CIDS - Assignment 2

1) Describe the role of Confusion Matrix.

A confusion matrix is a table that is used to define the performance of a classification algorithm. It visualizes and summarizes the performance of classification problem.

Predicted class \ Actual class		
	True positives	False positives
	False negatives	True negatives

2) Discuss the usage of Covariance matrix.

Covariance matrix is a type of matrix that is used to represent the covariance values blw the pair of elements given in a random vector. In this matrix, diagonal elements represent the variance and other elements represent covariance.

Structure of covariance matrix:

$$\begin{bmatrix} Var(X_1) & Cov(X_1,X_1) & \cdots & Cov(X_n,X_1) \\ \vdots & & Var(X_2) & \\ Cov(X_n,X_1) & \cdots & \cdots & Var(X_n) \end{bmatrix}$$

3) What is univariate and multivariate data?

Univariate data

This type of data consists of only one variable. It is the simplest form of analysis as it deals with only 1 quantity. The main purpose of this analysis is to describe the data and find the patterns that exist within it.

Ex: Height of a person in cm.

Multivariate data

When the data involves three or more variables, it is known as multivariate data

Example: Advertiser wants to compare the popularity of 4 advertisements on a website

4) What are different tools in Data Visualization?

1) Tableau
2) Looker
3) Zoho Analytics
4) Sisense
5) IBM Cognos Analytics
6) Qlik Sense
7) Domo
8) Microsoft Power BI
9) Klipfolio
10) SAP Analytics cloud
11) Yellowfin
12) Whatagraph
13) Dundas BI

5) Explain the importance of EDA in data science

Exploratory Data Analysis is used to analyze and summarize data sets and visualize the data to find the insights of data set

It provides a better understanding of the dataset variables and the relationship between data

Scientists use EDA to ensure results they produce are valid and applicable to any desired business goals and outcomes

6a) Discuss Ridge Regression in detail.

Ridge Regression decreases the standard error by adding penalty term to the regression coefficients. It is used in getting more accurate estimates, this regression is done by penalizing the needs of the future coefficients (L_2 regularization) and decreasing the value between the actual and predicted observations, further It prevents for ridge regression minimum $RSS + \alpha^2 \|\beta\|^2$

RSS : Residual sum of squares

β = weight of coefficient of independent last variables

α = regularization parameter that controls the strength of penalty term

$$RSS_{ridge}(w,n) = \sum_{i=1}^{n} (y_i - (w_i x_i + (b)^2)) + \alpha \sum_{j=1}^{p} w_j^2$$

fix training data well — keep parameter small

A trade off between fitting the training data well and keeping parameter small.

b) Explain the task of classification using Random Forest.

Random Forest is a supervised learning technique used for both classification and regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model

→ In random forest, we have no of decision trees on various subsets of given data sets and takes the average to improve the predictive accuracy instead of depending on 1 decision tree.

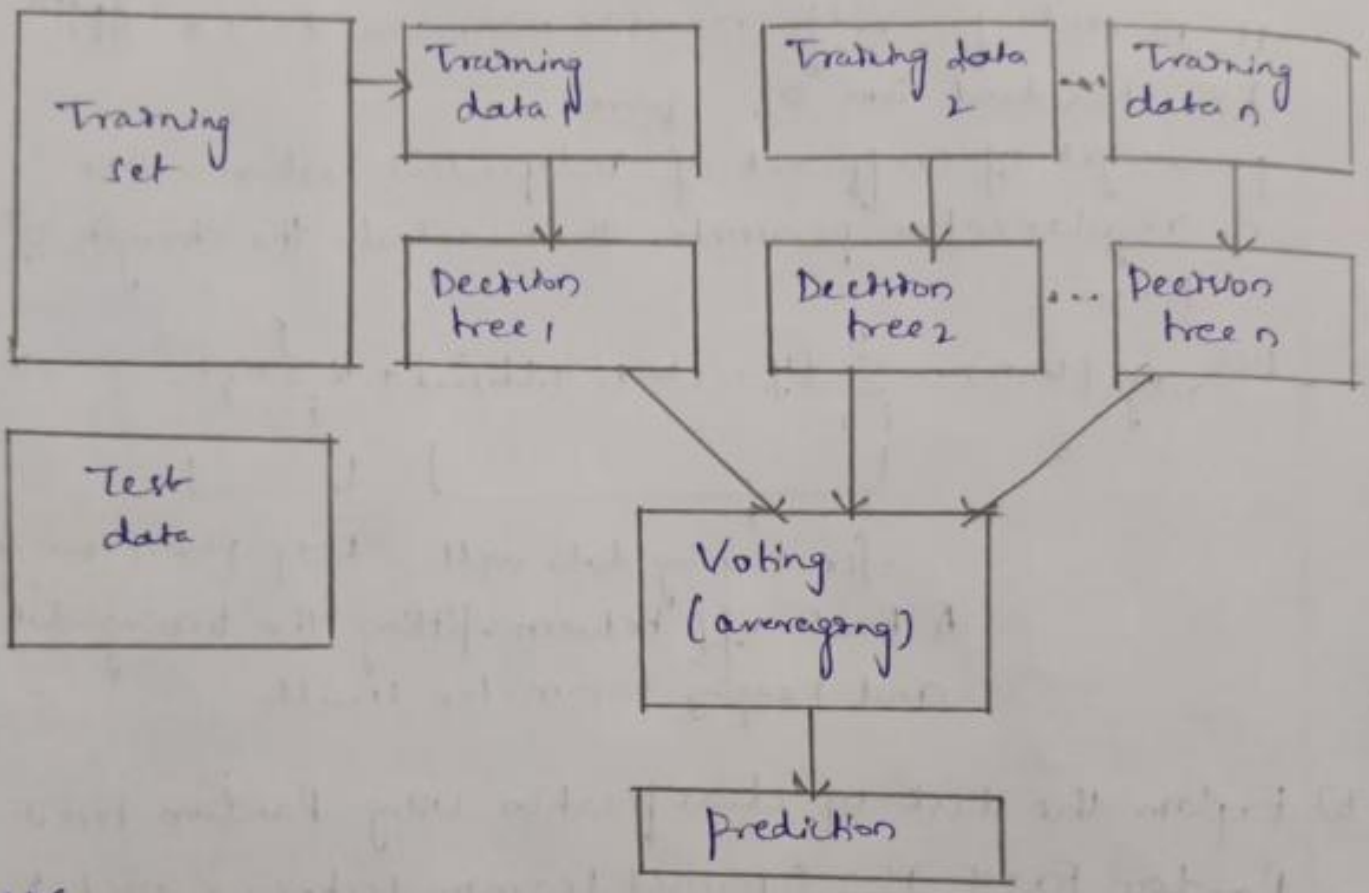
→ Based on the majority vote of prediction it predicts the final output.

→ The greater no. of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

→ This works in 2 phases:

i) It creates forest by combining n decision tree

ii) It makes prediction for each tree created in first phase.



Steps

i) In the random forest model, a subset of data points and a subset of features are selected for constructing each decision tree.

ii) Individual decision trees are constructed

iii) Each decision tree will generate an output

iv) Final output is based on majority voting or average for classification and regression. representation.

7) Explain K-Nearest Neighbor classifier with suitable example.

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on supervised learning technique.

→ K-NN algorithm stores all the available data and classifies a new data based on the similarity. This means when new data appears then it can be easily classified into a well-suite category by using K-NN algorithm.

→ K-NN algorithm can be used for regression and classification but mostly used for classification problems.

→ K-NN is a non-parametric algorithm, which means it doesn't make any assumption on underlying data.

Algorithm

Step-1 :- Select the number K of the neighbors

Step-2 :- Calculate the Euclidean distance of K number of neighbors

Step-3 :- Take the K nearest neighbors as per calculated Euclidean distance.

Step-4 :- Among these k-neighbors, count the no of data points in each category

Step-5 :- Assign the new data points to that category for which the no of neighbor is maximum

Step-6 :- Our model is ready.

Example

Consider age, income and a credit category of high or low for a bunch of people & Let's use age and income to

predict the credit label of high or low for a new person.
Consider the dataset with income represented in thousands

Age	Income	Credit
69	3	low
66	57	low
49	79	low
49	17	low
58	26	high
44	71	high

Consider a new person who is 57 years old and makes $70000
To calculate his credit label :
i) Decide a distance metric
ii) Split the original labeled dataset into training and test data
iii) Pick an evaluation metric
iv) Run k-NN few times, changing k and checking evaluation measure
v) Optimize k by picking the one with best evaluation measure
v) Create a new test with people's ages and incomes
From the above steps, the output by majority vote is a low credit score when k=5

Some Similarity Measures

1) Euclidean Distance

The dissimilarity (or similarity) blw the objects described by interval-scaled variables is typically computed based on distance blw each pair of objects

$$d(i,j) = \sqrt{(x_{i_1} - y_{i_1})^2 + (x_{i_2} - y_{i_2})^2 + \cdots + (x_{in} - y_{in})^2}$$

where $i = (x_{i_1}, x_{i_2}, \ldots x_{in})$, $j = (y_{j_1}, y_{j_2} \ldots y_{jn})$ are two n-dimensional objects

2) Manhattan (or city block) distance

The distance between 2 points is the sum of the absolute differences of their Cartesian coordinates

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \cdots |x_{i_n} - x_{j_n}|$$

3) Minkowski distance

It is generalization of both Euclidean and Manhattan distance. It is defined as

$$d(i,j) = \left(|x_{i_1} - x_{j_1}|^p + |x_{i_2} - x_{j_2}|^p + \cdots |x_{i_n} - x_{j_n}|^p\right)^{1/p}$$

4) Cosine Distance

This distance metric is used to calculate similarity blw 2 vectors. It is measured by the cosine of the angle blw 2 vectors and determines whether they are pointing in same direction

$$\text{Similarity}(A,B) = \cos\theta = \frac{\bar{A} \cdot \bar{B}}{\|\bar{A}\| \cdot \|\bar{B}\|}$$

θ = angle blw 2 vectors

$\|\bar{A}\| = \sqrt{A_1^2 + A_2^2 + \cdots A_n^2}$

5) Jaccard Distance

It is a common proximity measurement used to compute similarity blw 2 objects such as 2 text documents.

- It is used to compute similarity blw 2 asymmetric binary variables

$$J(i,j) = \text{sim}(i,j) = \frac{a}{a+b+c}$$

6) Mahalanobis Distance

It can be used blw 2 real-valued vectors and has the advantage over euclidean distance that it considers correlation and scale irrelevant

$$d(\bar{x}, \bar{y}) = \sqrt{(\bar{x}-\bar{y})^T S^{-1} (\bar{x}-\bar{y})}$$

S^{-1} covariance matrix

7) Hamming Distance

It is used to find distance b/w 2 strings or pair of words or DNA sequences of same length.

Ex) Distance b/w Olive and Ocean = 4 as except O, other letters are different

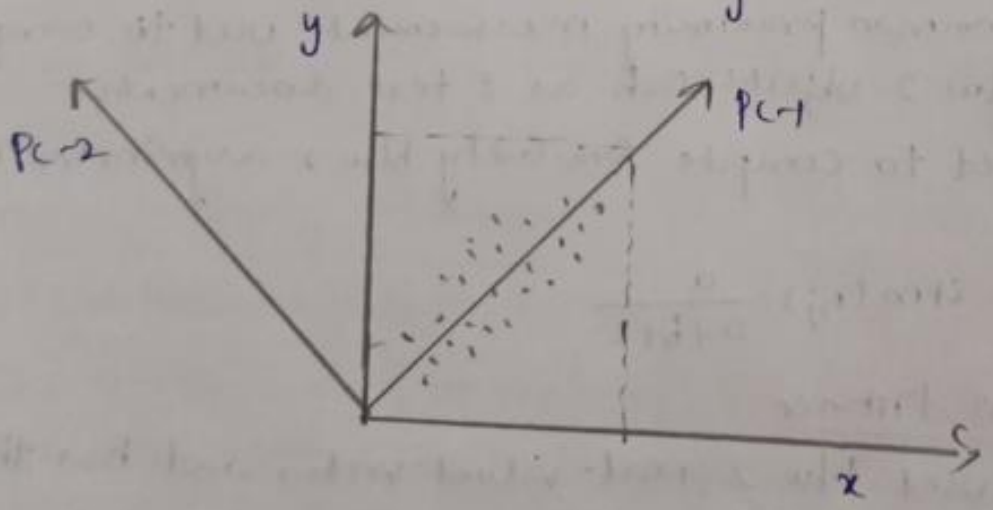
8) Describe the role of PCA in dimensionality reduction in detail.

Principle component Analysis (PCA) is an unsupervised learning technique to reduce the dimensionality. It increases interpretability at the same time minimizes information loss.

It is used to find the most significant features in the dataset and makes the data easy for plotting in the geometrical plane. While reducing dimensions, it preserves the most crucial data.

The original variables are converted into new set of variables called principle components which are the linear combinations of original variables.

The reduced dimensionality depends on how many principle components are used in the study.



Steps for PCA algorithm

1) Standardize the data to ensure that all the variables have a mean of 0 and standard deviation of 1

2) Calculate the covariance matrix. This matrix shows how each variable is related to every other variable in the dataset

3) Calculate eigen vectors and eigen values. The eigen vectors represent the direction in which data varies the most. Eigen values represent the amount of variation along each eigen vector

4) Choose the principle components: They are eigen vectors with highest eigen values. They represent the direction in which the data varies the most and are used to transform the original data into a lower dimensional space.

5) The final step is to transform original data into lower dimensional space defined by principle components.

10) Discuss basic principles, ideas and tools for data visualization

Basic Principles

i) Consistency

Maintain consistency in the use of colors, symbols, scales and so on across the visualization. Consistent use of elements makes it easier to understand the information

ii) Intuitive interpretation

Design visualizations in a way that allows intuitive interpretation users should be able to understand interpretation without more explanation. Use familiar symbols and conventions to enhance the understanding.

iii) Effective use of colors

Choose a color visually attractive & able in conveying the message.

iv) Proper Labelling

Label all axes, data points and any other relevant elements in the visualization. Title should always represent the main message of the visualization

v) Report

Arrange the elements of visualization in a logical sequence to convey the information

Ideas of Data Visualization

1) Choose the right chart type
2) Use correct plotting directions based on positive and negative values.
3) Do not use 'smoothed' line charts
4) Avoid confusing dual axis
5) Limit no.of slices displayed on a pie chart.
6) Label directly on the chart
7) Do not label on top of slices
8) Order pie slices for faster scanning
9) Avoid randomness
10) Focus on readability

Data Visualization tools

1) Tableau

It can take data and produce the required data visualization output in very short time.
Tableau also allow users to prepare, clean & format their data and then create data visualizations to obtain actionable insights.

2) Looker

It is a tool that can go indepth into the data & analyze it to obtain useful insights. It provides realtime dashboards of the data for more indepth analysis.

3) Zoho Analytics

It is a business intelligence and data analytics software. We can obtain data from multiple sources and mesh it together to create a multi-dimensional data visualizations.

4) Sisense

It is a business intelligence based data visualization system & it provides various tools that allows data analytics to simplify complex data and obtain insights for organization.

5) IBM cognos analytics

It is an AI based business intelligence platform that supports data analytics among other things. We can visualize and analyze data with anyone in organization.

6) Qlik Sense

It is a data visualization platform that helps companies to become data driven enterprises by providing associative data analytics engine, sophisticated AI system and suitable multi-cloud architecture.

II) Explain basic tools of Exploratory Data Analysis

1) Clustering and dimensionality reduction technique helps to create graphical displays of high dimensional data that contains many variables

2) Univariate visualization of each field in the raw data set with summary statistics

3) Bivariate visualizations and summary statistics - relationship blw each pair of variables

4) Multivariate visualization for mapping and understanding interactions between different fields in the data.

5) K-means clustering - It an unsupervised clustering method in which data points are divided into k-groups.
K-means clustering is usually utilized in market segmentation, image compression, & pattern recognition

6) Predictive models like linear regression, some statistics are used to accomplish EDA.

7) R: It an open source programming language widely used among statisticians in developing statistical observations and data analysis

8) Python: an interpreted, object oriented programming language. Python and EDA are often used together to spot missing values in the data set which is vital in machine learning

9) It is utilized in univariate, multivariate and bivariate visualization for summary statistics establishing relationships blw each variable & understanding how different fields interact with each other

9) Explain in detail about the following

i) Mahalanobis Distance

Mahalanobis distance is a statistical measure used to determine similarity blw 2 data points in a multidimensional space. It is instrumental in data analysis, pattern recognition & classification tasks.

Mahalanobis Distance (D) = $\sqrt{(x-m)^T \cdot \Sigma^{-1} (x-m)}$

where x: vector of the data

m: Vector of mean values of independent variables

Σ: Covariance matrix

Σ^{-1} : Inverse of covariance matrix

$x - m$: Distance of the data (vector) from the mean.

Applications

1) Outlier Detection
2) Credit Scoring
3) Image Recognition
4) Healthcare
5) Market Research

ii) Multivariate Normal Distribution

A multivariate normal distribution is a probability distribution that generalizes the univariate normal distribution to higher dimensions. In this, a set of random variables are jointly normally distributed.

Probability Density function

$$f(x, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where d = dimensionality of distribution

x = Vector of random variables

μ = mean vector

Σ : Covariance matrix

Applications

→ Widely used in statistics, finance and machine learning

→ It is a fundamental assumption in many statistical methods and models.