Machine learning is a type of artificial intelligence that involves teaching computers to learn and make decisions based on data. It allows machines to automatically improve their performance at a task by learning from examples.

Supervised learning is a type of machine learning in which the algorithm is trained on labeled data, and the goal is to learn a mapping between the input features and the target variable to make predictions on new, unseen data.

linear regression to predict a person's salary based on their years of experience. We have a dataset that contains the years of experience and corresponding salaries for a group of people.

logistic regression predict whether a person will buy a certain product based on their age. We have a dataset that contains the age and corresponding information whether a person bought the product or not.

decision trees person will play golf on a particular day based on the weather conditions. We have a dataset that contains the weather conditions (outlook, temperature, humidity, and wind) and whether the person played golf or not.

support vector machines to classify whether an email is spam or not based on the presence of certain keywords in the email.

we have a dataset of housing prices. we want to predict the price of a new house based on its size in square feet

Notation X = [x1 x2 ... xn]

where xi is a row vector of input features for the i th data point, such as the size of the house.

y = [y1, y2, ..., yn]

where yi is the corresponding target value for the ith data point, such as the price of the house.

Unsupervised learning is a type of machine learning in which an algorithm is trained on a dataset that doesn't have labeled target variables. The goal of unsupervised learning is to find patterns, relationships, or structures in the data without any prior knowledge of what the output should look like.

Clustering(a marketing company may use clustering to group customers with similar buying behaviors to create targeted marketing campaigns)

Dimensionality Reduction principal component analysis (PCA), which identifies the most important features of a dataset

Anomaly Detection banks may use anomaly detection to identify fraudulent transactions based on unusual patterns of spending.

Generative Models used for image generation and data augmentation.

Prediction refers to the task of making a prediction or forecast of a future outcome based on a set of input variables. For example, predicting the price of a stock based on historical prices, or predicting whether a customer is likely to churn based on their purchase history.

inference aims to uncover the causal relationships between the input and output variables. For example, understanding which features are the most important in predicting the price of a house, or which genes are most important in determining a disease outcome.

Inference is often used when the goal is to gain insights into the underlying data and build a better understanding of the system being studied. Prediction, on the other hand, is often used when the goal is to make accurate predictions of future outcomes.

Regression refers to the task of predicting a continuous output variable based on one or more input variables. For example, predicting the price of a house based on its size, location, and other features. In regression problems, the output variable is a numerical value, and the goal is to build a model that can predict the output value as accurately as possible.

Classification, on the other hand, refers to the task of predicting a categorical output variable based on one or more input variables. For example, classifying an email as spam or not spam based on its content, or classifying an image as containing a cat or a dog. In classification problems, the output variable is a discrete value or label, and the goal is to build a model that can accurately predict the correct label for a new input.

difference between regression and classification is the type of output variable being predicted: continuous in regression, and categorical in classification. The choice of which type of problem to use depends on the specific application and goals of the project.

2. Data Exploration……..

gene expression data is often used to identify patterns or relationships between gene expression levels and specific biological processes or disease states.

Clustering analysis identify sets of genes that are co-regulated and involved in similar biological processes.

Differential gene expression analysis: This involves identifying genes that are differentially expressed between different groups

Classification analysis: This involves building predictive models that use gene expression data to classify samples into different groups or predict disease outcomes.

1.Which of the following statement about machine learning is correct? Clustering is one of the most-often used unsupervised learning methods.

2.**The major difference between supervised learning and unsupervised learning is** The correct answers are included in the data for supervised learning.

3. What is a model in supervised learning? A mathematical mapping between labels and features.

4. What terms do we use to refer to the correct answers in a dataset? labels dependent variable target variable response **(All)**

5. You want build a machine learning model to predict whether a customer would choose a product on your website. Which of the following statements is correct? (all)

In a dataset, the **rows** are commonly referred to as observations or instances, while the **columns** are referred to as variables or features. Observations refer to the individual data points in the dataset, while variables refer to the attributes or characteristics that are measured for each observation.

What terms do we use for label and features in supervised learning? Label: Output variable that the model is trying to predict in supervised learning. **Features:** Input variables used to predict the label in supervised learning

To describe the distribution of a single variable: For categorical variables: use frequency count and bar charts to show the number of occurrences of each category. For interval variables: use summary statistics, histograms, and box plots to provide a numerical and visual summary of the distribution.

To describe the relationship between two variables: For two interval variables: use covariance, correlation, and scatterplots to examine the linear relationship. For an interval variable and a categorical variable: use comparison of statistics and box plots across categories to examine the relationship.

3. Linear Regression……..

Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a linear equation to the observed data. The goal is to find the best-fitting line through the data points that can predict the value of the dependent variable (Y) based on the values of the independent variable(s) (X).To estimate a linear regression model, one needs to first specify the linear combination of the independent variables that will be used to predict the dependent variable. Once the model specification is determined, the coefficients (or parameters) of the linear equation are estimated using a method called least squares estimation. This method calculates the values of the coefficients that minimize the sum of the squared differences between the observed values of Y and the predicted values of Y based on the linear equation.

factors are important in predicting Y? The significance of the coefficients indicates the importance of each independent variable in predicting the dependent variable. A coefficient that is statistically significant (has a p-value less than the chosen significance level) indicates that the corresponding independent variable is important in predicting Y.

How does each factor affect Y? The interpretation of the coefficients depends on the scale of the independent variable. For interval variables, the coefficient represents the change in Y for a one-unit increase in the corresponding independent variable, holding all other independent variables constant. For categorical variables, the coefficient represents the difference in Y between the reference category and the category of interest.

How well does the model fit the data? The goodness of fit of the model is typically assessed using the R-squared statistic, which measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. Higher values of R-squared indicate a better fit.

How to make predictions of Y? What is the difference between confidence interval and prediction interval? To make predictions of Y, one plugs the values of the independent variables into the linear equation and calculates the predicted value of Y. The confidence interval provides a range of values within which the true value of Y is likely to fall with a certain level of confidence. The prediction interval provides a range of values within which an individual observation of Y is likely to fall with a certain level of confidence. The prediction interval is wider than the confidence interval because it includes the uncertainty associated with the estimated error term.

Extension of basic linear model:

Basic linear model assumes linear relationship between dependent and independent variables

In case of non-linear relationship, the model can be extended

Categorical (qualitative) variables:

Categorical variables cannot be directly included in linear regression model

Dummy variables are used to represent categorical variables

Coefficient of a dummy variable represents the difference between the reference category and the category of interest

Interaction:

Interaction term is a product of two or more independent variables

Captures the combined effect of interacting variables on the dependent variable.

Coefficient of an interaction term represents the change in slope of regression line

Hierarchy principle:

Main effects of interacting variables should be included in the model. Avoids confounding effect and correctly interprets the coefficients of the interaction term.

4.Logistic regression:…….

Used for binary classification

Suitable for dependent variables that are categorical and binary

Model of logistic regression:

Based on the relationship between the dependent variable and independent variables

Probability of the dependent variable taking a particular value is expressed as the odds ratio

Log of the odds ratio is known as the logit

Estimation:

Finding coefficients of independent variables that best fit the observed data

Maximize the likelihood function using iterative optimization algorithm

Model accuracy can be evaluated using various measures such as confusion matrix, accuracy, precision, recall, and F1-score.

Logistic regression:

What factors are important in predicting Y?

The factors that are important in predicting Y can be determined by examining the magnitude and direction of the coefficients of the independent variables in the logistic regression model. Variables with larger coefficients have a stronger impact on the probability of the dependent variable taking a particular value.

How does each factor affect Y? The interpretation of coefficients – odds ratio

The coefficients of the logistic regression model represent the log-odds ratio of the dependent variable taking the value of 1 for a one-unit increase in the independent variable, holding all other variables constant. The exponent of the coefficient is the odds ratio, which represents the change in odds of the dependent variable taking the value of 1 for a one-unit increase in the independent variable.

How to classify?

To classify an observation into one of the two classes, we use the logistic regression model to predict the probability of the dependent variable taking the value of 1. If the predicted probability is above a certain threshold, we classify the observation as belonging to the positive class (1), otherwise we classify it as belonging to the negative class (0). The threshold can be adjusted to optimize the tradeoff between sensitivity and specificity, depending on the particular problem at hand.

5. Generative models: LDA, QDA, and Naïve Bayes ……..

Bayers Theorem: Gives the probability of an event based on prior knowledge of related conditions.

Estimation: LDA/QDA: estimates probability distribution of independent variables for each class and uses Bayes theorem to compute posterior probability of each class.

Naive Bayes: estimates joint probability distribution of independent variables and dependent variable, assuming independence given the dependent variable.

Assumptions:

LDA: independent variables have multivariate normal distribution with equal covariance matrices across classes.

QDA: allows for different covariance matrices across classes.

Naive Bayes: assumes conditional independence of independent variables given dependent variable.

Bias-variance trade-off: Trade-off between model complexity and ability to generalize to new data.

Model suitability:LDA/QDA: suitable for classes with similar covariance matrices and linear/quadratic decision boundaries.

Naive Bayes: suitable for high-dimensional problems with many independent variables and conditional independence.

6.Model performance evaluation……..

There are several metrics to evaluate the performance of classification models, including:
Accuracy: measures the proportion of correct predictions among all predictions.
Precision: measures the proportion of true positive predictions among all positive predictions.
Recall: measures the proportion of true positive predictions among all actual positives.
F1-score: the harmonic mean of precision and recall.
Confusion matrix: a table that summarizes the number of correct and incorrect predictions.
ROC curve: a graphical representation of the trade-off between true positive rate (TPR) and false positive rate (FPR) for different classification thresholds.
AUC: the area under the ROC curve, which measures the overall performance of the classifier.

Accuracy rate is a performance metric that measures the proportion of correct predictions among all predictions in a classification model. It is calculated by dividing the number of correct predictions by the total number of predictions. While commonly used, accuracy may not always be the best measure of model performance, especially when classes are imbalanced. Other performance metrics such as precision, recall, F1-score, and the ROC curve should also be considered when evaluating classification models**.**

Confusion matrix is a performance metric that summarizes the number of correct and incorrect predictions in a classification model. It is a square matrix that displays the number of true positives, true negatives, false positives, and false negatives.

Why do we need confusion matrix?
A confusion matrix provides a more detailed evaluation of a classification model than just the overall accuracy. It allows us to see where the model is making errors and which types of errors it is making.

What is Type I and II error?
Type I error (false positive) occurs when the model predicts positive but the actual class is negative, while type II error (false negative) occurs when the model predicts negative but the actual class is positive.

How to evaluate model performance based on costs/benefits of different classes?
Depending on the application, different classes may have different costs or benefits associated with correct or incorrect predictions. In such cases, we can adjust the classification threshold to minimize the total cost or maximize the total benefit, taking into account the costs and benefits of each class.

What is sensitivity and specificity? How to calculate them?
Sensitivity (true positive rate) measures the proportion of actual positives that are correctly predicted as positive, while specificity (true negative rate) measures the proportion of actual negatives that are correctly predicted as negative. They are calculated as TP/(TP+FN) and TN/(TN+FP), respectively, where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of false positives.

How to use ROC curve to select models? AUC?
The ROC curve is a graphical representation of the trade-off between true positive rate (TPR) and false positive rate (FPR) for different classification thresholds. The area under the ROC curve (AUC) measures the overall performance of the classifier. AUC values range from 0.5 (random guessing) to 1 (perfect classifier). A classifier with higher AUC is generally considered better.

7.R Programming……..

Defining objects (number, string, vector):
In R, we can define objects of different types, including numeric (e.g. 1, 2.5), character or string (e.g. "hello"), and logical (TRUE or FALSE) values. We can also create vectors that contain multiple elements of the same type using functions like c() or seq().

What are the rules for variable names?
Variable names in R can consist of letters, numbers, and underscores, but they cannot start with a number. They are case sensitive, so "Var1" and "var1" are different variable names.

What is a function? How to call a function: e.g. seq()?
A function in R is a piece of code that performs a specific task. It takes one or more input values (arguments) and produces an output value. We can call a function by typing its name, followed by parentheses and any necessary arguments. For example, the function seq() generates a sequence of numbers and takes arguments for the start, end, and increment.

Read in a data set file (read.csv()):
In R, we can read in data from a CSV file using the read.csv() function. We specify the file path and name as the first argument and can specify other arguments to control how the data is read in.

Display variable names (names()) and data types (class()), etc:
In R, we can display the names of variables in a data frame using the names() function and the data types using the class() function. Other useful functions for exploring data include summary() to get summary statistics and head() or tail() to display the first or last few rows of data.

R basic plots: plot(), boxplot(), hist(), barplot():
R provides a variety of basic plotting functions that can be used to explore data. The plot() function creates scatter plots, line plots, or other types of graphs depending on the input data. The boxplot() function creates box and whisker plots to display the distribution of a variable. The hist() function creates histograms to display the frequency distribution of a variable. The barplot() function creates bar charts to display counts or other data summaries.

ggplot2: ggplot() + geom_xxx(aes()):
ggplot2 is a popular package in R for creating high-quality graphics. The basic syntax involves calling the ggplot() function and specifying the data frame to use, followed by adding one or more geometric objects (geoms) using functions like geom_point(), geom_line(), or geom_bar(), and specifying aesthetic mappings (aes) between variables and plot properties. For example, ggplot(data = my_data) + geom_point(aes(x = var1, y = var2)) creates a scatter plot of var1 and var2. ggplot2 offers a wide range of customization options and themes to create visually appealing plots.

lm() function to fit the model:
In R, the lm() function is used to fit a linear regression model to the data. The syntax for lm() is lm(formula, data), where formula specifies the relationship between the response variable and one or more explanatory variables, and data specifies the data frame containing the variables. For example, to fit a linear regression model with one explanatory variable (x) and one response variable (y), the syntax would be lm(y ~ x, data = my_data).

predict() function to predict the results:
Once the linear regression model has been fitted using lm(), the predict() function can be used to make predictions on new data. The syntax for predict() is predict(model, newdata), where model is the fitted linear regression model and newdata is the data frame containing the new data. For example, if we want to predict the values of y for a new set of x values (new_x), the syntax would be predict(my_model, newdata = data.frame(x = new_x)). The predict() function returns a vector of predicted values based on the input data.

Logistic regression

glm() function in R is used to estimate logistic regression models. It takes the form glm(y ~ x1 + x2 + ..., family = binomial(link = "logit")) where y is the binary response variable and x1, x2, etc. are the predictor variables. The family argument should be set to "binomial" for logistic regression models, and the link argument should be set to "logit".

predict() function in R is used to generate predicted probabilities of the response variable based on a fitted logistic regression model. It takes the form predict(model, newdata, type = "response") where model is the fitted logistic regression model, newdata is a data frame of predictor variables for which we want to make predictions, and type argument is set to "response" to obtain predicted probabilities.
ifelse() function in R can be used for binary classification based on predicted probabilities. It takes the form ifelse(predicted_prob >= threshold, 1, 0) where predicted_prob is the predicted probability of the response variable and threshold is a value between 0 and 1 that we use to classify observations as 0 or 1. If predicted_prob is greater than or equal to threshold, the function returns 1, otherwise it returns 0

LDA, QDA, and Naïve Bayes are all generative models for classification tasks.
In R, these models can be fitted using the lda(), qda(), and naiveBayes() functions, respectively.
The MASS package in R is commonly used for LDA and QDA, while the e1071 package is commonly used for Naïve Bayes.
LDA assumes that the classes have equal covariance matrices, while QDA assumes that the classes have different covariance matrices. Naïve Bayes assumes that the features are conditionally independent given the class.
Once these models are fitted, the predict() function can be used to make predictions on new data.
Performance of these models can be evaluated using a confusion matrix and various metrics such as accuracy, precision, recall, and F1 score.