypothesis

#### Alternative Hypothesis

#### One-tailed Hypothesis

#### Two-tailed Hypothesis

#### Null Hypothesis:

#### A null hypothesis is a basic hypothesis that states that no link exists between the independent and dependent variables. In other words, it assumes the independent variable has no influence on the dependent variable. It is symbolized by the symbol H0. If the p-value falls outside the significance level, the null hypothesis is typically rejected (). If the null hypothesis is correct, the coefficient of determination is the probability of rejecting it. A null hypothesis is involved in test findings such as t-tests and ANOVA.

#### Alternative Hypothesis:

An alternative hypothesis is a hypothesis that contradicts the null hypothesis. It assumes that there is a relationship between the independent and dependent variables. In other words, it assumes that there is an effect of the independent variable on the dependent variable. It is denoted by Ha. An alternative hypothesis is generally accepted if the p-value is less than the significance level (α). An alternative hypothesis is also known as a research hypothesis.

#### One-tailed Hypothesis:

A one-tailed test is a type of significance test in which the region of rejection is located at one end of the sample distribution. It denotes that the estimated test parameter is more or less than the crucial value, implying that the alternative hypothesis rather than the null hypothesis should be accepted. It is most commonly used in the chi-square distribution, where all of the crucial areas, related to, are put in either of the two tails. Left-tailed or right-tailed one-tailed tests are both possible.

#### Two-tailed Hypothesis:

The two-tailed test is a hypothesis test in which the region of rejection or critical area is on both ends of the normal distribution. It determines whether the sample tested falls within or outside a certain range of values, and an alternative hypothesis is accepted if the calculated value falls in either of the two tails of the probability distribution. α is bifurcated into two equal parts, and the estimated parameter is either above or below the assumed parameter, so extreme values work as evidence against the null hypothesis. Overall, the hypothesis plays a critical role in the machine learning model. It provides a starting point for the model to make predictions and helps to guide the learning process. The accuracy of the hypothesis is evaluated using various metrics like mean squared error or accuracy.

1. **Difference between z and t test in regression:-**

In regression analysis, both the Z-test and the t-test are used to assess the statistical significance of coefficients (slope parameters) in a regression model. However, there are important differences between these two tests, mainly related to the assumptions and situations in which they are used

**Assumption about Population Variance:**

Z-test: The Z-test assumes that you know the population standard deviation, σ (sigma). This test is typically used when you have a large sample size or when you have information about the population standard deviation.

t-test: The t-test, on the other hand, does not assume that you know the population standard deviation. Instead, it estimates the standard error of the coefficient based on the sample data. This test is commonly used when you have a small sample size or when you do not have information about the population standard deviation.

**Sample Size:**

Z-test: Typically, the Z-test is more appropriate when your sample size is large (e.g., n > 30). When the sample size is large, the sampling distribution of the sample mean is approximately normally distributed, allowing you to use the standard normal distribution for hypothesis testing.

t-test: The t-test is preferred when your sample size is small (e.g., n < 30) or when the population standard deviation is unknown. In such cases, the t-distribution has heavier tails than the normal distribution, which accounts for the added uncertainty due to the smaller sample size.

**Hypothesis Testing:**

Z-test: It is often used for hypothesis tests involving population parameters, such as testing the significance of a regression coefficient (slope) in a simple linear regression model.

t-test: The t-test is more commonly used in the context of estimating population parameters when you have a small sample or when the population standard deviation is unknown.

**Degrees of Freedom**:

Z-test: The Z-test does not involve degrees of freedom because it relies on the known population standard deviation.

t-test: The t-test takes into account the degrees of freedom, which depend on the sample size. The degrees of freedom affect the shape of the t-distribution and the critical values used for hypothesis testing.

In summary, the choice between a Z-test and a t-test in regression analysis depends on factors such as the sample size, the availability of population standard deviation information, and the specific hypothesis being tested. A Z-test is appropriate for larger samples with known population standard deviation, while a t-test is more suitable for smaller samples or when the population standard deviation is unknown.

1. **Normal distribution and its types:-**

The normal distribution, commonly known as the Gaussian distribution, is essential to machine learning (ML) and statistics because it can accurately represent a variety of real-world occurrences. The bell-shaped curve, which is symmetrical and centred around the mean (average) value, is what distinguishes the normal distribution. Its form and dispersion are determined by two parameters, the mean and the standard deviation. There are a few important types or variations of the normal distribution that are commonly used in ML and statistics:

1. **(Z-Distribution) Standard Normal Distribution**

1. A specific example of the normal distribution, the standard normal distribution has a mean of 0 and a standard deviation of 1.

2. The random variables that follow this distribution are commonly referred to as Z-scores, and it is often denoted as Z N(0, 1).

3. Z-scores are used to standardize and compare data in a number of statistical calculations and tests.

1. **Multiple Variable Normal Distribution**
2. You work with multivariate data in many ML applications, where each observation comprises of many variables.
3. The normal distribution's application to multiple dimensions is expanded by the multivariate normal distribution.
4. It has a mean vector () and a covariance matrix () that explain the means, variances, and correlations between variables instead of a single mean and standard deviation.
5. It is frequently applied in methods like probabilistic modeling for data with numerous features and Gaussian Mixture Models (GMMs).
6. **The t-Distribution of students**

The Student's t-distribution is employed when the population standard deviation needs to be determined from a sample and is connected to the normal distribution. Similar to the normal distribution, it features a bell-shaped curve, but it has heavier tails, making it more resistant to outliers. The degrees of freedom (df) parameter determines the shape of the t-distribution, and larger df values cause it to resemble a normal distribution.

1. **Skewed Distribution:**

A generalisation of the normal distribution that takes into consideration skewness (asymmetry) in data is the skew-normal distribution.The amount and direction of the skewness are controlled by an extra parameter ().When working with skewly data, skew-normal distributions are employed.

1. **Log-Normal Distribution**:

Despite not being a normal distribution, the log-normal distribution is one.

When the natural logarithm of the data is calculated, it describes data that has a normal distribution.

It is frequently used to simulate data that is naturally right-skewed and positive, like income or stock prices.

The normal distribution and its variations are used in machine learning (ML) in a number of algorithms, such as Gaussian Naive Bayes, Gaussian Mixture Models, and probability density estimation. For data modelling and statistical analysis in machine learning, it is essential to comprehend the features of these distributions.

1. **Cost function in regression:-**

In regression analysis, a cost function, also known as a loss function or objective function, is a mathematical function that measures the error or discrepancy between the predicted values generated by a regression model and the actual observed values in the dataset. The goal of regression analysis is to find the model parameters (coefficients) that minimize this cost function. Different types of regression models use different cost functions. Here are some commonly used cost functions in regression:

**Mean Squared Error (MSE):**

MSE is one of the most widely used cost functions in regression. It calculates the average of the squared differences between the predicted values and the actual values. The formula for MSE is: MSE = (1/n) Σ(yᵢ - ŷᵢ)², where yᵢ represents the actual values, ŷᵢ represents the predicted values, and n is the number of data points.

**Mean Absolute Error (MAE):**

MAE measures the average absolute differences between the predicted values and the actual values. The formula for MAE is: MAE = (1/n) Σ|yᵢ - ŷᵢ|.

**Root Mean Squared Error (RMSE):**

RMSE is the square root of the MSE and is often used to provide error in the same units as the dependent variable. The formula for RMSE is: RMSE = √(MSE).

**R-Squared (R²) or Coefficient of Determination**:

R-squared measures the proportion of the variance in the dependent variable (target) that is explained by the independent variables (features) in the regression model. It is not a traditional error metric like MSE or MAE but is commonly used to assess the goodness of fit. R² values range from 0 to 1, with higher values indicating a better fit. An R² of 1 means that the model explains all the variance in the data.

**Huber Loss:**

Huber loss is a combination of the MSE and MAE loss functions. It is less sensitive to outliers than MSE because it penalizes large errors linearly (like MAE) for data points that are far from the prediction and quadratically (like MSE) for data points that are close to the prediction.

**Quantile Loss (Quantile Regression):**

Quantile loss is used in quantile regression to estimate conditional quantiles of the response variable. It allows modeling different quantiles of the response distribution, making it useful for scenarios where the distribution of errors is not necessarily Gaussian.

**Logarithmic Loss (Log Loss):**

Logarithmic loss is used in logistic regression, where the target variable represents binary or categorical outcomes. It measures the dissimilarity between predicted probabilities and the true binary outcomes.

The choice of cost function depends on the nature of the data, the goals of the regression analysis, and the assumptions made about the distribution of errors. Different cost functions may be more appropriate for specific situations. The regression model aims to find the model parameters that minimize the chosen cost function through techniques like ordinary least squares (OLS), gradient descent, or other optimization methods.

1. **Model evaluation in regression:-**

Model evaluation in regression involves assessing the performance and quality of a regression model's predictions. Several metrics and techniques can be used to evaluate the accuracy and goodness of fit of a regression model. Here are some common methods for model evaluation in regression:

**Mean Squared Error (MSE):**

MSE measures the average squared difference between the predicted values and the actual target values. It penalizes larger errors more heavily than smaller ones. A lower MSE indicates a better fit.

**Formula:** MSE = (1/n) Σ(yᵢ - ŷᵢ)², where yᵢ represents the actual values, ŷᵢ represents the predicted values, and n is the number of data points.

**Root Mean Squared Error (RMSE):**

RMSE is the square root of the MSE and is in the same units as the target variable. It provides a more interpretable measure of the error.

**Formula:** RMSE = √(MSE).

**Mean Absolute Error (MAE):**

MAE measures the average absolute difference between the predicted values and the actual target values.It is less sensitive to outliers compared to MSE.

**Formula:** MAE = (1/n) Σ|yᵢ - ŷᵢ|.

**R-Squared (R²) or Coefficient of Determination:**

R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables in the regression model. It ranges from 0 to 1, with higher values indicating a better fit. R² = 1 indicates a perfect fit, while R² = 0 suggests that the model does not explain any of the variance. Interpretation: R² represents the percentage of variability in the dependent variable that can be accounted for by the independent variables.

**Adjusted R-Squared**:

Adjusted R-squared is a modification of R-squared that penalizes the inclusion of irrelevant predictors. It accounts for the number of predictors in the model and is useful for comparing models with different numbers of variables.

**Residual Analysis:**

Residuals are the differences between the actual and predicted values. Analyzing the residuals can help identify patterns or trends that the model may have missed. Plotting the residuals against predicted values or predictor variables can reveal potential issues with the model, such as hetero-scedasticity or nonlinearity.

**Distribution of Residuals**:

Examining the distribution of residuals can help ensure that the assumptions of regression (e.g., normality of errors) are met. Q-Q plots and histogram plots of residuals are common tools for assessing the distribution.

**Cross-Validation:**

Cross-validation techniques, such as k-fold cross-validation, can help estimate the model's performance on unseen data. It involves splitting the dataset into multiple subsets, training the model on some subsets, and evaluating it on the remaining data. Cross-validation provides a more robust estimate of a model's performance and helps identify potential over-fitting issues.

**Outlier Detection:**

Identifying and handling outliers is important in regression. Outliers can significantly impact the model's performance and should be investigated and potentially addressed.

**Predictive Power:**

Ultimately, the model's predictive power on new, unseen data is a crucial measure of its performance. It is important to assess how well the model generalizes to new observations.

The choice of evaluation metric depends on the specific goals of the regression analysis and the nature of the data. It's often a good practice to use a combination of these evaluation methods to obtain a comprehensive understanding of the model's performance.

1. **Correlation | Causation | Co- variance:-**

Correlation, causation, and covariance are fundamental concepts in statistics and machine learning, each serving a different purpose and providing insights into data relationships.

**Correlation:-**

**Definition:** Correlation measures the statistical relationship between two variables. It quantifies the degree to which two variables move together, indicating whether there is a linear association between them.

**Range:** Correlation coefficients range from -1 to 1.

**Interpretation:**

1. A correlation coefficient of 1 (positive correlation) indicates a perfect positive linear relationship: as one variable increases, the other also increases.

2. A correlation coefficient of -1 (negative correlation) indicates a perfect negative linear relationship: as one variable increases, the other decreases.

3. A correlation coefficient of 0 (no correlation) means there is no linear relationship between the variables.

Note: Correlation does not imply causation. Even if two variables are highly correlated, it does not mean that one variable causes the other. Correlation only measures the strength and direction of their association.

**Causation:-**

**Definition:** Causation, or causality, refers to a cause-and-effect relationship between two variables. It asserts that one variable directly influences or causes changes in another.

**Establishing Causation:** Establishing causation typically requires more than just observing a statistical relationship (correlation) between variables. It often involves conducting controlled experiments, considering temporal precedence (cause precedes effect), and ruling out confounding variables.

**Example:** While there may be a correlation between ice cream sales and the number of drownings in a given period (both increase in the summer), one does not cause the other. The common factor is the hot weather.

**Covariance:**

**Definition:** Covariance measures how two variables change together. It is a measure of joint variability between two random variables.

**Range:** Covariance can take on any real value.

**Interpretation:**

1. A positive covariance indicates that when one variable is above its mean, the other tends to be above its mean as well (and vice versa).
2. A negative covariance suggests that when one variable is above its mean, the other tends to be below its mean (and vice versa).
3. A covariance near zero suggests little to no linear relationship.

Use in ML: Covariance is often used in the context of feature selection and understanding the relationships between features in a dataset. However, it can be challenging to interpret directly because it depends on the scale of the variables.

**Conclusion:-** correlation measures the strength and direction of a linear association between two variables, but it does not imply causation. Causation, on the other hand, asserts a cause-and-effect relationship and requires more rigorous analysis to establish. Covariance quantifies how two variables change together but does not provide the same interpretability as correlation. In machine learning and statistics, understanding these concepts is crucial for making informed decisions about data relationships and model building.

1. **Importance of P-value in Regression:-**

In regression analysis, the p-value (probability value) plays a crucial role in assessing the significance of the relationship between independent variables (predictors) and the dependent variable (response). The p-value is an essential statistical measure, and its importance lies in several key aspects of regression analysis: In regression analysis, the p-value (probability value) plays a crucial role in

**Hypothesis Testing**:

The p-value is central to hypothesis testing in regression. It helps you determine whether there is a statistically significant relationship between the independent variables and the dependent variable. The null hypothesis (H0) typically posits that there is no significant relationship between the independent variables and the dependent variable, i.e., the coefficients of the predictors are zero. The alternative hypothesis (Ha) suggests that there is a significant relationship. The p-value associated with a regression coefficient tests the null hypothesis. A low p-value (usually less than a chosen significance level, often 0.05) indicates evidence against the null hypothesis, suggesting that the variable is likely to have a significant impact on the dependent variable.

**Coefficient Significance**:

The p-values associated with individual coefficients in a regression model indicate whether each predictor variable is statistically significant. If a predictor's p-value is less than the chosen significance level, it suggests that the predictor is likely to have a significant effect on the dependent variable, all else being equal.

**Model Selection:**

P-values help in the selection of the most appropriate model when you have multiple predictors. By comparing the p-values of different predictors, you can decide which variables to include or exclude from the model. Variables with high p-values (greater than the chosen significance level) may be candidates for removal from the model to simplify it.

**Assessment of Model Fit:**

In addition to assessing individual coefficients, you can use overall goodness-of-fit tests, such as the F-statistic, which relies on the p-value. A low p-value associated with the F-statistic suggests that the overall model is a good fit for the data.

**Control for Type I Error:**

By choosing a significance level (alpha) for hypothesis testing (commonly 0.05), you control the risk of making a Type I error (false positive). A low p-value provides evidence against the null hypothesis and supports the claim of a significant relationship.

**Interpretability:**

The p-value offers a quantitative measure of the strength of evidence against the null hypothesis. A very low p-value indicates strong evidence against the null hypothesis, making it easier to interpret the results of a regression analysis.

**Conclusion:-**

While p-values are useful for evaluating the significance of specific predictors and overall model fit, they shouldn't be the only criterion used to make judgments in regression analysis. When evaluating the findings and choosing which variables to include in a model, other factors including effect sizes, domain expertise, and practical significance should be taken into account. Furthermore, the regression model's assumptions are vulnerable to scrutiny, and deviations from these assumptions may affect the dependability of p-values.

1. **Sampling and Data Sampling in Probability and Regression (Sampling in Data Science):-**

Sampling is a fundamental concept in statistics and data science that involves selecting a subset (sample) of observations or data points from a larger population or dataset. It is used for various purposes, including estimating population parameters, reducing data size for analysis, and making data collection more efficient. In the context of probability, statistics, and regression analysis, sampling can be explained as follows:

**Sampling in Probability:-**

**Probability Samples:**

In probability theory, sampling is often used to model random experiments or events. A sample space represents all possible outcomes of an experiment, and an event is a subset of the sample space. Probability measures the likelihood of an event occurring, and random sampling helps generate empirical probability distributions.

**Sampling in Statistics**:-

* **Simple Random Sampling**:

In statistics, simple random sampling is a method where each member of the population has an equal chance of being selected for the sample. It is often used for estimating population parameters, such as means and variances, based on sample statistics.

* **Stratified Sampling**:

Stratified sampling divides the population into distinct subgroups or strata based on certain characteristics (e.g., age, gender). Samples are then randomly selected from each stratum in proportion to the population distribution of that stratum. It is useful when you want to ensure representation from different subgroups.

* **Systematic Sampling**:

Systematic sampling selects every nth member from a population, starting from a random point.It is useful when there is a natural order to the population (e.g., a list of students in a class) and the sampling interval is determined systematically.

* **Cluster Sampling**:

Cluster sampling divides the population into clusters (e.g., geographical areas) and randomly selects a subset of clusters. Then, all members within the selected clusters are included in the sample. It is efficient for large populations when it is costly to sample individual elements.

**Sampling in Data Science and Regression**:-

* **Data Sampling**:

In data science and machine learning, data sampling is used to create representative datasets for training, testing, and validation. Common techniques include random sampling, stratified sampling, and oversampling/undersampling for imbalanced datasets. Proper data sampling is crucial to ensure that the model's performance generalizes well to unseen data.

* **Regression Analysis:**

In regression analysis, you may use sampling techniques to create training and testing datasets from a larger dataset. The training dataset is used to build the regression model, while the testing dataset is used to assess its performance. Techniques like cross-validation involve repeated sampling to evaluate model performance across multiple subsets of data.

**Conclusion:-**

sampling is a foundational concept in probability, statistics, and data science. It is used for various purposes, from estimating probabilities in probability theory to creating representative datasets for regression analysis and machine learning. Properly designed and executed sampling methods are essential for making valid inferences and building accurate predictive models.