Unit VI INTRODUCTION TO ARTIFICIAL NEURAL NETWORK

Perceptron Learning - Biological Neuron, Introduction to ANN, McCulloch Pitts Neuron, Perceptron and its Learning Algorithm, Sigmoid Neuron, Activation Functions : Tanh, ReLu.

Multi-layer Perceptron Model - Introduction, Learning parameters : Weight and Bias, Loss function : Mean Square Error.

Introduction to Deep Learning (Chapter - 6)

TABLE OF CONTENTS

Unit III

Unit IV

Unit V

Unit VI

Unit III

3 Regression

3.1 : Introduction of Regression

Q.1 Define and explain regression with its model.

[SPPU : Dec-17, End Sem, Marks 4]

Ans. : • Regression finds correlations between dependent and independent variables. If the desired output consists of one or more continuous variable, then the task is called as regression.

• Therefore, regression algorithms help predict continuous variables such as house prices, market trends, weather patterns, oil and gas prices etc.

• Fig. Q.1.1 shows regression.

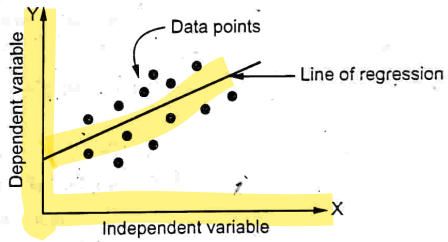


Fig. Q.1.1 Regression

• When the targets in a dataset are real numbers, the machine learning task is known as regression and each sample in the dataset has a real-valued output or target.

• Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of

the relationship between variables and for modelling the future relationship between them.

- The two basic types of regression are linear regression and multiple linear regression.

3.2 : Univariate Regression

Q.2 Explain univariate regression.

Ans. : • Univariate data is the type of data in which the result depends only on one variable. If there is only one input variable then we call it 'Single Variable Linear Regression' or 'Univariate Linear Regression.

- The function that we are trying to develop looks like this :

$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$y = mx + b$$

- That is because linear regression is essentially the algorithm for finding the line of best fit for a set of data.
- The algorithm finds the values for θ_0 and θ_1 that best fit the inputs and outputs given to the algorithm. This is called univariate linear regression because the ?? parameters only go up to 1.
- The univariate linear regression algorithm is much simpler than the one for multivariate.

Q.3 When is it suitable to use linear regression over classification ?

☞ [SPPU : Dec-16, End Sem, Marks 5]

Ans. : • Linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.

- The objective of a linear regression model is to find a relationship between the input variables and a target variable.
  1. One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
  2. The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

- Regression models predict a continuous variable, such as the sales made on a day or predict temperature of a city. Let's imagine that we fit a line with the training points that we have. If we want to add another data point, but to fit it, we need to change existing model.
- This will happen with each data point that we add to the model; hence, linear regression isn't good for classification models.
- Regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. Classification predicts categorical labels (classes), prediction models continuous - valued functions. Classification is considered to be supervised learning.
- Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. Prediction means models continuous - valued functions, i.e. predicts unknown or missing values.

Q.4 Why do we need to regularize in regression? Explain.

☞ [SPPU : Dec.-16, End Sem, Marks 5]

Ans. : • Regression model fails to generalize on unseen data. This could happen when the model tries to accommodate all kinds of changes in the data including those belonging to both the actual pattern and also the noise.

- As a result, the model ends up becoming a complex model having significantly high variance due to overfitting, thereby impacting the model performance (accuracy, precision, recall, etc.) on unseen data.

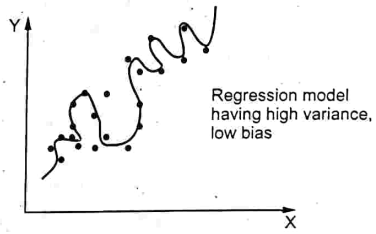


Fig. Q.4.1 Regression model before regularization

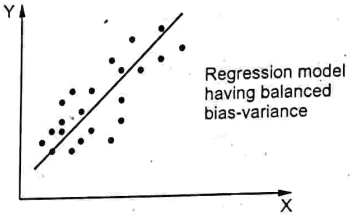


Fig. Q.4.2 Regression model after regularization

- Regularization needed for reducing overfitting in the regression model.
- Regularization techniques are used to calibrate the coefficients of the determination of multi-linear regression models in order to minimize the adjusted loss function.
- Regularization methods provide a means to control our regression coefficients, which can reduce the variance and decrease our of sample error.
- The goal is to reduce the variance while making sure that the model does not become biased (underfitting). After applying the regularization technique, the following model could be obtained.

Q.5 Explain least square method. ☞[SPPU : May-16, End Sem, Marks 5]

OR What do you mean by least square method? Explain least square method in the context of linear regression.

☞[SPPU : Dec-19, End Sem, Marks 5]

Ans. : • The method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data, on the one hand, and their expected values on the other.

- Considering an arbitrary straight line, $y = b_0 + b_1 x$, is to be fitted through these data points. The question is "Which line is the most representative" ?
- What are the values of b_0 and b_1 such that the resulting line "best" fits the data points ? But, what goodness-of-fit criterion to use to determine among all possible combinations of b_0 and b_1 ?

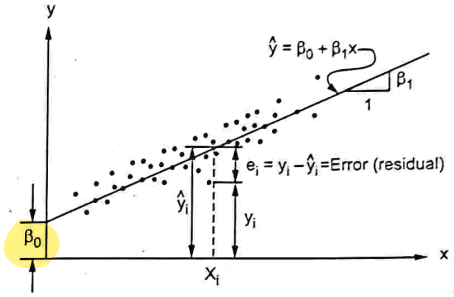


Fig. Q.5.1

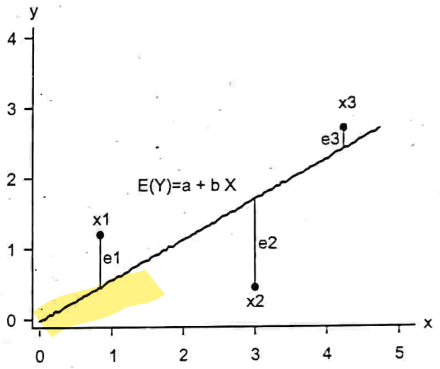


Fig. Q.5.2

- The Least Squares (LS) criterion states that the sum of the squares of errors is minimum. The least-squares solutions yields **y(x)** whose elements sum to 1, but do not ensure the outputs to be in the range [0,1].
- How to draw such a line based on data points observed ? Suppose a imaginary line of $y = a + bx$.
- Imagine a vertical distance between the line and a data point $E = Y - E(Y)$.

- This error is the deviation of the data point from the imaginary line, regression line. Then what is the best values of a and b ? A and b that minimizes the sum of such errors.
- Deviation does not have good properties for computation. Then why do we use squares of deviation ? Let us get a and b that can minimize the sum of squared deviations rather than the sum of deviations. This method is called **least squares.**
- Least squares method minimizes the sum of squares of errors. Such a and b are called least squares estimators i.e. estimators of parameters α and β.
- The process of getting parameter estimators (e.g., a and b) is called **estimation.** Lest squares method is the estimation method of Ordinary Least Squares (OLS).

Disadvantages of least square

1. Lack robustness to outliers
2. Certain datasets unsuitable for least squares classification
3. Decision boundary corresponds to ML solution

Q.6 Consider following data for 5 students.

Each Xi (i = 1 to 5) represents the score of i^{th} student in standard X and corresponding Yi (i = 1 to 5) represents the score of i^{th} student in standard XII.

i) What linear regression equation best predicts standard XII^{th} score ?

ii) Find regression line that fits best for given sample data.

iii) How to interpret regression equation ?

iv) If a student's score is 80 in std X, then what is his expected score in XII standard ?

Student	*Score in X standard (Xi)*	*Score in XII standard (Yi)*
1	*95*	*85*
2	*85*	*95*

3	*80*	*70*
4	*70*	*65*
5	*60*	*70*

Ans. : • The mean of the x values denoted by $\bar{X}$

- The mean of the y values denoted by $\bar{Y}$
- The standard deviation of the x values (denoted S_X)
- The standard deviation of the y values (denoted S_Y)

X_i	Y_i	X^2	XY	Y^2	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$
95	85	9025	8075	7225	17	8
85	95	7225	8075	9025	7	18
80	70	6400	5600	4900	2	– 7
70	65	4900	4550	4225	– 8	– 12
60	70	3600	4200	4900	– 18	– 7
ΣX= 390	ΣY=385	ΣX^2 = 31150	ΣXY = 30500	ΣY^2 = 30275		

Find : $\bar{X}$, $\bar{Y}$, S_X, S_Y

$$\bar{X} = \frac{\sum X_i}{n}, \quad \bar{Y} = \frac{\sum Y_i}{n}, \quad S_X = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}, \quad S_Y = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

$$\bar{X} = 390/5 = 78, \quad \bar{Y} = 385/5 = 77,$$

$$S_X = \sqrt{(390-78)^2 / 4} = \sqrt{24336} = 156$$

$$S_Y = \sqrt{(385-77)^2 / 4} = \sqrt{23716} = 154$$

We also need to compute the squares of the deviation scores :

X_i	Y_i	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X}) \times (Y_1 - \bar{Y})$
95	85	289	64	136
85	95	49	324	136
80	70	4	49	– 14
70	65	64	144	96
60	70	324	49	126
ΣX= 390	ΣY= 385	730	630	$\Sigma(X_i - \bar{X}) \times (Y_i - \bar{Y})$ = 480

The regression equation is a linear equation of the form : $\hat{Y} = \beta_0 + \beta_1 X$

First, we solve for the regression coefficient (β_1) :

$$\beta_1 = \sum[(X_i - \overline{X}) \times (Y_i - \overline{Y})] / \sum [(X_i - \overline{X})^2]$$

$$= 480 / 730 = 0.657$$

Once we know the value of the regression coefficient (β_1), we can solve for the regression slope (β_0) :

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

$$\beta_0 = 385 - 0.657 * 390 = 128.77$$

Therefore, the regression equation is $\hat{Y} = 128.77 + 0.657\ X$

Q.7 Consider following data :

i) Find values of β_0 *and* β_1 w.r.t. linear regression model which best fits given data.

ii) Interpret and explain equation of regression line.

iii) If new person rates "Bahubali -Part I" as 3 then predict the rating of same person for "Bahubali -Part II".

Person	X_i = Rating for movie "Bahubali Part - I" by i^{th} person	Y_i = Rating for movie "Bahubali Part - II" by i^{th} person
1st	4	3
2nd	2	4
3rd	3	2
4th	5	5
5th	1	3
6th	3	1

☞ [SPPU : Aug.-18, In Sem, Marks 6]

Ans. :

X_i	Y_i	X^2	XY	Y^2	$(X_i - \overline{X})$	$(Y_i - \overline{Y})$
4	3	16	12	9	1	0
2	4	4	8	16	− 1	1
3	2	9	6	4	0	− 1
5	5	25	25	25	2	2
1	3	1	3	9	− 2	0
3	1	9	3	1	0	− 2
$\sum X = 18$	$\sum Y = 18$	$\sum X^2 = 64$	$\sum XY = 57$	$\sum Y^2 = 64$		

Find : $\overline{X}$, $\overline{Y}$, S_X, S_Y

$$\overline{X} = \frac{\sum X_i}{n}, \quad \overline{Y} = \frac{\sum Y_i}{n}, \quad S_X = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}, \quad S_Y = \sqrt{\frac{\sum (Y_i - \overline{y})^2}{n-1}}$$

$$\overline{X} = 18/6 = 3, \quad \overline{Y} = 18/6 = 3,$$

$$S_X = \sqrt{(18-3)^2 / 5} = \sqrt{45}$$

$$S_Y = \sqrt{(18-3)^2 / 5} = \sqrt{45}$$

We also need to compute the squares of the deviation scores :

X_i	Y_i	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X}) \times (Y_1 - \overline{Y})$
4	3	1	0	0
2	4	1	1	− 1
3	2	0	1	0
5	5	4	4	4
1	3	4	0	0
3	1	0	4	0
ΣX= 18	ΣY= 18	10	10	$\Sigma(X_i - \overline{X}) \times (Y_i - \overline{Y}) = 3$

The regression equation is a linear equation of the form: $\hat{Y} = \beta_0 + \beta_1 X$

First, we solve for the regression coefficient (β_1) :

$$\beta_1 = \sum[(X_i - \overline{X}) \times (Y_i - \overline{Y})] / \sum [(X_i - \overline{X})^2]$$

$$= 3/10 = 0.3$$

Once we know the value of the regression coefficient (β_1), we can solve for the regression slope (β_0):

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

$$\beta_0 = 18 + 0.3 * 18 = 12.6$$

Therefore, the regression equation is $\hat{y} = 12.6 + 0.3\ X$

3.3 : Cost Function

Q.8 Define and explain Squared Error (SE) and Mean Squared Error (MSE) w.r.t Regression. ☞ [SPPU : May-18, End Sem, Marks 5]

Ans. : • The most common measurement of overall error is the sum of the squares of the errors, or SSE (sum of squared errors). The line with the smallest SSE is called the least-squares regression line.

- Mean Squared Error (MSE) is calculated by taking the average of the square of the difference between the original and predicted values of the data. It can also be called the quadratic cost function or sum of squared errors.
- The value of MSE is always positive or greater than zero. A value close to zero will represent better quality of the estimator/predictor. An MSE of zero (0) represents the fact that the predictor is a perfect predictor.

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (\text{Actual values} - \text{Predicted values})^2$$

- Here N is the total number of observations/rows in the dataset. The sigma symbol denotes that the difference between actual and predicted values taken on every i value ranging from 1 to n.
- Mean squared error is the most commonly used loss function for regression. MSE is sensitive towards outliers and given several examples with the same input feature values, the optimal prediction will be their mean target value. This should be compared with Mean

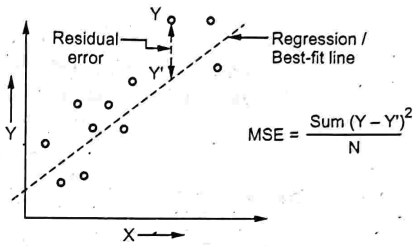


Fig. Q.8.1 Representation of MSE

Absolute Error, where the optimal prediction is the median. MSE is thus good to use if you believe that your target data, conditioned on the input, is normally distributed around a mean value, and when it's important to penalize outliers extra much.

- MSE incorporates both the variance and the bias of the predictor. MSE also gives more weight to larger differences. The bigger the error, the more it is penalized.
- Example : You want to predict future house prices. The price is a continuous value, and therefore we want to do regression. MSE can here be used as the loss function

Q.9 How the performance of a regression function is measured ? ☞ [SPPU : Dec-17, End Sem, Marks 4]

Ans. : • Following are the performance metrics used for evaluating a regression model :

a) Mean Absolute Error (MAE)
b) Mean Squared Error (MSE)
c) Root Mean Squared Error (RMSE)
d) R-squared
e) Adjusted R-squared

1. MAE :

- MAE is the sum of absolute differences between our target and predicted variables. So it measures the average magnitude of errors in a set of predictions, without considering their directions.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_j|$$

2. Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

3. Root Mean Square Error (RMSE)

- Root Mean Square Error (RMSE)is a standard way to measure the error of a model in predicting quantitative data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

4. R-squared

- R-squared is also known as the coefficient of determination. This metric gives an indication of how good a model fits a given dataset. It indicates how close the regression line is to the actual data values.
- The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

$$\text{R-squared} = 1 - \frac{\text{First sum of errors}}{\text{Second sum of errors}}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

5) Adjusted R-squared :

- The adjusted R-squared shows whether adding additional predictors improve a regression model or not.

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

Q.10 For a given data having 100 examples, if squared errors SE_1, SE_2, and SE_3 are 13.33, 3.33 and 4.00 respectively, calculate Mean Squared Error (MSE). State the formula for MSE.

☞ [SPPU : May-16, End Sem, Marks 5]

Ans. :

$$\text{Mean Squared Error} = \frac{\text{Squared Error}_1 + \text{Squared Error}_2 + \ldots + \text{Squared Error}_N}{\text{Number of data samples}}$$

$$\text{Mean Squared Error} = \frac{13.33 + 3.33 + 4.00}{100} = 0.2066$$

Q.11 What do you mean by coefficient of regression? Explain SST, SSE, SSR, MSE in the context of regression.

☞ [SPPU : Dec.-19, End Sem, Marks 5]

Ans. : • The quantities multiplied by the variables in a regression equation are called regression coefficients. Linear regression is the most common type of regression. The goal of linear regression is to determine which regression coefficients provide the best-fitting line.

- In linear regression, the regression coefficients assist in estimating the value of an unknown variable using a known variable. The regression coefficients analyses how the variables are dependent on other.
- **Regression line of X on Y** gives the best estimate for the value of X for any specific given values of Y :

$$\mathbf{X = a + b\,Y}$$

where a = X - intercept

b = Slope of the line

X = Dependent variable

Y = Independent variable

- Also Refer Q.8 and Q.9.

3.4 : Performance Evaluation

Q.12 Write short note on clustering.

Ans. : • Clustering groups data points based on their similarities. Each group is called a cluster and contains data points with high similarity and low similarity with data points in other clusters.

- The objective of clustering is to segregate groups with similar traits and bundle them together into different clusters.
- Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure has a range of [– 1, 1].
- Silhouette coefficients near + 1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.
- Many clustering algorithms use distance measures to determine the similarity or dissimilarity between any pair of data points. A valid distance measure should be symmetric and obtains its minimum value (usually zero) in case of identical data points. By computing the distance or (dis) similarity between each pair of observations, a dissimilarity or distance matrix is obtained.

3.5 : Optimizing Simple Linear Regression with Gradient Descent Algorithm

Q.13 List the derivative based optimization methods.

Ans. : • Derivative based optimization methods are used for :

1. Optimization of nonlinear neuro-fuzzy models
2. Neural network learning
3. Regression analysis in nonlinear models

Q.14 Explain gradient descent algorithm. Also explain its limitation.

Ans. : • Gradient descent is a first-order optimization algorithm. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient of the function at the current point.

- Gradient descent is popular for very large-scale optimization problems because it is easy to implement, can handle black box functions, and each iteration is cheap.
- Given a differentiable scalar field f (x) and an initial guess x_1 , gradient descent iteratively moves the guess toward lower values of "f" by taking steps in the direction of the negative gradient – ∇ f (x).
- Locally, the negated gradient is the steepest descent direction, i.e., the direction that x would need to move in order to decrease "f" the fastest. The algorithm typically converges to a local minimum, but may rarely reach a saddle point, or not move at all if x_1 lies at a local maximum.
- The gradient will give the slope of the curve at that x and its direction will point to an increase in the function. So we change x in the opposite direction to lower the function value :

$$X_{k+1} = x_k - \lambda \nabla f(x_k)$$

- The $\lambda > 0$ is a small number that forces the algorithm to make small jumps

Limitations of gradient descent :

- Gradient descent is relatively slow close to the minimum : technically, its asymptotic rate of convergence is inferior to many other methods.
- For poorly conditioned convex problems, gradient descent increasingly 'zigzags' as the gradients point nearly orthogonally to the shortest direction to a minimum point

Q.15 Explain steepest descent method.

Ans. : • Steepest descent is also known as gradient method.

- This method is based on first order Taylor series approximation of objective function. This method is also called saddle point method. Fig. Q.15.1 shows steepest descent method.

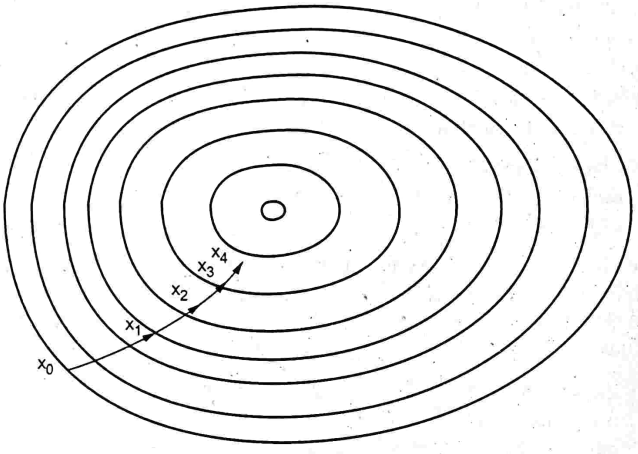


Fig. Q.15.1 Steepest descent method

- The steepest descent is the simplest of the gradient methods. The choice of direction is where f decreases most quickly, which is in the direction opposite to $\nabla f(x_i)$. The search starts at an arbitrary point x0 and then go down the gradient, until reach close to the solution.
- The method of steepest descent is the discrete analogue of gradient descent, but the best move is computed using a local minimization rather than computing a gradient. It is typically able to converge in few steps but it is unable to escape local minima or plateaus in the objective function.
- The gradient is everywhere perpendicular to the contour lines. After each line minimization the new gradient is always orthogonal to the previous step direction. Consequently, the iterates tend to zig-zag down the valley in a very inefficient manner.

- The method of steepest descent is simple, easy to apply and each iteration is fast. It also very stable; if the minimum points exist, the method is guaranteed to locate them after at least an infinite number of iterations.

3.6 : Multivariate Regression

Q.16 What is multivariate regression ?

☞ [SPPU : Oct-16, In Sem, Marks 5]

Ans. : • Multivariate regression involves multiple data variables for analysis. It is a supervised machine learning algorithm.

- Multivariate regression is used when there is a need to predict the value of a variable which is based on the value of two more variables. The value of the variable which is needed to be predicted is called the dependent variable.
- If multiple independent variables affect the response variable, then the analysis calls for a model different from that used for the single predictor variable. In a situation where more than one independent factor (variable) affects the outcome of a process, a multiple regression model is used. This is referred to as multiple linear regression model or multivariate least squares fitting.
- Let $z_1 ; z_2 ; \ldots ; z_r$ be a set of r predictors believed to be related to a response variable Y. The linear regression model for the j^{th} sample unit has the form

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \ldots + \beta_r z_{jr} + \varepsilon_j$$

where ε is a random error and $\beta_i, i = 0, 1, \ldots, r$ are un-known regression coefficients.

- With n independent observations, we can write one model for each sample unit so that the model is now

$$Y = Z\beta + \varepsilon$$

where Y is $n \times 1$, Z is $n \times (r+1)$, β is $(r+1) \times 1$ and ε is $n \times 1$

- In order to estimate β , we take a least squares approach that is analogous to what we did in the simple linear regression case.
- In matrix form, we can arrange the data in the following form :

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ : & : & & & : \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ : \\ y_N \end{bmatrix} \quad \hat{a} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ : \\ \hat{\beta}_k \end{bmatrix}$$

- where $\hat{\beta}_j$ are the estimates of the regression coefficients.

Q.17 What is multivariate regression? How will it be different from Univariate regression ?

☞ [SPPU : May-16, End Sem, Marks 5, Oct.-18, In Sem, Marks 4]

Ans. :

Simple regression	Multiple regression
One dependent variable Y predicted from one independent variable X	One dependent variable Y predicted from a set of independent variables $(X_1, X_2 \ldots X_k)$
One regression coefficient	One regression coefficient for each independent variable
r^2 : Proportion of variation in dependent variable Y predictable from X	R^2 : Proportion of variation in dependent variable Y predictable by set of independent variables (X's)

- Also Refer Q.16.

Q.18 What do you mean by zero-centering ?

☞ [SPPU, OCT.-16, In Sem, Marks 5]

OR What do you mean by zero centered and un-correlated features ? What is the use of it in the solution of multivariate linear regression ?

☞ [SPPU : May-18, End Sem, Marks 6]

Ans. : • Feature normalization is often required to neutralize the effect of different quantitative features being measured on different scales. If the

features are approximately normally distributed, we can convert them into z-scores by centring on the mean and dividing by the standard deviation. If we don't want to assume normality we can centre on the median and divide by the interquartile range.

- Sometimes feature normalization is understood in the stricter sense of expressing the feature on a [0,1] scale. If we know the feature's highest and lowest values h and l , then we can simply apply the linear scaling.
- Feature calibration is understood as a supervised feature transformation adding a meaningful scale carrying class information to arbitrary features. This has a number of important advantages. For instance, it allows models that require scale, such as linear classifiers, to handle categorical and ordinal features. It also allows the learning algorithm to choose whether to treat a feature as categorical, ordinal or quantitative.
- The goal of both types of normalization is to make it easier for your learning algorithm to learn. In feature normalization, there are two standard things to do :
  1. Centering : Moving the entire data set so that it is centered around the origin.
  2. Scaling : Rescaling each feature so that one of the following holds :
  a) Each feature has variance 1 across the training data.
  b) Each feature has maximum absolute value 1 across the training data.
- The goal of centering is to make sure that no features are arbitrarily large.

Q.19 List the characteristics of multivariate regression.

Ans. : Characteristics of multivariate regression are as follows :

- Multivariate regression allows one to have a different view of the relationship between various variables from all the possible angles.
- It helps you to predict the behavior of the response variables depending on how the predictor variables move.

- Multivariate regression can be applied to various machine learning fields, economic, science and medical research studies.

Q.20 What is difference between regression and correlation ?

Ans. :

Regression	Correlation
Regression tells us how to draw the straight line described by the correlation	Correlation describes the strength of a linear relationship between two variables
For regression only the dependent variable Y must be random.	For correlation, both variables should be random variables
Main goal is use the measure of relation to predict values of the random variable based on values of the fixed variable.	Main goal is simply to find a number that expresses the relation between the variables

3.7 : Introduction to Polynomial Regression

Q.21 What is a polynomial regression? How it can be represented in a form of a matrix. [SPPU : May-18, End Sem, Marks 5]

Ans. : • A polynomial regression consists of constants and variables that are combined using operations such as addition, subtraction and multiplication. Polynomial regression is a technique based on a trick that allows the use of linear models even when the dataset has strong non-linearities.

- With polynomial regression, the data is approximated using a polynomial function. A polynomial is a function that takes the form $f(x) = c_0 + c_1 x + c_2 x_2 \ldots c_n x_n$ where n is the degree of the polynomial and c is a set of coefficients.
- A polynomial of degree 0 (zero) is just a constant because $f(x) = c_0 x_0 = c_0$. Likewise preforming polynomial regression with a degree of 0 (zero) on a set of data returns a single constant value. It is the same as the mean average of that data.

- Linear regression is polynomial regression of degree 1, and generally takes the form $y = m x + b$ where m is the slope, and b is the y-intercept. It could just as easily be written $f(x) = c_0 + c_1 x$ with c_1 being the slope and c_0 the y-intercept.
- The order of the polynomial regression model depends on the number of features included in the model, so a model with m features is an n^{th} - degree or n^{th} - order polynomial regression. The general form of the design matrix with m-degrees looks like this :

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix}$$

- The matrix is always invertible as they follow the statistical rule of $m < n$ and thus become Vandermonde matrix. While it might be tempting to fit the curve and decrease error, it is often required to analyze whether fitting all the points makes sense logically and avoid overfitting.
- This is a highly important step as Polynomial Regression despite all its benefit is still only a statistical tool and requires human logic and intelligence to decide on right and wrong.

Q.22 Explain overfitting and underfitting.

[SPPU : Aug.-17, In Sem, Marks 5]

Ans. : Overfitting

- Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to overfitting and poor generalization.
- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Low error rates and a high variance are good indicators of overfitting. Fig. Q.22.1 shows overfitting.
- Overfitting is when a classifier fits the training data too tightly. Such a classifier works well on the training data but not on independent test data. It is a general problem that plagues all machine learning methods.

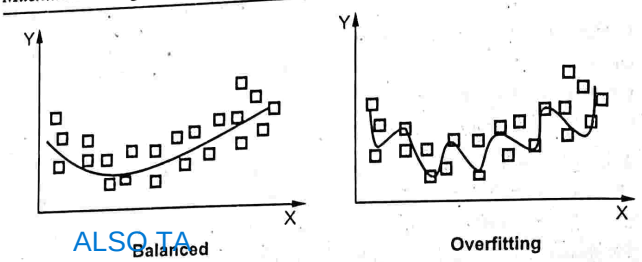


Fig. Q.22.1

- Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.
- We can determine whether a predictive model is under-fitting or overfitting the training data by looking at the prediction error on the training data and the evaluation data.
- Reasons for overfitting
  1. Noisy data
  2. Training set is too small
  3. Large number of features

ALSO TAKE Q24 IN CONSIDERATION

Underfitting :

- Underfitting happens when the learner has not found a solution that fits the observed data to an acceptable level.
- Underfitting : If we put too few variables in the model, leaving out variables that could help explain the response, we are underfitting.

Consequences :

1. Fitted model is not good for prediction of new data -prediction is biased
2. Regression coefficients are biased

Underfitting examples :

1. The learning time may be prohibitively large, and the learning stage was prematurely terminated.
2. The learner did not use a sufficient number of iterations.

3. The learner tries to fit a straight line to a training set whose examples exhibit a quadratic nature.

- The more difficult a criterion is to predict, the more noise exists in past information that need to be ignored. The problem is determining which part to ignore. Fig. Q.22.2 shows underfitting.

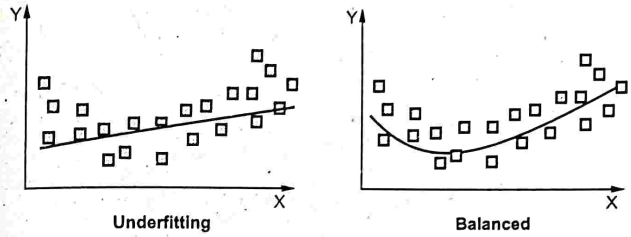


Fig. Q.22.2

- How do we know if we are underfitting or overfitting ?
  1. If by increasing capacity we decrease generalization error, then we are underfitting, otherwise we are overfitting.
  2. If the error in representing the training set is relatively large and the generalization error is large, then underfitting.
  3. If the error in representing the training set is relatively small and the generalization error is large, then overfitting;
  4. There are many features but relatively small training set.

Q.23 Explain bias-variance trade off.

Ans. : • In the experimental practice we observe an important phenomenon called the bias variance dilemma.

- In supervised learning, the class value assigned by the learning model built based on the training data may differ from the actual class value. This error in learning can be of two types, errors due to 'bias' and error due to 'variance'.
- Fig. Q.23.1 shows bias-variance trade off.
- Give two classes of hypothesis (e.g. linear models and k-NNs) to fit to some training data set, we observe that the more flexible hypothesis

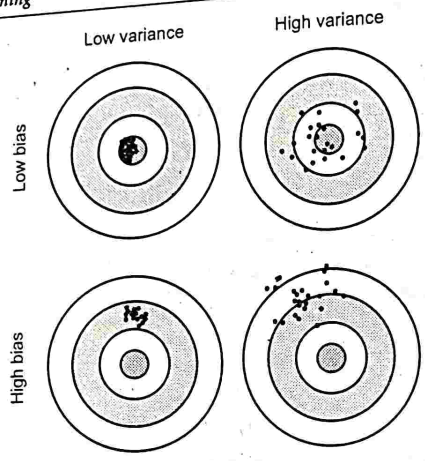


Fig. Q.23.1 Bias-variance trade off

class has a low bias term but a higher variance term. If we have parametric family of hypothesis, then we can increases the flexibility of the hypothesis but we still observe the increase of variance.

- The bias-variance-dilemma is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithm from generalizing beyond their training set :
  1. The bias is error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs.
  2. The variance is error from sensitivity to small fluctuations in the training set. High variance can cause overfitting : modeling the random noise in the training data, rather than the intended outputs.
- In order to reduce the model error, the designer can aim at reducing either the bias or the variance, as the noise components is irreducible.
- As the model increases in complexity, its bias is likely to diminish. However, as the number of training examples is kept fixed, the

parametric identification of the model may strongly vary from one DN to another. This will increase the variance term.

- At one stage, the decrease in bias will be inferior to the increase in variance, warning that the model should not be too complex. Conversely, to decrease the variance term, the designer has to simplify its model so that it is less sensitive to a specific training set. This simplification will lead to a higher bias.

Q.24 Explain the above Fig. Q.24.1 (a), (b) and (c).

☞ [SPPU : Oct.-19, In Sem, Marks 5]

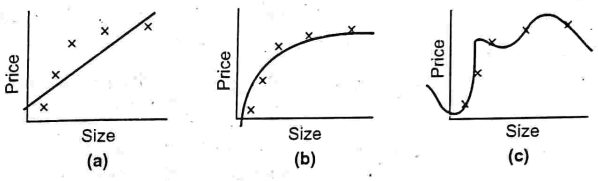


Fig. Q.24.1

Ans. : • Given Fig. Q.24.1 is related to overfitting and underfitting.

Underfitting (High bias and low variance) :

- A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.
- It usually happens when we have less data to build an accurate model and also when we try to build a linear model with a non-linear data.

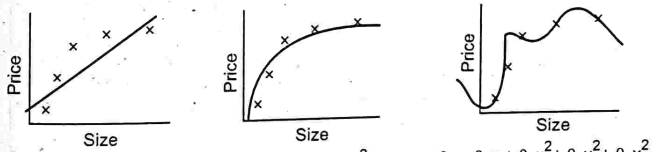


Fig. Q.24.2

- In such cases the rules of the machine learning model are too easy and flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions.
- Underfitting can be avoided by using more data and also reducing the features by feature selection.

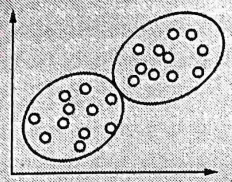
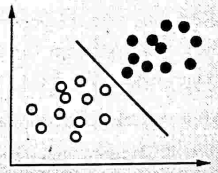
Overfitting (High variance and low bias) :

- A statistical model is said to be overfitted, when we train it with a lot of data.
- When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set.
- Then the model does not categorize the data correctly, because of too many details and noise.
- The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.
- A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Q.25 Explain difference between clustering and classification.

Ans. :

Clustering	Classification
This function maps the data into one of several clusters which is the grouping of data items based on the similarities between them.	This model function classifies the data into one of several predefined categorical classes.
Involved in unsupervised learning	Involved in supervised learning
Training sample is not provided	Training sample is provided

The number of cluster is not known before clustering, These are identified after the completion of clustering.	The number of classes is known before classification as there is predefined output based on input data.
Data is not labeled.	Labeled data points.
Asks how can I group this set of items?	Asks what class does this item belong to?
Unknown number of classes	Known number of classes
Used to understand data	Used to classify future observations
	

END...

Unit IV

4 Tree Based and Probabilistic Models

4.1 : Tree Based Model : Decision Tree

Q.1 What is tree based model ?

Ans. : Tree-based models use a series of if-then rules to generate predictions from one or more decision trees. All tree-based models can be used for either regression (predicting numerical values) or classification (predicting categorical values)

Q.2 What is decision tree ? Explain.

Ans. : • A decision tree is a simple representation for classifying examples. A decision tree or a classification tree is a tree in which each internal node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

- In this method a set of training examples is broken down into smaller and smaller subsets while at the same time an associated decision tree get incrementally developed. At the end of the learning process, a decision tree covering the training set is returned.
- The key idea is to use a decision tree to partition the data space into cluster (or dense) regions and empty (or sparse) regions.
- Decision tree consists of
  1. **Nodes :** Test for the value of a certain attribute.
  2. **Edges :** Correspond to the outcome of a test and connect to the next node or leaf.
  3. **Leaves :** Terminal nodes that predict the outcome.

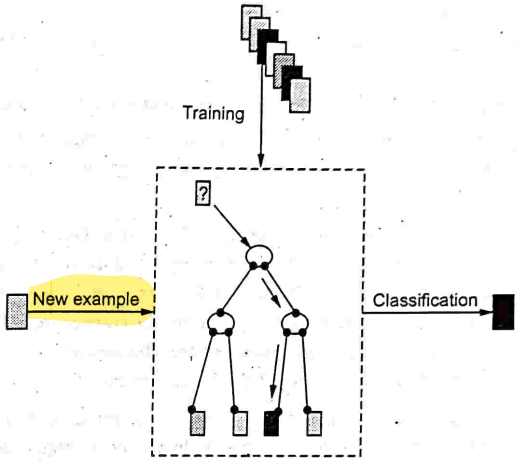


Fig. Q.2.1

- In Decision Tree Learning, a new example is classified by submitting it to a series of tests that determine the class label of the example. These tests are organized in a hierarchical structure called a decision tree.

Q.3 Write short on regression tree with its algorithm.

Ans. : • Regression tree models are known for their simplicity and efficiency when dealing with domains with large number of variables and cases. Regression trees are obtained using a fast divide and conquer greedy algorithm that recursively partitions the given training data into smaller subsets.

- When the complexity of the model is dependant on the learning sample size, both bias and variance decrease with the learning sample size. E.g. regression trees. Small bias, a tree can approximate any non linear function.
- Regression trees are among the machine learning method that present the highest variance. Even a small change of the learning sample can result in a very different tree. Even small trees have a high variance.

- Possible sources of variance :
  1. **Discretization of numerical attributes :** The selected threshold has a high variance.
  2. **Structure choice :** Sometimes, attribute scores are very close.
  3. **Estimation at leaf nodes :** Because of the recursive partitioning, prediction at leaf nodes are based on very small samples of objects.
- Regression trees are constructed using a Recursive Partitioning (RP) algorithm. This algorithm builds a tree by recursively splitting the training sample into smaller subsets. The RP algorithm receives as input a set of n data points and if certain termination criteria are not met it generates a test node t, whose branches are obtained by applying the same algorithm with two subsets of the input data points.
- At each node the best split test is chosen according to some local criterion which means that this is a greedy hill-climbing algorithm.

Algorithm : Recursive Partitioning Algorithm

Input : A set of n datapoints
Output : A regression tree
IF termination criterion THEN
Create Leaf Node and assign it a Constant Value
Return Leaf Node
ELSE
*Find Best Splitting Test s**
*Create Node t with s**
Left_branch (t) = RecursivePartitioningAlgorithm $([<x_i, y_i> : x_i \rightarrow s^*])$
Right_branch (t) = RecursivePartitioningAlgorithm $([<x_i, y_i> : x_i \rightarrow s^*])$
Return Node t
ENDIF

- The algorithm has three main components :
  1. A way to select a split test

2. A rule to determine when a tree node is terminal.
3. A rule for assigning a value to each terminal node.

Q.4 Explain advantages and disadvantages of decision tree.

Ans. : Advantages :

1. Rules are simple and easy to understand.
2. Decision trees can handle both nominal and numerical attributes.
3. Decision trees are capable of handling datasets that may have errors.
4. Decision trees are capable of handling datasets that may have missing values.
5. Decision trees are considered to be a nonparametric method.
6. Decision trees are self-explantory.

Disadvantages :

1. Most of the algorithms require that the target attribute will have only discrete values.
2. Some problem are difficult to solve like XOR.
3. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
4. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.

Q.5 Define feature tree.

Ans. : Feature tree is a tree such that each internal node is labelled with a feature, and each edge emanating from an internal node is labelled with a literal. The set of literals at a node is called a split. Each leaf of the tree represents a logical expression, which is the conjunction of literals encountered on the path from the root of the tree to the leaf.

Q.6 How empirical probabilities can be used in ranking and probability estimation trees ? Explain the purpose of pruning the subtree of a decision tree. ☞ **[SPPU : Dec-17, End Sem, Marks 8]**

Ans. :

- The class distributions in the leaves of an unlabelled feature tree can be used to turn one and the same tree into a decision tree, a ranking tree, or a probability estimation tree :

a) To turn a feature tree into a ranker, we order its leaves on non-increasing empirical probabilities, which is provably optimal on the training set;

b) To turn the tree into a probability estimator, we predict the empirical probabilities in each leaf, applying Laplace or m-estimate smoothing to make these estimates more robust for small leaves;

c) To turn the tree into a classifier, we choose the operating conditions and find the operating point that is optimal under those operating conditions.

- A tree is defined as a set of logical conditions on attributes ; a leaf represents the subset of instances corresponding to the conjunction of conditions along its branch or path back to the root. A simple approach to ranking is to estimate the probability of an instance's membership in a class and assign that probability as the instance's rank. A decision tree can easily be used to estimate these probabilities.
- Rule learning is known for its descriptive and therefore comprehensible classification models which also yield good class predictions. In some application areas, we also need good class probability estimates.
- A probabilistic rule is an extension of a classification rule, which does not only predict a single class value, but a set of class probabilities, which form a probability distribution over the classes. This probability distribution estimates all probabilities that a covered instance belongs to any of the class in the data set, so we get one class probability per class.
- Building decision trees with accurate probability estimates, called probability estimation trees. A small tree has a small number of leaves, thus more examples will have the same class probability. That prevents the learning algorithm from building an accurate PET.
- **Purpose of pruning the subtree of a decision tree.**
- Pruning simplifies a classifier by merging disjuncts that are adjacent in instance space. This can improve the classier's performance by eliminating error-prone components.

- Pruning of the decision tree is done by replacing a whole sub-tree by a leaf node. The replacement takes place if a decision rule establishes that the expected error rate in the sub-tree is greater than in the single leaf.

Q.7 Write the Grow Tree algorithm to generate feature tree? Explain the role of best split in this algorithm.

☞ [SPPU : May - 16, End Sem, Marks 9]

Ans. : • A feature tree is a tree such that each internal node is labelled with a feature, and each edge emanating from an internal node is labelled with a literal. The set of literals at a node is called a split.

- Each leaf of the tree represents a logical expression, which is the conjunction of literals encountered on the path from the root of the tree to the leaf. The extension of that conjunction is called the instance space segment associated with the leaf.
- Algorithm GrowTree(D, F)

  Input : data D; set of features F .

  Output : feature tree T with labelled leaves.

  1. if Homogeneous(D) then return Label(D);
  2. S ←BestSplit(D, F) ;
  3. split D into subsets D_i according to the literals in S ;
  4. for each i do
  5. if $D_i \neq \phi$; then $T_i \leftarrow$ GrowTree(D_i, F) ;
  6. else T_i is a leaf labelled with Label(D);
  7. end
  8. return a tree whose root is labeled with S and whose children are T_i.

- Algorithm gives the generic learning procedure common to most tree learners. It assumes that the following three functions are defined :
  1. Homogeneous(D) returns true if the instances in D are homogeneous enough to be labelled with a single label, and false otherwise;
  2. Label(D) returns the most appropriate label for a set of instances D;
  3. BestSplit(D, F) returns the best set of literals to be put at the root of the tree.

- These functions depend on the task at hand : For instance, for classification tasks a set of instances is homogeneous if they are of a single class, and the most appropriate label would be the majority class. For clustering tasks a set of instances is homogenous if they are close together, and the most appropriate label would be some exemplar such as the mean.

Q.8 If S is a collection of 14 examples with 9 YES and 5 NO examples then calculate entropy.

Ans. :

Entropy(S) = $\Sigma - p(I)\log_2 p(I)$

Where p(I) is the proportion of S belonging to class I.

Σ is over c.

$$\text{Entropy(S)} = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) = \mathbf{-0.940}$$

Q.9 Consider following splits having four features :

Length = [3, 4, 5] **[2 +, 0 –] [1+, 3 –] [2+, 2–]**
Gills = [Yes, No] **[0+, 4 –] [5+, 1–]**
Beak = [Yes, No] **[5+, 3 –] [0+, 2–]**
Teeth = [many, few] **[3+, 4 –] [2+, 1–]**

Find : Total weighted entropy and gini-index of all features.

☞ [SPPU : Dec.-18, In Sem, Marks 8]

Ans. : • Lets calculate the impurity of the first split. We have three segments : the first one is pure and so has entropy 0;

- The second one has entropy

$$= -(1/4)\log_2(1/4)\log_2 -(3/4)\log_2(3/4) = 0.5 + 0.31$$

$$= 0.81$$; the third one has entropy

- Similar calculations for the other three features give the following entropies :

$$\text{Gills} = (4/10)\times 0 + (6/10)\times(-5/6)\log_2(5/6) - (1/6)\log_2(1/6)) = 0.39$$

$$\text{Beak} = (8/10)\times(-5/8\log_2(5/8) - (3/8)\log_2(3/8)) + (2/10)\times 0 = 0.76$$

$$\text{Teeth} = (7/10)\times(-3/7\log_2(3/7) - (4/7)\log_2(4/7)) + (3\times 10)\times(-2/3)\log_2(2/3) - (1/3)\log_2(1/3)0 = 0.97$$

We thus clearly see that 'Gills' is an excellent feature to split on ; 'Teeth' is poor and the other two are somewhere in between.

4.2 : Probabilistic Model

Q.10 What is a probabilistic model ? Give an example of it.

☞ [SPPU : Dec.-17, End Sem, Marks 4]

Ans. : • Probabilistic modeling is a statistical technique used to take into account the impact of random events or actions in predicting the potential occurrence of future outcomes.

- In machine learning, we train the system by using a limited data set called 'training data' and based on the confidence level of the training data we expect the machine learning algorithm to depict the behaviour of the larger set of actual data.
- Probability theory provides a mathematical foundation for quantifying uncertainty of the knowledge.
- ML is focused on making predictions as accurate as possible, while traditional statistical models are aimed at inferring relationships between variables.
- We make observations using the sensors in the world. Based on the observations, we intend to make decisions. Given the same observations, the decision should be the same. However, the world changes, observations change, our sensors change, the output should not change.
- We build models for predictions; can we trust them? Are they certain? Many applications of machine learning depend on good estimation of the uncertainty :

a) Forecasting
b) Decision making
c) Learning from limited, noisy, and missing data
d) Learning complex personalised models
e) Data compression
f) Automating scientific modelling, discovery, and experiment design

- A signal is called random if its occurrence can not be predicted. Such signal can not be by any mathematical equation.
- The random signals are represented collectively by a random variable takes its value will be taken at particular time is not known.

Q.11 What is Bayes theorem ? Explain.

Ans. : • Bayes' theorem is a method to revise the probability of an event given additional information. Bayes's theorem calculates a conditional probability called a posterior or revised probability.

- Bayes' theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, P(A|B) denotes the conditional probability of A occurring, given that B occurs. The two conditional probabilities P(A|B) and P(B|A) are in general different.
- Bayes theorem gives a relation between P(A|B) and P(B|A). An important application of Bayes' theorem is that it gives a rule how to update or revise the strengths of evidence-based beliefs in light of new evidence a posteriori.
- A **prior probability** is an initial probability value originally obtained before any additional information is obtained.
- A **posterior probability** is a probability value that has been revised by using additional information that is later obtained.
- Suppose that $B_1, B_2, B_3 \ldots B_n$ partition the outcomes of an experiment and that A is another event. For any number, k, with $1 \leq k \leq n$, we have the formula :

$$P(B_k/A) = \frac{P(A/B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A/B_i) \cdot P(B_i)}$$

Q.12 Write short note on class probability estimation.

☞ [SPPU : May - 16, End Sem, Marks 8, Oct.-16, In Sem, Marks 5]

Ans. : • Probability estimates are also important when the classification outputs are not used in isolation but are combined with information from other components in a system.

- Let us consider handwritten character recognition example. Here the output from the classifier is used as input to a high level system which incorporates domain information.
- For many supervised learning tasks it is very costly to produce training data with class labels.
- Active learning acquires data incrementally, at each stage using the model learned so far to help identify especially useful additional data for labeling.
- Many applications require more than simple classification. Decision making often requires estimates of the probability of class membership.
- Class probability estimates can be combined with decision making costs/benefits to minimize expected cost.
- For example, in target marketing the estimated probability that a customer will respond to an offer is combined with the estimated profit. Other applications require ranking of cases, to add flexibility to user processing.
- Given the observation x and the class label y, we assume that the estimated pair-wise class probabilities r_{ij} of $\mu_{ij} = p(y = i \mid y = i \text{ or } j, x)$ are available.
- From the i^{th} and j^{th} classes of a training set, we obtain a model which, for any new x, calculates r_{ij} as an approximation of μ_{ij}.
- Then, using all r_{ij}, the goal is to estimate $p_i = p(y = i / x)$, for $i = 1, \ldots, k$.

Q.13 Write and explain Naïve Bayes classification algorithm.

☞ [SPPU, May-18, End Sem, Marks 8]

OR Explain Naïve Bayes classification algorithm.

☞ [SPPU, May-17, End Sem, Marks 8]

OR Is Naïve Bayes algorithm supervised or unsupervised task ? Explain how it achieves the task you specified.

☞ [SPPU : Dec.-16, End Sem, Marks 8]

Ans. : • Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

- Naïve Bayes assumes conditional independence over the training dataset. The classifier separates data into different classes according to the Bayes' Theorem. But assumes that the relationship between all input features in a class is independent. Hence, the model is called Naïve.
- Essentially, the Naive Bayes model is a conditional probability classification with Bayes Theorem applied.
- Conditional probability defines the probability of an event occurring given the occurrence of another event. The conditional probability is use it to calculate the joint probability, which is the probability of two or more simultaneous events. On the flip side, the joint probability could also be used to calculate the conditional probability, but it's often quite difficult to calculate the joint probability hence we use Bayes Theorem to calculate the conditional probability.
- A Naive Bayes classifier is a program which predicts a class value given a set of attributes.
- For each known class value,
  1. Calculate probabilities for each attribute, conditional on the class value.
  2. Use the product rule to obtain a joint conditional probability for the attributes.
  3. Use Bayes rule to derive conditional probabilities for the class variable.
- Once this has been done for all class values, output the class with the highest probability.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.
- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

Q.14 At a certain university, 4 % men are over 6 feet tall and 1 % women are over 6 feet tall. The total student population is divided in the ratio 3 : 2 in favor of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman ? ☞ [SPPU : Dec.-19, End Sem, Marks 8]

Ans. : • Let M be the event that student is male and F be the event that the student is female.

- Let T be the event that student is taller than 6 ft.

$$P(M) = 2/5$$

$$P(F) = 3/5$$

$$P(T|M) = 4/100$$

$$P(T|F) = 1/100$$

$$P(F/T) = \frac{P(T/F)\times P(F)}{(P(T/F)\times P(F)+P(T/M)\times P(M))}$$

$$= \frac{(1/100)\times(3/5)}{(1/100)\times(3/5)+(4/100)\times(2/5))} = \frac{3}{11}$$

Q.15 Consider following dataset and predict the class of new instance X using Navie Bayes classification algorithm.

Tid	*Refund*	*Material Status*	*Taxable Amount*	*Evade*
1	*Yes*	*Single*	*125 K*	*No*
2	*No*	*Married*	*100 K*	*No*
3	*No*	*Single*	*70 K*	*No*
4	*Yes*	*Married*	*120 K*	*No*
5	*No*	*Divorced*	*95 K*	*Yes*
6	*No*	*Married*	*60 K*	*No*
7	*Yes*	*Married*	*220 K*	*No*

8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes

X = (Refund = No, Martial status = Married, Income = 120 K)

☞ [SPPU : June-19, End Sem, Marks 10]

Ans. :

P(Refund = Yes|No) = 3/7
P(Refund = No|No) = 4/7
P(Refund = Yes|Yes) = 0
P(Refund = No|Yes) = 1
P(Marital Status = Single|No) = 2/7
P(Marital Status = Divorced|No) =1/7
P(Marital Status = Married|No) = 4/7
P(Marital Status = Single|Yes) = 2/7
P(Marital Status = Divorced|Yes) = 1/7
P(Marital Status = Married|Yes) = 0

For taxable income :

If class = No : sample mean = 110
sample variance = 2975
If class=Yes: sample mean = 90
sample variance = 25

$P(X|Class = No) = P(Refund=No|Class=No) \times P(Married|Class=No) \times P(Income = 120 K|Class=No)$

$= 4/7 \times 4/7 \times 0.0072 = 0.0024$

$P(X|Class=Yes) = P(Refund=No|\ Class=Yes) \times P(Married|\ Class=Yes) \times P(Income=120\ K|\ Class=Yes)$

$= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

Since $P(X|No)P(No) > P(X|Yes)P(Yes)$

Therefore $P(No|X) > P(Yes|X) \Rightarrow$ Class = No

Q.16 Consider following dataset and predict the class of new instance X using Navie Bayes.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Classification algorithm.

X = (Outlook = Sunny, Temp. = Cool, Humidity = High, Wind = Strong).

☞ [SPPU : May-19, End Sem, Marks 10]

Ans. : Learning phase

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5

Overcast	4/9	0/5
Rain	3/9	2/5

Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

P (Play = yes) = 9/14 P (Play = No) = 5/14

Text Phase

- Given a new instance, predict its label

x'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

- Look up tables achieved in the learning phrase

P(Outlook=Sunny|Play=Yes) = 2/9

P(Outlook=Sunny|Play=No) = 3/5

P(Temperature=Cool|Play=Yes) = 3/9

P(Temperature=Cool|Play=No)=1/5

P(Huminity=High|Play=Yes)=3/9

P(Huminity=High|Play=No)=4/5

P(Wind=Strong|Play=Yes)=3/9 P(Wind=Strong|Play=No)=3/5

P(Play=Yes)=9/14

P(Play=No) = 5/14

- Decision making

P(Yes|x') ≈ [P(Sunny|Yes)P(Cool|Yes)P(High|Yes) P(Strong|Yes)] P(Play=Yes)=0.0053

P(No|x') ≈ [P(Sunny|No) P(Cool|No) P(High|No) P(Strong|No)] P(Play=No)=0.0206

Given the fact P(Yes|x') < P (No |x'), we label x' to be "No".

END...

Unit V

5 Distance and Rule Based Models

5.1 : Distance Based Models

Q.1 Discuss various distance measures.

☞ [SPPU : May-17, (End Sem), Marks 9]

Ans. : 1) Euclidean distance :

- The Euclidean distance is the most common distance metric used in low dimensional data sets. It is also knows the L_2 norm. The Euclidean distance is the ususal manner in which distance is measured in real world.

$$d_{euclidean}(x, y) = \sqrt{\sum i(x_i - y_i)^2}$$

where x and y are m-dimensional vectors and denoted by $x = (x_1, x_2, x_3, \ldots, x_m)$ and $y = (y_1, y_2, y_3, \ldots, y_m)$ represent the m attribute values of two records.

- While Euclidean metric is useful in low dimensions, it doesn't work well in high dimensions and for categorical variables. The drawback of Euclidean distance is that it ignores the similarity between attributes. Each attributes is treated as totally different from all of the attributes.

2) Minkowski distance metric :

- Minkowski Distance is the generalized form of Euclidean and Manhattan Distance.
- The Minkowski distance between two variabes X and Y is defined as

$$D = \left(\sum_{i=1}^{n} |p_i - q_i|^P \right)^{1/p}$$

- The case where p = 1 is equivalent to the Manhattan distance and the case where p = 2 is equivalent to the Euclidean distance.

- Although p can be any real value, it is typically set to a value between 1 and 2. For values of p less than 1, the formula above does not define a valid distance metric since the triangle inequality is not satisfied.

3) Manhattan distance :

- Manhattan distance is a distance metric between two points in a N dimensional vector space. It is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes. In simple terms, it is the sum of absolute difference between the measures in all dimensions of two points.
- Manhattan Distance
  [{a, b, c}, {x, y, z}] → Abs [a – x] + Abs [b – y] + Abs [c – z]
- Manhattan distance is frequently used in:
- **Regression analysis :** It is used in linear regression to find a straight line that fits a given set of points
- **Compressed sensing :** In solving an underdetermined system of linear equations, the regularization term for the parameter vector is expressed in terms of Manhattan distance. This approach appears in the signal recovery framework called compressed sensing
- **Frequency distribution :** It is used to assess the differences in discrete frequency distributions

Q.2 What is hamming distance ?

Ans. : • Hamming bits are inserted into the message at the random locations. Hamming code is a single error correcting code. It is most complex from the stand point of creating and interpreting the error bits. Let us consider a frame which consists of m data bits and r check bits. The total length of message is then n = m + r. An n-bit unit containing data and checkbits is often referred to as an **n-bit codeword.**

- If 10001001 and 10110001 are two codewords, then the corresponding bits differ in these two codewrods is 3 bits. The number of bit positions in which two codewords differ is called the **hamming distance.**

Q.3 What do you mean distance metric and exemplar ? Explain different types of distance measures.

☞ [SPPU : Dec-19, (End Sem), Marks 9]

Ans. : • Distance metric uses distance function which provides a relationship metric between each elements in the dataset.

- When assessing how similar two data points or observations are, we need to calculate some sort of metric to be able to compare them. Distance Metrics allow us to numerically quantify how similar two points are by calculating the distance in between them.
- The distance calculation on complex/structured types requires three types of functions :
  1. A function to generate pairs of objects of the simpler constitutive types i.e. pairing function.
  2. Distance functions on the simpler types.
  3. An aggregation function that is applied to the distance values obtained from above steps.
- Distance measures are as Hamming Distance, Euclidean Distance, Manhattan Distance and Minkowski Distance. Also Refer Q.1.

5.2 : Neighbors and Examples

Q.4 What are neighbors ? Why is it necessary to use nearest neighbor while classifying ? ☞ [SPPU : May - 16, (End Sem), Marks 9]

Ans. : • To find a predefined number of training samples closest in distance to the new point, and predict the label from these.

- The number of samples can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning).
- The distance can, in general, be any metric measure : standard Euclidean distance is the most common choice.
- In the nearest neighbor algorithm, we classified a new data point by calculating its distance to all the existing data points, then assigning it the same label as the closest labeled data point.
- Despite its simplicity, nearest neighbors has been successful in a large number of classification and regression problems, including handwritten digits or satellite image scenes.

- Being a non-parametric method, it is often successful in classification situations where the decision boundary is very irregular
- Neighbors-based classification is a type of instance-based learning or non-generalizing learning : it does not attempt to construct a general internal model, but simply stores instances of the training data.
- Classification is computed from a simple majority vote of the nearest neighbors of each point : a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.
- The basic nearest neighbors classification uses uniform weights: that is, the value assigned to a query point is computed from a simple majority vote of the nearest neighbors.
- Under some circumstances, it is better to weight the neighbors such that nearer neighbors contribute more to the fit. This can be accomplished through the weights keyword.
- The default value, weights = 'uniform', assigns uniform weights to each neighbor. weights = 'distance' assigns weights proportional to the inverse of the distance from the query point.
- Alternatively, a user-defined function of the distance can be supplied which is used to compute the weights.

Q.5 Explain kNN algorithm with its advantages and disadvantages.

Ans. : • The k-nearest neighbor (kNN) is a classical classification method and requires no training effort, critically depends on the quality of the distance measures among examples.

- It is the simplest machine learning algorithms based on supervised learning technique. kNN algorithm assumes the similarity between the new data and available data and put the new data into the category that is most similar to the available categories.
- The k-nearest neighbour classification is one of the most popular distance-based algorithms. This classification is based on measuring the distances between the test sample and the training samples to determine the final classification output. The traditional k-NN classifier works naturally with numerical data. Fig. Q.5.1 shows kNN.

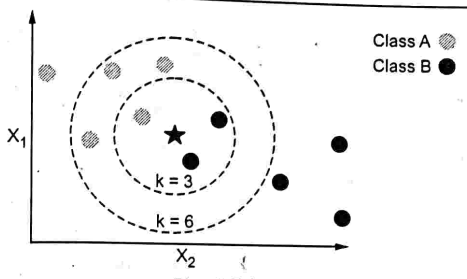


Fig. Q.5.1 kNN

- kNN stores all available data and classifies a new point based on the similarity. This algorithm also used for regression as well as for Classification but mostly it is used for the classification problems.
- kNN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- kNN, k is the number of nearest neighbors. The number of neighbors is the core deciding factor. k is generally an odd number if the number of classes is 2. When k=1, then the algorithm is known as the nearest neighbor algorithm.
- k-NN algorithm gives user the flexibility to choose distance while building k-NN model.
  a) Euclidean distance b) Hamming distance
  c) Manhattan distance d) Minkowski distance

The performance of the kNN algorithm is influenced by three main factors :

1. The distance function or distance metric used to determine the nearest neighbors.
2. The decision rule used to derive a classification from the k-nearest neighbors.
3. The number of neighbors used to classify the new example.

Advantages :

1. The kNN algorithm is very easy to implement.
2. Nearly optimal in the large sample limit.
3. Uses local information, which can yield highly adaptive behavior.
4. Lends itself very easily to parallel implementations.

Disadvantages :

1. Large storage requirements.
2. Computationally intensive recall.
3. Highly susceptible to the curse of dimensionality.

Q.6 Consider following data set. ☞ **[SPPU : Dec. - 17, (End Sem), Marks -10]**

X_1	X_2	Y
2	1	4
6	35	2
2	5	2
6	7	3
10	7	3
4	4	2
7	6	3

Model this function using the K - nearest neighbor regression. What will be the value of Y for the instance $(X_1, X_2) = (4, 5)$ and K = 3.

Ans. : • When K = 2, the nearest points are $X_1 = 2$; $X_2 = 5$ and $X_1 = 6$ and $X_2 = 7$.

• Taking the average of the outputs of these two points, we have,

$$Y = \frac{2+3}{2} = 2.5$$

• Similarly, when K = 3, we additionally consider the point $X_1 = 6$; $X_2 = 3$ to get output,

$$Y = \frac{2+3+2}{3} = 2.333$$

Q.7 Let on a scale of 1 to 10 (where 1 is lowest and 10 is highest), a student is evaluated by internal examiner and external examiner and accordingly student result can be pass or fail. A sample data is collected for 4 students. If a new student is rated by internal and external examiner as 3 and 7 respectively (test instance) decide new student's result using KNN classifier.

Student No.	(Xi1) Rating by internal examiner	(Xi2) Rating by external examiner	(Y) Result
S_1	7	7	Pass
S_2	7	4	Pass
S_3	3	4	Fail
S_4	1	4	Fail
Snew	3	7	?

☞ **[SPPU : Dec.-19, (End Sem), Marks 9]**

Ans. :

x_i = Input vector of dimension 2 = $\{x_{i1}, x_{i2}\}$

where, x_{i1} = Rating by internal examiner

x_{i2} = Rating by external examiner

For i = 1, 2, 3, 4 (i.e. 4 number of sample instances)

a) y_i = Result of a student and $y_i \in$ {pass, fail}

b) $x_0 = (x_{01}, x_{02}) = (3, 7)$

Step 1 : Let K = 3 (because number of classes = 2 and K should be odd)

Step 2 : Calculation of Euclidean distance between x_0 and x_1, x_2, x_3, x_4 as

$$d_{i0} = \sqrt{\sum_{i=1}^{2} (x_{i1} - x_{01})^2}$$

where, d = Number of features in input vector = 2

$$d_{10} = \sqrt{\sum_{i=1}^{2}(x_{i1} - x_{o1})^2}$$

$$= \sqrt{(x_{i1} - x_{o1})^2 + (x_{i2} - x_{02})^2}$$

$$= \sqrt{(7-3)^2 + (7-7)^2} = 4$$

Similarly $d_{10} = 4$

$d_{20} = 5$

$d_{30} = 3$

$d_{40} = 3.60$

Step 3 : Arranging all above distances in non decreasing order.

$(d_{30}, d_{40}, d_{10}, d_{20}) = (3, 3.60, 4, 5)$

Step 4 : Select K = 3 distance from above as (d_{30}, d_{40}, d_{60})

(3, 3.60, 4, 5)

Step 5 : Decide instances corresponding to 3-nearest instances. For given test instance 3, nearest neighbors are 3rd student, 4th student, 1st student, i.e. (3,4, fail), (1, 4, fail), (7, 7, pass).

Step 6 : Decide K_{pass} and K_{fail}

$K_{pass} = 1$

$K_{fail} = 2$

$K_{fail} > K_{pass}$

Step 7 : New student or test instance x_0 is classified to "fail" because K_{fail} is maximum.

5.3 : Clustering as a Learning Task

Q.8 Define the following :
Cluster, clustering, cluster analysis.

Ans. :

- **Cluster :** Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.

- **Clustering :** Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user understand the natural grouping or structure in a data set.
- **Cluster analysis :** Cluster analysis is process of grouping a set of data-objects into clusters.

Q.9 Write a note on clustering trees.

☞ [SPPU : Dec. - 16, (End Sem), Marks 9]

Ans. :
- Clustering is an exploratory data analysis task. It aims to find the intrinsic structure of data by organizing data objects into similarity groups or clusters.
- Let P be a finite set of points in an om-space. If the distances between different pairs of points in P are of different orders of magnitude, then the om-space imposes a unique tree-like hierarchical structure on P.
- The points will naturally fall into clusters, each cluster C being a collection of points all of which are much closer to one another than to any point in P outside C.
- The collection of all the clusters over P forms a strict tree under the subset relation.
- Moreover, the structure of this tree and the comparative sizes of different clusters in the tree captures all of the order-of-magnitude relations between any pair of points in P.
- A cluster tree is a tree T such that :
  1. Every leaf of T is a distinct symbol.
  2. Every internal node of T has at least two children.
  3. Each internal node of T is labelled with a non - negative value. Two or more nodes may given the same value.
  4. Every leaf of the tree is labelled 0.
  5. The label of every internal node in the tree is less than the label of its parent.

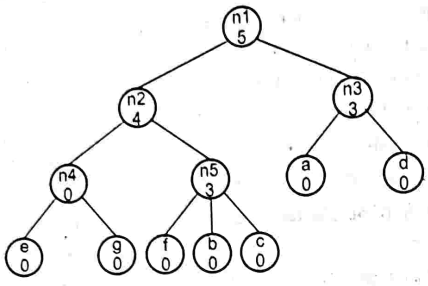


Fig. Q.9.1

- For any node N of T, the field "N.symbols" gives the set of symbols in the leaves in the subtree of T rooted at N, and the field "N.label" gives the integer label on node N.
- **Cluster tree construction :** This step uses a modified decision tree algorithm with a new purity function to construct a cluster tree to capture the natural distribution of the data without making any prior assumptions.
- **Cluster tree pruning :** After the tree is built, an interactive pruning step is preformed to simplify the tree to find meaningful / useful clusters. The final clusters are expressed as a list of hyper - rectangular regions.

Q.10 Write K-means algorithm. ☞ **[SPPU : May - 16, 17, (End Sem), Marks 9]**

Ans. : • K-Means clustering is heuristic method. Here each cluster is represented by the center of the cluster.

- The "K" stands for number of clusters, it is typically a user input to the algorithm; some criteria can be used to automatically estimate K.
- This method initially takes the number of components of the population equal to the final required number of clusters.
- In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart.

- Given K, the K-means algorithm consists of four steps :
  1. Select initial centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.
- The $x_1, \ldots, x_N$ are data points or vectors of observations. Each observation (vector x_i) will be assigned to one and only one cluster. The C(i) denotes cluster number for the i^{th} observation. K - means minimizes within cluster scatter :

$$W(C) = \frac{1}{2} \sum_{K=1}^{K} \sum_{C(i)=K} \sum_{C(j)=K} \|x_i - x_j\|^2$$

$$= \sum_{K=1}^{K} N_k \sum_{C(i)=K} \|x_i - x_j\|^2$$

where, m_k is the mean vector of the K^{th} cluster.

N_k is the number of observations in K^{th} cluster.

K-Means Algorithm Properties :

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

The K-Means Algorithm Process :

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point-
   a. Calculate the distance from the data point to each cluster.
   b. If the data point is closest to its own cluster, leave it where it is.
   c. If the data point is not closest to its own cluster, move it into the closest cluster.

3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intra-cluster distances and cohesion.

- K-means algorithm is iterative in nature. It converges, however only a local minimum is obtained. It works only for numerical data. This method easy to implement.

Q.11 Explain single linkage, complete linkage and average linkage.

☞ **[SPPU : Dec. - 16, (End Sem), Marks 9]**

Ans. : • In the agglomerative hierarchical approach, we start by defining each data point to be a cluster and combine existing clusters at each step.

• Here are four different methods for doing this :

1. Single linkage :

- Smallest pairwise distance between elements from each cluster. Also referred to as nearest neighbour or minimum method.
- This measure defines the distance between two clusters as the minimum distance found between one case from the first cluster and one case from the second cluster.
- For example, if cluster 1 contains cases a and b, and cluster 2 contains cases c, d, and e, then the distance between cluster 1 and cluster 2 would be the smallest distance found between the following pairs of cases : (a, c), (a, d), (a, e), (b, c), (b, d), and (b, e).

2. Complete linkage : Largest distance between elements from each cluster.

- Also referred to as furthest neighbor or maximum method. This measure is similar to the single linkage measure described above, but instead of searching for the minimum distance between pairs of cases, it considers the furthest distance between pairs of cases.
- Although this solves the problem of chaining, it creates another problem.

- Imagine that in the above example cases a, b, c, and d are within close proximity to one another based upon the pre-established set of variables; however, if case e differs considerably from the rest, then cluster 1 and cluster 2 may no longer be joined together because of the difference in scores between (a, e) and (b, e).
- In complete linkage, outlying cases prevent close clusters to merge together because the measure of the furthest neighbour exacerbates the effects of outlying data.

3. Average linkage : The average distance between elements from each cluster

- Also referred to as the unweighted pair-group method using Arithmetic averages.
- To overcome the limitations of single and complete linkage, Sokal and Michener proposed taking an average of the distance values between pairs of cases.
- This method is supposed to represent a natural compromise between the linkage measures to provide a more accurate evaluation of the distance between clusters.
- For average linkage, the distances between each case in the first cluster and every case in the second cluster are calculated and then averaged.

Q.12 List the problems associated with clustering.

Ans. : Problems with clustering are as follows :

1. Current clustering techniques do not address all the requirements adequately.
2. Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.
3. The effectiveness of the method depends on the definition of "distance".
4. If an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces.
5. The result of the clustering algorithm can be interpreted in different ways.

Q.13 What is hierarchical clustering ? List its types.

Ans. : • Hierarchical clustering arranges items in a hierarchy with a tree like structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a dendrogram.

- The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level.
- The hierarchical clustering algorithm is an unsupervised Machine Learning technique.
- Hierarchical clustering starts with k = N clusters and proceed by merging the two closest objects into one cluster, obtaining k = N – 1 clusters. The process of merging two clusters to obtain k – 1 clusters is repeated until we reach the desired number of clusters K.
- Hierarchical clustering algorithms are of two types : Divisive and Agglomerative
- Divisive methods initialize with all examples as members of a single cluster, and split this cluster recursively.
- Agglomerative methods begin instead with each point defining its own cluster, and iteratively link nearby points into larger and larger clusters.

Q.14 Explain steps for an agglomerative hierarchical cluster analysis.

Ans. : Steps are as follows :

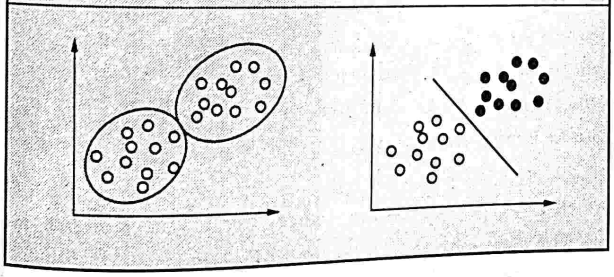
1. Find the similarity or dissimilarity between every pair of objects in the data set. In this step, we calculate the distance between objects using the pdist function. The pdist function supports many different ways to compute this measurement.
2. Group the objects into a binary, hierarchical cluster tree. In this step, we link pairs of objects that are in close proximity using the linkage function. The linkage function uses the distance information generated in step 1 to determine the proximity of objects to each other. As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed.
3. Determine where to cut the hierarchical tree into clusters. In this step, we use the cluster function to prune branches off the bottom of the

hierarchical tree, and assign all the objects below each cut to a single cluster. This creates a partition of the data. The cluster function can create these clusters by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point.

Q.15 Explain difference between clustering and classification.

Ans. :

Clustering	Classification
This function maps the data into one of several clusters which is the grouping of data items based on the similarities between them.	This model function classifies the data into one of several predefined categorical classes.
Involved in unsupervised learning	Involved in supervised learning
Training sample is not provided	Training sample is provided
The number of cluster is not known before clustering, These are identified after the completion of clustering.	The number of classes is known before classification as there is predefined output based on input data.
Data is not labeled.	Labeled data points.
Asks how can I group this set of items?	Asks what class does this item belong to?
Unknown number of classes	Known number of classes
Used to understand data	Used to classify future observations



Q.16 Consider following instance given as input to K-means clustering algorithm for k = 3. Find members of these 3 clusters after 2 iterations. X = {(2,10), (2,5), (8,4), (5,8), (7,5), (6,4), (1,2), (4,9)}

☞ [SPPU : Dec.-19, (End Sem), Marks 9]

Ans. : We rewrite the given data set :

Data set : $A_1(2,10)$, $A_2(2,5)$, $A_3(8,5)$, $B_1(5,8)$, $B_2(7,5)$, $B_3(6,4)$, $C_1(1,2)$, $C_2(4,9)$

Centroids : $A_1(2,10)$, $B_1(5,8)$, $C_1(1,2)$.

Iteration 1 : We need to calculate the distance between each data points and the centroids using the Euclidean distance.

Two points (x_1, y_1), (x_2, y_2)

Euclidean distance formula = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

or = $|x_2 - x_1| + |y_2 - y_1|$

Mean formula : $((x_1 + x_2)/2, (y_1 + y_2)/2)$

1st ROW :

Distance calculate between the A_2 data point and the Centroids A_1, B_1, C_1

Distance between $A_2(2,5)$ and $A_1(2,10)$ = $|2 - 2| + |5 - 10| = 0 + 5 = 5$

Distance between $A_2(2,5)$ and $B_1(5,8)$ = $|2 - 5| + |5 - 8| = 3 + 3 = 6$

Distance between $A_2(2,5)$ and $C_1(1,2)$ = $|2 - 1| + |5 - 2| = 1 + 3 = 4$

The A_2 nearby Cluster Center is C_1.

2nd ROW : Distance calculate between the A_3 data point and the Centroids A_1, B_1, C_1

Distance between $A_3(8,5)$ and $A_1(2,10)$ = 11

Distance between $A_3(8,5)$ and $B_1(5,8)$ = 6

Distance between $A_3(8,5)$ and $C_1(1,2)$ = 10

The A_3 nearby Cluster Center is B_1.

3rd ROW : Distance calculate between the B_2 data point and the Centroids A_1, B_1, C_1

Distance between $B_2(7,5)$ and $A_1(2,10)$ = 10

Distance between $B_2(7,5)$ and $B_1(5,8)$ = 5

Distance between $B_2(7,5)$ and $C_1(1,2)$ = 9

The B_2 nearby Cluster Center is B_1.

4th ROW : Distance calculate between the B_3 data point and the Centroids A_1, B_1, C_1

Distance between $B_3(6,4)$ and $A_1(2,10)$ = 10

Distance between $B_3(6,4)$ and $B_1(5,8)$ = 5

Distance between $B_3(6,4)$ and $C_1(1,2)$ = 7

The B_3 nearby Cluster Center is B_1.

5th ROW : Distance calculate between the C_2 data point and the Centroids A_1, B_1, C_1

Distance between $C_2(4,9)$ and $A_1(2,10)$ = 3

Distance between $C_2(4,9)$ and $B_1(5,8)$ = 2

Distance between $C_2(4,9)$ and $C_1(1,2)$ = 10

The C_1 nearby Cluster Center is B_1.

The above calculations are shown in the form of below table :

Data Sets	Centroids			Cluster
	$A_1(2,10)$	$B_1(5,8)$	$C_1(1,2)$	
$A_2(2,5)$	5	6	4	C_1
$A_3(8,5)$	11	6	10	B_1
$B_2(7,5)$	10	5	9	B_1
$B_3(6,4)$	10	5	7	B_1
$C_2(4,9)$	3	2	10	B_1

The we need to calculate the cluster mean values

Cluster $B_1(5,8)$ nearby points are $A_3(8,5)$, $B_2(7,5)$, $B_3(6,4)$, $C_2(4,9)$

B_1 Mean value = (6, 6.2)

Cluster $C_1(1,8)$ nearby points are $A_2(2,5)$

C_1 Mean value = (1.5, 3.5)

The updated Cluster points are : $A_1(2,10)$, $B_1(6, 6.2)$, $C_1(1.5, 3.5)$

Now we need to go for the next iteration with the updated cluster points.

Iteration 2 : Now we need to calculate the distance between the each data points to centroids

Data Sets	Centroids			Cluster
	$A_1(2,10)$	$B_1(6,6.2)$	$C_1(1.5,3.5)$	
$A_2(2,5)$	5	5.2	2	C_1
$A_3(8,5)$	11	3.2	8	B_1
$B_2(7,5)$	10	2.2	7	B_1
$B_3(6,4)$	10	2.2	5	B_1
$C_2(4,9)$	3	4.8	8	A_1

So, after completion of the iteration 2 the cluster points are not equal to the iteration 1 cluster points and then we need to go for the iteration 3 before that we need to calculate the cluster mean values.

Cluster $A_1(2,10)$ nearby points are $C_2(4,9)$

A_1 Mean value = (3, 9.5)

Cluster $B_1(6,6.2)$ nearby points are $A_3(8,5)$, $B_2(7,5)$, $B_3(6,4)$

B_1 Mean value = (6.7, 4)

Cluster $C_1(1.5,3.5)$ nearby points are $A_2(2,5)$

C_1 Mean value = (1.7, 4.2)

The updated Cluster points are : $A_1(3,9.5)$, $B_1(6.7, 4)$, C_1 = (1.7, 4.2)

Q.17 Define cluster tree ? Write and explain agglomerative clustering algorithm. ☞[SPPU : May-18, (End Sem), Marks 9]

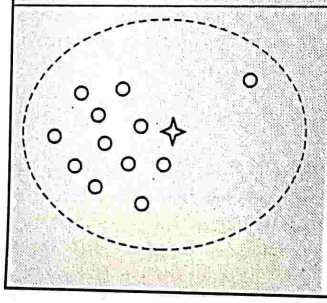
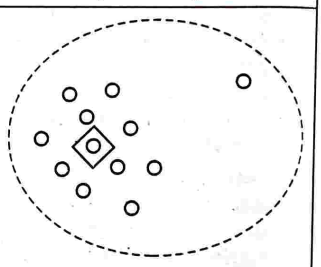
Ans. : • A cluster tree is a tree T such that : Every leaf of T is a distinct symbol. Every internal node of T has at least two children. Each internal node of T is labelled with a non-negative value. Two or more nodes may be given the same value.

- The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.

- This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed? For this there are many available methods. Some of them are :
  1) Single-nearest distance or single linkage.
  2) Complete-farthest distance or complete linkage.
  3) Average-average distance or average linkage.
  4) Centroid distance.
- This way we go on grouping the data until one cluster is formed. Now on the basis of dendogram graph we can calculate how many number of clusters should be actually present.
- Steps for an agglomerative hierarchical cluster analysis :
  1. Find the similarity or dissimilarity between every pair of objects in the data set. In this step, we calculate the distance between objects using the pdist function. The pdist function supports many different ways to compute this measurement.
  2. Group the objects into a binary, hierarchical cluster tree. In this step, we link pairs of objects that are in close proximity using the linkage function. The linkage function uses the distance information generated in step 1 to determine the proximity of objects to each other. As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed.
  3. Determine where to cut the hierarchical tree into clusters. In this step, we use the cluster function to prune branches off the bottom of the hierarchical tree and assign all the objects below each cut to a single cluster. This creates a partition of the data. The cluster function can create these clusters by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point.

Q.18 Explain the difference between K-mean and K-medoids clustering.

Ans. :

K-means	K-medoids
Each cluster is represented by the center of the cluster.	Each cluster is represented by one of the objects in the cluster.
Simple centroid-based method	Data point are chosen be the medoids.
K-means clustering is a non-hierarchical cluster analysis method that attempts to partition existing objects into one or more clusters or groups of objects based on their characteristics.	K-Medoid is a classic partition clustering technique that groups data sets of ni objects into k groups known a priori.
K-means algorithm is very prone to the effects of outliers.	Normally less delicate to outliers than K-means.
Convex shape is required.	Convex shape is not required.
	

Q.19 What is differences between k-NN and K-means clustering ?

Ans. :

k-NN	K-means clustering
KNN is a supervised learning algorithm.	k-Means clustering is an unsupervised learning.
kNN used for classification.	k-Means is used for clustering.
K-nearest neighbors needs labelled data to train on.	K-means clustering needs unlabelled data to train.
'K' in KNN is the number of nearest neighbours used to classify or a test sample.	K' in K-Means is the number of clusters the algorithm is trying to identify/learn from the data.
Centroid is the point X to be classified.	Centroids are not necessarily data points

5.4 : Association Rule Mining

Q.20 What is subgroup discovery ? ☞ [SPPU : May - 16, Dec. - 16, Marks 9]

Ans. : • Subgroup discovery is a method to identify relations between a dependent variable (target variable) and usually many explaining, independent variables.

- For example, consider the subgroup described by "smoker = true AND family history = positive" for the target variable coronary heart disease = true.
- Subgroup discovery does not necessarily focus on finding complete relations; instead partial relations, i.e., (small) subgroups with "interesting" characteristics can be sufficient.
- Subgroup discovery is a data analysis approach that aims at finding descriptions of subgroups of data instances with unusual statistical distribution of the property of interest.

- The discovered subgroup patterns must essentially satisfy two conditions. First, they have to be interpretable for the analyst, and second they need to be interesting according to the criteria of the user.
- Interestingness is typically defined by a quality function, which can take certain statistical or other user-defined quality criteria into account.
- For example, important parameters are the difference in the distribution of a (binary) target variable concerning the subgroup and the general population, and the subgroup size.
- The deviations of a subgroup from the performance of the general population are usually not simply due to statistical fluctuations, but are caused by local factors.
- Identifying these factors helps to understand the data in general and thus can provide useful insights for the analyst. Therefore, the main application areas of subgroup discovery are exploration and descriptive induction.
- Subgroup discovery task mainly relies on the following four properties : the target variable, the subgroup description language, the quality function, and the search strategy
- The target variable may be binary, nominal or numeric. Depending on its type, there are different analytic questions, e.g., we can search for significant deviations of the mean of a numeric target variable. The description language specifies the individuals from the genera population belonging to the subgroup.
- Subgroup description languages can be either single-relational or multi-relational. In the case of single-relational propositional languages a subgroup description can be defined as follows : Let Ω_A the set of all attributes with an associated domain dom(a) of values. V_A is defined as the (universal) set of attribute values of the form $(a = v)$, $a \in \Omega_A$, $v \in dom(a)$.
- Example applications of subgroup discovery include the analysis of clinical data, marketing analytics, gene expression data analysis, analysis of e-learning, and analysis of traffic accidents.

- In each of these applications, the problem was conveniently represented with data in the attribute-value form, and classification rules were used to describe the subgroups.

Q.21 Define :
i) Frequent itemset ii) Support iii) Confidence
iv) Market basket analysis. ☞ [SPPU : May - 16, (End Sem), Marks 8]

Ans. :

i) **Frequent itemset :** The set of items a customer will buy is referred to as an itemset.

ii) **Support :** The probability that a customer will buy beer without a bar meal (i.e. that the antecedent is true) is referred to as the support for the rule.

iii) **Confidence :** The conditional probability that a customer will purchase crisps is referred to as the confidence.

iv) **Market basket analysis :** Market basket analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another grouo of items. For example, if you are in an English pub and you buy a pint of beer and don't buy a bar meak, you are more likely to buy crisps at the same time than somebody who didn't buy beer.

Q.22 Using the following data, find 2-item-itemsets which have minimum support = 2. ☞ [SPPU : Dec. - 16, (End Sem), Marks 8]

Transaction	Items
1	nappies
2	beer, crisps
3	apples, nappies
4	beer, crisps, nappies
5	apples
6	apples, beer, crisps, nappies
7	apples, crisps
8	crisps

Ans. : Any give association rule has a support level and a confidence level. Support it the percentage of the population which satisfies the rule. If the percentage of the population in which the attendant is satisfied is s, then the confidence is that percentage in which the consequent is also satisfied.

- This idea of association rules was later generalized to any data in the tabular, attribute-value form. So data describing properties (values of attributes)of some examples can be analyzed in order to find associations between con-junctions of attribute-value pairs (categories) called antecedent (Ant) and succedent (Suc).
- The two basic characteristics of an association rule are support and confidence. Let a be the number of examples that fulfill both Ant and Suc, b the number of examples that fulfill Ant but do not fulfill Suc, c the number of examples that fulfill Suc but do not fulfill Ant, and d the number of examples that fulfill neither Ant nor Suc. Then Support is then defined as

$$sup = P(Ant \wedge Suc) = \frac{a}{a+b+c+d}$$

and confidence is defined as

$$con\ f = P(Suc|Ant) = \frac{a}{a+b}$$

Transaction	Items
1	nappies
2	beer, crisps
3	apples, nappies
4	beer, crisps, nappies
5	apples
6	apples, beer, crisps, nappies
7	apples, crisps
8	crisps

- Each transaction in this table involves a set of items; conversely, for each item we can list the transactions in which it was involved :

transactions 1, 2, 3 and 6 for nappies, transactions 3, 5, 6 and 7 for apples, and so on

Total support = 8

Support (beer, crisps) = 3/8 = 0.375

Confidence = 3/5 = 0.625

Q.23 Define with respect to association rule mining :
i) Support ii) Confidence iii) Lift ☞ **[SPPU : May - 18, (End Sem), Marks 8]**

Ans. : i) Support :

Usefulness of a rule can be measured with a minimum support threshold

This parameter lets to measure how many events have such itemsets that match both sides of the implication in the association rule

Rules for events whose itemsets do not match boths sides sufficiently often (defined by a threshold value) can be excluded.

The support supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set.

$$Supp(X) = \frac{\text{Number of transactions which contain the item set X}}{\text{Total number of transactions}}$$

ii) Confidence :

Certainty of a rule can be measured with a threshold for confidence

This parameter lets to measure how often an event's itemset that matches the left side of the implication in the association rule also matches for the right side

Rules for events whose itemsets do not match sufficiently often the right side while mathching the left (defined by a threshold value) can be excluded

The **confidence** of a rule is defined :

$$conf\ (X \rightarrow Y) = \frac{supp\ (X \cup Y)}{supp\ (X)}$$

- Let's assume D = {(1,2,3), (2,3,4), (1,2,4), (1,2,5), (1,3,5)}

$$supp\ (1,2) = \frac{3}{5}$$

- From itemset 3 => {(1,2,3), (1,2,4), (1,2,5)} That is : Relation of number of events containing itemset (1,2) to number of all events in database
- Let's assume D = {(1,2,3), (2,3,4), (1,2,4), (1,2,5), (1,3,5)}
- The confidence for rule 1 → 2

$$\text{conf}((1,2)) = \frac{\sup(1 \cup 2)}{\sup(1)} = \frac{\frac{3}{5}}{\frac{4}{5}} = \frac{3}{4}$$

- That is: relation of number of events containing both itemsets X_a and X_b to number of events containing an itemset X_a.
- If confidence gets a value of 100 % the rule is an exact rule.
- Even if confidence reaches high values the rule is not useful unless the support value is high as well.
- Rules that have both high confidence and support are called **strong rules.**
- Some competing alternative approaches (other that Apriori) can generate useful rules even with low support values .

iii) Lift or lift ratio : It is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.

$$\text{Lift ratio} = \frac{(A+B)/A}{(B/\text{Total})} = \frac{\text{Confidence}}{(B/\text{Total})}$$

Q.24 What is Apriori algorithm ? Explain use and limitation of Apriori algorithm.

Ans. : • Discovery methods for frequent concept sets in text mining build on the Apriori algorithm of by Agrawal and Srikant and used in data mining for market basket association problems.

- The Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules.
- Apriori is a classic algorithm for learning association rules. It is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).
- In text mining, **associations** specifically refer to the directed relations between concepts or sets of concepts. An association rule is generally an expression of the form A ⇒ B, where A and B are sets of features.
- An association rule A ⇒ B indicates that transactions that involve A tend also to involve B.
- Following the original definition by Agrawal the problems of association rule mining is defined as :
- Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I.
- A rule is defined as an implication of the form XY where $X, Y \subseteq I$ and $X \cap Y = \phi$
- The sets of items X and Y are called antecedent (left-hand-side) and consequent (right-hand-side) of the rule respectively.
- Use of Apriori algorithm :
  1. Initial information : transactional database D and user-defined numeric minimun support threshold min_sup
  2. Algortihm uses knowledge from previous iteration phase to produce frequent itemsets
  3. This is reflected in the Latin origin of the name that means "from what comes before".

Limitations of Apriori algorithm :

1. Needs several iterations of the data
2. Uses a **uniform** minimum support threshold

3. Difficulties to find rarely occuring events
4. Alternative methods (other than appriori) can address this by using a **non-uniform** minimum support thresold
5. Some competing alternative approaches focus on **partition** and **sampling.**

Q.25 Explain with example frequent itemset and close itemset.

Ans. :

1. A set of items is referred to as an itemset. An itemset is a unordered set of distinct items. An itemset that contains k items is a k-itemset.
2. The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency, support count, or count of the itemset.
3. Frequent itemsets that cannot be extended with any item without making them infrequent are called maximal frequent itemsets. Exact support counts of the subsets cannot be directly derived from support of the maximal frequent itemset.

Closed itemsets :

- An alternative approach is to try to retain some of the support information in the compacted representation.

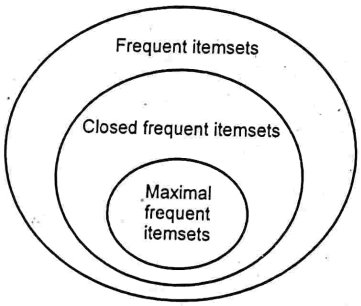


Fig. Q.25.1

- A closed itemset is an itemset whose all immediate supersets have different support count.

- A closed frequent itemset is a closed itemset that satisfies the minimum support threshold.
- Maximal frequent itemsets are closed by definition.
- An itemset X is closed in a data set S if there exists no proper super-itemset Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is both closed and frequent in S.
- An itemset is closed if none of its immediate supersets has the same support as the itemset.
- Closed itemset example 1 :

TID	Items
1	{A, B}
2	{B, C, D}
3	{A, B, C, D}
4	{A, B, D}
5	{A, B, C, D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A, B}	4
{A, C}	2
{A, D}	3
{B, C}	3
{B, D}	4
{C, D}	3

Itemset	Support
{A, B, C}	2
{A, B, D}	3
{A, C, D}	2
{B, C, D}	3
{A, B, C, D}	2

- Closed itemset are : {B}, {A, B}, {B, D}, {A, B, D}, {B, C, D}, {A, B, C, D}

- **Closed itemset example 2 :**

TID	Items
100	a, c, d, e, f
200	a, b, e
300	c, e, f
400	a, c, d, f
500	c, e, f

- Total Frequent itemsets : 20

{a}, {c}, {d}, {e}, {f}, {a, c} {a, d}, {a, e}, {a, f}, {c, d}, {c, e}, {c, f}, {d, f}, {e, f}, {a, c, d},

{a, c, f}, {a, d, f}, {c, d, f}, {c, e, f} {a, c, d, f}

Closed frequent itemsets :

{a, c, d, f}, {c, e, f}, {a, e}, {c, f}, {a}, {e}

Q.26 Find all association rules in the following database in the following database with minimum support = 2 and minimum confidence = 65 %.

Transactions	*Data Items*
T_1	*Milk, Bread, Cornflakes*
T_1	*Bread, Jam*
T_1	*Milk, Bread, Cornflakes, Jam*
T_1	*Milk, Cornflakes*
T_1	*Bread, Butter, Jam*
T_1	*Bread, Butter*
T_1	*Milk, Bread, Butter*

☞ **[SPPU : Dec.-18, (End Sem), Marks 10]**

Ans. :

TID	Items
T1	Milk, Bread, Cornflakes
T2	Bread, Jam
T3	Milk, Bread, Cornflakes, Jam
T4	Milk, Cornflakes
T5	Bread, Butter, Jam
T6	Bread, Butter
T7	Milk, Bread, Butter

Itemset	Sup
Milk	4
Bread	6
Cornflakes	3
Jam	3
Butter	3

Itemset	Sup
Milk, Bread	3
Milk, Cornflakes	3
Milk, Jam	1
Milk, Butter	1
Bread, Cornflakes	2
Bread, Jam	3
Bread, Butter	3
Cornflakes, Jam	1
Butter, Jam	1

Itemset	Sup
Milk, Bread	3
Milk, Cornflakes	3
Bread, Cornflakes	2
Bread, Jam	3
Bread, Butter	3

Internet	Sup
Milk, Bread, Cornflakes	2

Therefore, the most frequent and highest itemset data mining sub-itemset is {Milk, Bread, Cornflakes}.

Q.27 Apply apriori algorithm for following set of transactions and find all the association rules with min support = 1 and min confidence = 60 %. ☞ **[SPPU : May-18, (End Sem), Marks 10]**

Tr. ID	Transactions
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

Ans. :

Database D

Tr. ID	Transactions
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

C_1

Itemset	Supp
(1)	2
(2)	3
(3)	3
(4)	1
(5)	3

C_2

Itemset	Supp
(1, 2)	1
(1, 3)	2
(1, 4)	1
(1, 5)	1
(2, 3)	2
(2, 5)	3
(3, 4)	1
(3, 5)	2

C_3

Itemset	Supp
(1, 2, 3)	1
(1, 2, 5)	1
(1, 3, 4)	1
(1, 3, 5)	1
(2, 3, 5)	2

C_4

Itemset	Supp
(1, 2, 3, 5)	1

END...

Unit VI

6 Introduction to Artificial Neural Network

6.1 : Perceptron Learning : Biological Neuron

Q.1 What is perceptron ?

Ans. : • An arrangement of one input layer of McCulloch-Pitts neurons feeding forward to one output layer of McCulloch-Pitts neurons is known as a perceptron.

• The perceptron is a feed - forward network with one output neuron that learns a separating hyper - plane in a pattern space. Fig. Q.1.1 shows perceptron.

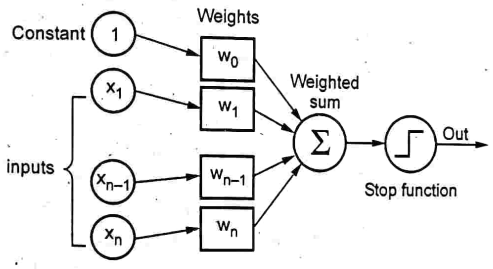


Fig. Q.1.1 Perceptron

• Perceptron is an Artificial Neuron. It is the simplest possible Neural Network. The original Perceptron was designed to take a number of binary inputs, and produce one binary output (0 or 1).

• The idea was to use different weights to represent the importance of each input, and that the sum of the values should be greater than a threshold value before making a decision like true or false (0 or 1).

Q.2 What is biological neuron ? Explain with diagram and its components.

Ans. : • Biological neurons, consisting of a cell body, axons, dendrites and synapses, are able to process and transmit neural activation.

- Artificial neural systems are inspired by biological neural systems. The elementary building block of biological neural systems is the neuron.
- The brain is a collection of about 10 billion interconnected neurons. Each neuron is a cell [right] that uses biochemical reactions to receive, process and transmit information. Fig. Q.2.1 shows biological neural systems.

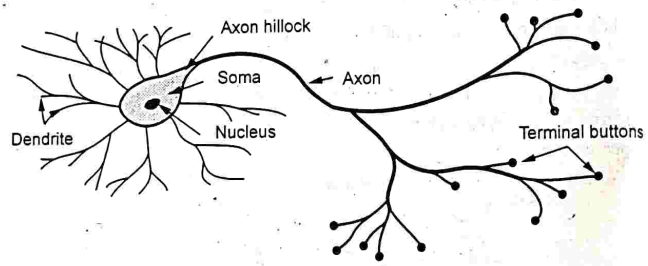


Fig. Q.2.1 Schematic of biological neuron

- The single cell neuron consists of the cell body or soma, the dendrites and the axon. The dendrites receive signals from the axons of other neurons. The small space between the axon of one neuron and the dendrite of another is the synapse. The afferent dendrites conduct impulses toward the soma. The efferent axon conducts impulses away from the soma.

Basic Components of Biological Neurons

1. The majority of ***neurons*** encode their activations or outputs as a series of brief electrical pulses (i.e. spikes or action potentials).
2. The neuron's ***cell body (soma)*** processes the incoming activations and converts them into output activations.
3. The neuron's ***nucleus*** contains the genetic material in the form of DNA. This exists in most types of cells, not just neurons.

4. ***Dendrites*** are fibres which emanate from the cell body and provide the receptive zones that receive activation from other neurons.
5. ***Axons*** are fibres acting as transmission lines that send activation to other neurons.
6. The junctions that allow signal transmission between the axons and dendrites are called ***synapses***. The process of transmission is by diffusion of chemicals called ***neurotransmitters*** across the synaptic cleft.

Q.3 Compare biological NN with artificial NN.

Ans. : Comparison between Biological NN and Artificial NN

Biological NN	Artificial NN
Soma	Unit
Axon, dendrite	Dendrite
Synapse	Weight
Potential	Weighted Sum
Threshold	Bias Weight
Signal	Activation

6.2 : Introduction to ANN

Q.4 What is artificial neural network ? List its characteristics.

Ans. : Artificial Neural Network (ANN) is a computational system inspired by the structure, processing method, learning ability of a biological brain. An artificial neural network is composed of many artificial neurons that are linked together according to specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs.

- ANNs do not execute programmed instructions; they respond in parallel to the pattern of inputs presented to it. There are also no separate memory addresses for storing data. Instead, information is contained in the overall activation 'state' of the network. 'Knowledge' is thus

1. Biological Neuron (Inspire)
2.composed of .. & to specific network architect
3. objective:- transform i/p to meaningful o/p
4. No seprate address to storing the data hence data is stored into activation sate
5. Elements :- processing Unit , topology and learning algorithm
6. Diagram
7. Applications
8.Characteristics

represented by the network itself, which is quite literally more than the sum of its individual components.

- Fig Q.4.1 shows artificial neural network.

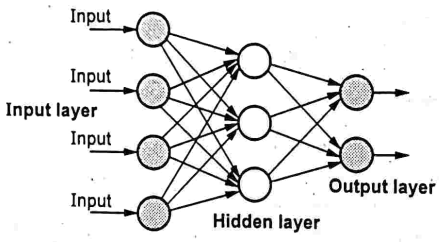


Fig. Q.4.1 Artificial neural network

- Elements of ANN are processing units, topology and learning algorithm.
- Tasks to be solved by artificial neural networks :
  1. Controlling the movements of a robot based on self-perception and other information;
  2. Deciding the category of potential food items in an artificial world;
  3. Recognizing a visual object;
  4. Predicting where a moving object goes, when a robot wants to catch it.
- Characteristics of artificial neural networks
  1. Large number of very simple processing neuron-like processing elements.
  2. Large number of weighted connections between the elements.
  3. Distributed representation of knowledge over the connections.
  4. Knowledge is acquired by network through a learning process.

Q.5 Explain advantages and application of neural network.

Ans. : Advantages of neural network

The advantages of neural networks are due to its adaptive and generalization ability.

a) Neural networks are adaptive methods that can learn without any prior assumption of the underlying data.

b) Neural network, namely the feed forward multilayer perception and radial basis function network have been proven to be universal functional approximations.

c) Neural networks are non-linear model with good generalization ability.

Applications of Neural Network

Neural network applications can be grouped in following categories :

1. **Clustering :** A clustering algorithm explores the similarity between patterns and places similar patterns in a cluster. Best known applications include data compression and data mining.
2. **Classification/Pattern recognition :** The task of pattern recognition is to assign an input pattern (like handwritten symbol) to one of many classes. This category includes algorithmic implementations such as associative memory.
3. **Function approximation :** The tasks of function approximation is to find an estimate of the unknown function f() subject to noise. Various engineering and scientific disciplines require function approximation.
4. **Prediction/Dynamical systems :** The task is to forecast some future values of a time-sequenced data. Prediciton has a significant impact on decision support systems. Prediction differs from function approximation by considering time factor.

Q.6 What are the characteristics of ANN ?

Ans. : Characteristics of Artificial Neural Networks :

1. Large number of very simple processing neuron-like processing elements.
2. Large number of weighted connections between the elements.
3. Distributed representation of knowledge over the connections.
4. Knowledge is acquired by network through a learning process.

Q.7 Explain with example, the challenges in assigning synaptic weights for the interconnection between neurons ? How can this challenge be addressed ?

Ans. : • Artificial neural networks are specifically designed for a particular function like binary classification, multi class classification,

pattern recognition and so on through learning processes. The weights of the synaptic connections of both the neural networks adjust with the learning process.

- Each connection between two neurons has a unique synapse with a unique weight attached to it. When we talk about updating weights in a network, we're really talking about adjusting the weights on these synapses.
- Weights are values that control the strength of the connection between two neurons. That is, inputs are typically multiplied by weights, and that defines how much influence the input will have on the output.
- In other words, when the inputs are transmitted between neurons, the weights are applied to the inputs along with an additional value (the bias).
- The connection is controlled by the strength or amplitude of a connection between both nodes, also called the synaptic weight. Multiple synapses can connect the same neurons, with each synapse having a different level of influence (trigger) on whether that neuron is "fired" and activates the next neuron.
- the synapse is represented as a weight vector. The weights control how the output of the previous layer is fed through the activation function for each node. Using the convenient scalability of linear algebra, all the weight vectors of the synapses for an entire node layer can be represented as a single weight matrix.
- The resulting output of all these nodes is used as the input for the next layer of nodes and so on throughout the neural network "brain" until the final output layer is reached.
- The synapse is defined by these weights and expressed algebraically (for a linear synapse) as :

$$y_i = \sum w_{ij} x_i \quad \text{or} \quad y = wx$$

- Typically a bias term will be added and the output values will be passed through a choice of activation function before being used as inputs into the subsequent layer

6.3 : McCulloch Pitts Neuron

Q.8 Explain McCulloch Pitts neuron with diagram.

Ans. : • The first mathematical model of a biological neuron was presented by McCulloch and Pitts. This model is known as McCulloch Pitt model. It is basic building block of neural network.

- Directed weight graph is used for connecting neurons.
- McCulloch and Pitts describe a neuron as a logical threshold element with two possible states. Such a threshold element has "N" input channels and one output channel. An input channel is either active (input 1) or silent (input 0).
- The activity states of all input channels thus encode the input information as a binary sequence of N bits. The state of the threshold element is then given by linear summation of all a different input signals x_i and comparison of the sum with a threshold value s.
- The system of neurons is static and acts synchronously. A processor (system) with multiple inputs and a single output.
- Effective input : Weighted sum of all inputs.
- Bias or threshold : If the effective input is larger than the bias, the neuron outputs a one, otherwise, it outputs a zero.
- Fig. Q.8.1 shows McCulloch Pitt model.

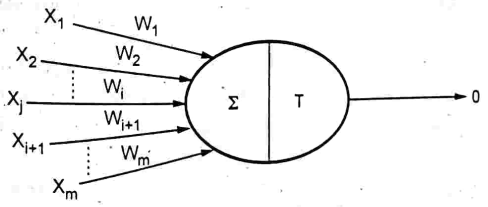


Fig. Q.8.1

- This model can be described in a mathematical formalism as follows :

$$0 = \theta(a)$$

Where, $a = \sum W_j X_j - T$

And

$\theta(x)$ is a function such that $\theta(x) = 1$ if $x > 0$, otherwise $\theta(x) = 1$.

- The parameters used to scale the inputs are called the weights. The effective input is the weighted sum of the inputs. The parameter to measure the switching level is the threshold or bias. Neuron fires (output of one) when its net input excitation exceeds a certain value called 'threshold.' Threshold is the minimum value of the sum of the weighted active inputs needed for the postsynaptic neuron to fire.
- The function for producing the final output is called the activation function, which is the step function in the McCulloch-Pitts model.

$$0 = f\left(\sum_{j=1}^{N} W_j X_j - T\right)$$

$$f(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- Their "neurons" operated under the following assumptions :
  1. They are binary devices (0,1) .
  2. Each neuron has a fixed threshold (theta).
  3. The neuron receives inputs from excitatory synapses, all having identical weights.
  4. Inhibitory inputs have an absolute veto power over any excitatory inputs.
  5. At each time step the neurons are simultaneously (synchronously) updated by summing the weighted excitatory inputs and setting the output to 1 iff the sum is greater than or equal to the threshold AND if the neuron receives no inhibitory input.

6.4 : Perceptron and its Learning Algorithm

Q.9 Briefly discuss Adaline network.

Ans. : ADALINE (Adaptive Linear Neuron) is an early single-layer artificial neural network.

- An important generalized of the perceptrons training algorithm was presented by Widrow and Hoff. as the least mean square learning procedure also known as the delta rule.
- The learning rule was applied to the **"adaptive linear element"** also named Adaline.
- The perceptron learning rule uses the output of the thersold function for learning. The delta rule uses the net output without further mapping into output values –1 or + 1.
- Fig. Q.9.1 shows adaline.

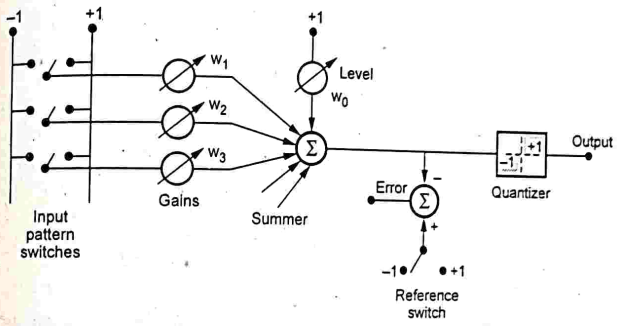


Fig. Q.9.1 Adaline

- If the input conductances are denoted by w_i where m i = 0, 1, 2, ..., n and input and output signals by x_i and y respectively, then the output of the central block is defined to be :

$$y = \sum_{i=1}^{n} w_i x_i + \theta$$

Where, $\theta \equiv w_0$

- In a simple physical implementation, this device consists of a set of controllable resistors connected to a circuit which can sum up currents caused by the input voltage signals. Usually the central block, the summer is also followed by a quantizer which outputs + 1 of – 1, depending on the polarity of the sum.

- The problem is to determine the coefficients w_i, where $i = 0, 1 \ldots, n$, in such way that the input output response is correct for a large number of arbitrarily chosen signal sets.
- If an exact mapping is not possible the average error must be minimized, for instance, in the sense of least squares.
- An adaptive operation means that there exists a mechanism by which the w_i can be adjusted, usually iteratively to attain the correct values.
- For the Adaline, Widrow introduced the delta rule to adjust the weights.
- For the p^{th} input-output pattern, the earor measure of a single-output Adaline can be expressed as,

$$E_p = (t_p - o_p)^2$$

Where

t_p = Target output

o_p = Actual output of the Adaline

- The derivation of E_p with respect to each weight w_i is

$$\frac{\partial E_p}{\partial w_i} = -2(t_p - o_p)\, x_i$$

- To decrease E_p by gradient descent, the update formula for w_i on the p^{th} input-output pattern is

$$\Delta_p w_i = \eta(t_p - o_p)x_i$$

- The delta rule tries to minimize squared errors, it is also referred to as the least mean square learning procedure or Widrow - Hoff learning rule.

Q.10 What is activation function ? Also explain logistic function and Arc tangent.

Ans. : • Activation functions also known as transfer function is used to map input nodes to output nodes in certain fashion.

- The activation function is the most important factor in a neural network which decided whether or not a neuron will be activated or not and transferred to the next layer.

- Activation functions help in normalizing the output between 0 to 1 or -1 to 1. It helps in the process of backpropagation due to their differentiable property. During backpropagation, loss function gets updated, and activation function helps the gradient descent curves to achieve their local minima.
- Activation function basically decides in any neural network that given input or receiving information is relevant or it is irrelevant.
- These activation function makes the multilayer network to have greater representational power than single layer network only when non-linearity is introduced.
- The input to the activation function is sum which is defined by the following equation.

$$\text{sum} = I_1 W_1 + I_2 W_2 + \ldots + I_n W_n = \sum_{j=1}^{n} I_j W_j + b$$

- **Activation Function : Logistic Function**

$$f(\text{sum}) = \frac{1}{(1 + e^{-s * \text{sum}})} = (1 + s^{-s \times \text{sum}})^{-1}$$

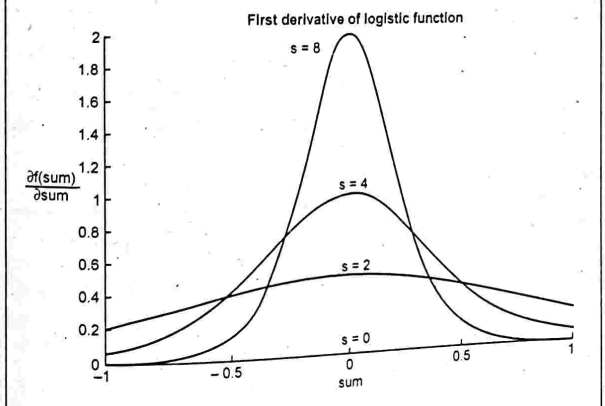


Fig. Q.10.1

- Logistic function monotonically increases from a lower limit (0 or – 1) to an upper limit (+1) as sum increases. In which values vary between 0 and 1, with a value of 0.5 when I is zero.
- **Activation Function : Arc Tangent**

$$f(sum) = \frac{2}{\pi} \tan^{-1} (s \times sum)$$

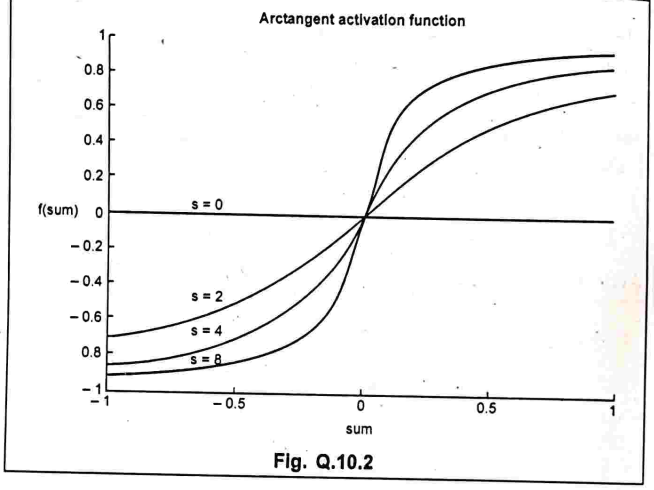


Fig. Q.10.2

Q.11 What are advantages and disadvantages of Sigmoid, Tanh and ReLU.

Ans. :

Function	Advantages	Disadvantages
Sigmoid	1. Output in range (0,1)	1. Saturated Neurons 2. Not zero entered 3. Small gradient 4. Vanishing gradient
Tanh	1. Zero centered, 2. Output in range (– 1,1)	1. Saturated Neurons
ReLU	1. Computational efficiency, 2. Accelerated convergence	1. Dead Neurons, 2. Not zero centered

Q.12 What is the function of a summation junction of a neuron? What is threshold activation function ?

Ans. : • Summation junction for the input signals is weighted by the respective synaptic weight. Because it is a linear combiner or adder of the weighted input signals, the output of the summation junction (y) can be expressed as follows:

$$y_{sum} = \sum_{i=1}^{n} w_i \; x_i$$

- This summed function is applied over an Activation function. The output from this neuron is multiplied with the weight W3 and supplied as input to the output layer.
- Fig. Q.12.1 shows summing function and activation function.

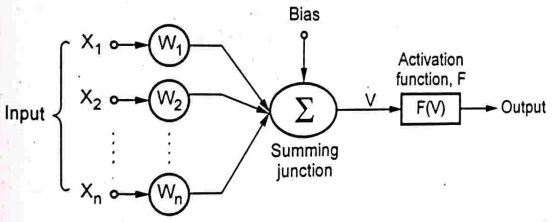


Fig. Q.12.1

- Threshold activation function also known as activation function, it results in an output signal only when an input signal exceeding a specific threshold value comes as an input. It is similar in behaviour to the biological neuron which transmits the signal only when the total input signal meets the firing threshold.

- Activation functions help in normalizing the output between 0 to 1 or – 1 to 1. It helps in the process of backpropagation due to their differentiable property. During backpropagation, loss function gets updated, and activation function helps the gradient descent curves to achieve their local minima.
- Activation function basically decides in any neural network that given input or receiving information is relevant or it is irrelevant.
- These activation function makes the multilayer network to have greater representational power than single layer network only when non-linearity is introduced.

6.5 : Multi-Layer Perceptron Model

Q.13 Briefly discuss multilayer perceptron NN.

Ans. : • Multilayer perceptron is a neural network that learns the relationship between linear and non-linear data. Multilayer Perceptron has input and output layers, and one or more hidden layers with many neurons stacked together.

- Multilayer perceptron falls under the category of feed-forward algorithms, because inputs are combined with the initial weights in a weighted sum and subjected to the activation function.

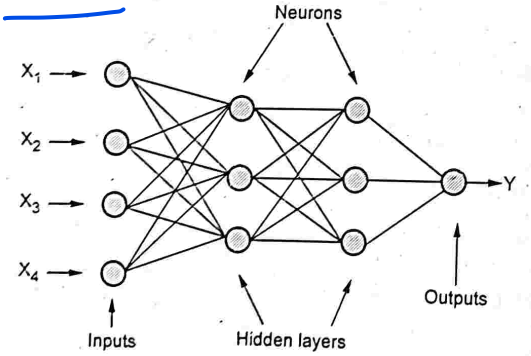


Fig. Q.13.1 Multilayer perceptron NN

- Fig. Q.13.1 shows Multilayer perceptron NN.
- Each layer is feeding the next one with the result of their computation, their internal representation of the data. This goes all the way through the hidden layers to the output layer.
- Multilayer perceptron is a typical example of a feedforward artificial neural network.
- The number of layers and the number of neurons are referred to as hyper-parameters of a neural network, and these need tuning. Cross-validation techniques must be used to find ideal values for these.
- The weight adjustment training is done via back-propagation. Deeper neural networks are better at processing data. However, deeper layers can lead to vanishing gradient problems. Special algorithms are required to solve this issue.
- The Multilayer Perceptron learning procedure is as follows:
  a) Starting with the input layer, propagate data forward to the output layer. This step is the forward propagation.
  b) Based on the output, calculate the error. The error needs to be minimized.
  c) Backpropagate the error. Find its derivative with respect to each weight in the network, and update the model.
  d) Repeat the three steps given above over multiple epochs to learn ideal weights.
  e) Finally, the output is taken via a threshold function to obtain the predicted class labels.

Q.14 Explain back-propagation neural networks.

Ans. : • Backpropagation is a training method used for a multi-layer neural network. It is also called the generalized delta rule. It is a gradient descent method which minimizes the total squared error of the output computed by the net.

- The backpropagation algorithm looks for the minimum value of the error function in weight space using a technique called the delta rule or

gradient descent. The weights that minimize the error function is then considered to be a solution to the learning problem.

- Backpropagation is a systematic method for training multiple layer ANN. It is a generalization of Widrow-Hoff error correction rule. 80 % of ANN applications uses backpropagation.
- Fig. Q.14.1 shows backpropagation network.

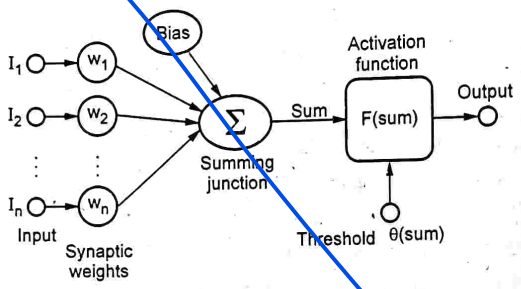


Fig. Q.14.1 Backpropagation network

- Consider a simple neuron :
  a. Neuron has a summing junction and activation function.
  b. Any non linear function which differentiable everywhere and increases everywhere with sum can be used as activation function.
  c. Examples : Logistic function, Arc tangent function, Hyperbolic tangent activation function.
- These activation function makes the multilayer network to have greater representational power than single layer network only when non-linearity is introduced.
- **Weights** connects unit (neuron) in one layer only to those in the next higher layer. The output of the unit is scaled by the value of the connecting weight and it is fed forward to provide a portion of the activation for the units in the next higher layer.
- Backpropagation can be applied to an artificial neural network with any number of hidden layers. The training objective is to adjust the weights so that the application of a set of inputs produces the desired outputs.

Q.15 Which factors are influencing backpropagation training ?

Ans. : Factors influencing backpropagation training :

- The training time can be reduced by using :

1. **Bias :** Networks with biases can represent relationships between inputs and outputs more easily than networks without biases. Adding a bias to each neuron is usually desirable to offset the origin of the activation function. The weight of the bias is trainable similar to weight except that the input is always +1.
2. **Momentum :** The use of momentum enhances the stability of the training process. Momentum is used to keep the training process going in the same general. In backpropagation with momentum, the weight change is a combination of the current gradient and the previous gradient.

Q.16 Explain advantages and disadvantages of backpropagation.

Ans. : Advantages of backpropagation

1. It is simple, fast and easy to program.
2. Only numbers of the input are tuned and not any other parameter.
3. No need to have prior knowledge about the network.
4. It is flexible.
5. A standard approach and works efficiently.
6. It does not require the user to learn special functions.

Disadvantages of backpropagation

1. Backpropagation possibly be sensitive to noisy data and irregularity.
2. The performance of this is highly reliant on the input data.
3. Needs excessive time for training.
4. The need for a matrix-based method for backpropagation instead of mini-batch.

Q.17 What is need of hidden layer in backpropagation ? How hidden units are selected ?

Ans. : • Need of hidden layers

1. A network with only two layers (input and output) can only represent the input with whatever representation already exists in the input data.

2. If the data is discontinuous or non-linearly separable, the innate representation is inconsistent and the mapping cannot be learned using two layers (Input and Output).
3. Therefore, hidden layers(s) are used between input and output layers.

- **Selection of number of hidden units :** The number of hidden units depends on the number of input units.
  1. Never choose h to be more than twice the number of input units.
  2. You can load p patterns of 1 elements into $\log_2 p$ hidden units.
  3. Ensure that we must have at least 1/e times as many training examples.
  4. Features extraction requires fewer hidden units than inputs.
  5. Learning many examples of disjointed inputs requires more hidden units than inputs.
  6. The number of hidden units required for a classification task increases with the number of classes in the task. Large network require longer training times.

Q.18 Write short note on recurrent neural network.

Ans. : • A recurrent neural network is a type of neural network that contains loops, allowing information to be stored within the network.

- A RNN is particularly useful when a sequence of data is being processed to make a classification decision or regression estimate but it can also be used on non-sequential data. Recurrent neural networks are typically used to solve tasks related to time series data.
- Applications of recurrent neural networks include natural language processing, speech recognition, machine translation, character-level language modeling, image classification, image captioning, stock prediction, and financial engineering.
- Fig. Q.18.1 shows architecture of recurrent neural network.
- Recurrent Neural Networks can be thought of as a series of networks linked together. They often have a chain-like architecture, making them applicable for tasks such as speech recognition, language translation, etc.

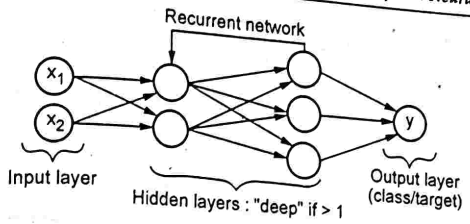


Fig. Q.18.1

- An RNN can be designed to operate across sequences of vectors in the input, output, or both. For example, a sequenced input may take a sentence as an input and output a positive or negative sentiment value. Alternatively, a sequenced output may take an image as an input, and produce a sentence as an output.

6.6 : Learning Parameter : Bias and Weight

Q.19 Discuss learning parameter weight and bias and compare bias with weight.

Ans. : • Weights and biases are the learnable parameters that help a neural network correctly learn a function. A machine learning model uses lots of examples to learn the correct weights and bias to assign to each feature in a dataset to help it correctly predict outputs.

- Weight is the parameter within a neural network that transforms input data within the network's hidden layers. A neural network is a series of nodes, or neurons. Within each node is a set of inputs, weight, and a bias value.
- As an input enters the node, it gets multiplied by a weight value and the resulting output is either observed, or passed to the next layer in the neural network. Often the weights of a neural network are contained within the hidden layers of the network.
- Choose the weights w_i to best fit the set of training examples.
- Minimize the squared error E between the train values and the values predicted by the hypothesis

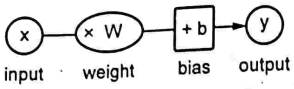


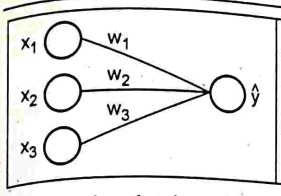
Fig. Q.19.1

$$E = \sum_{(b,\ Vtrain(b)) \in \text{training examples}} (V_{train}(b) - \hat{V}((b))^2$$

- Require an algorithm that will incrementally refine weights as new training values.
- Least Mean Squares (LMS) is one such algorithm.

Weight vs. Bias

- A teachable neural network will randomize both the weight and bias values before learning initially begins. As training continues, both parameters are adjusted toward the desired values and the correct output.
- The two parameters differ in the extent of their influence upon the input data. Simply, bias represents how far off the predictions are from their intended value.
- Biases make up the difference between the function's output and its intended output. A low bias suggest that the network is making more assumptions about the form of the output, whereas a high bias value makes less assumptions about the form of the output.
- Weights, on the other hand, can be thought of as the strength of the connection. Weight affects the amount of influence a change in the input will have upon the output. A low weight value will have no change on the input, and alternatively a larger weight value will more significantly change the output.
- In most learning networks, error is calculated as the difference between the actual output y and the predicted output $\hat{y}$. Loss functions are quantitative measures of how satisfactory the model predictions are.



Output :

$\hat{y} = w_1x_1 + w_2x_2 + w_3x_3 + b = w^T x + b$

Neural network parameters

$W = \{[w_1, w_2, w_3]b\}$

The function that is used to compute this error is known as Loss Function also known as Cost function.

6.7 : Introduction to Deep Learning

Q.20 Explain multi task learning.

☞ [SPPU : May-16, 17, (End Sem), Marks 8]

OR Explain multi-task learning with task grouping and overlap methodology. ☞ [SPPU : May-18, (End Sem), Marks 8]

Ans. : • Multi-task learning refers to learning multiple tasks simultaneously, in order to avoid tabula rasa learning and to share information between similar tasks during learning.

- The transfer of information between tasks is usually eased by learning the tasks with similar machine learning models or algorithms under common input and output representations. The aim may be to increase the performance on all tasks, or it may be to increase the performance on one task of primary interest given other tasks of secondary interests.

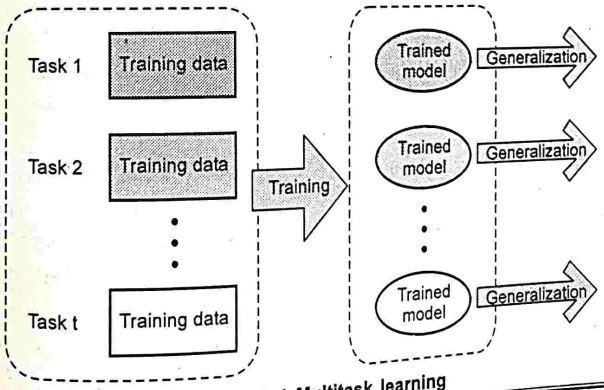


Fig. Q.20.1 Multitask learning

- Fig. Q.20.1 shows multitask learning.
- To differentiate between these two cases of multi-task learning, call them symmetric and asymmetric respectively. In the machine learning literature, it is more common to find symmetric multi-task learning. In contrast, asymmetric multi-task learning is more prevalent in the geo-statistics community, where the secondary tasks pertain to information more readily available than the primary task.
- For example, the primary task may be to predict the concentration of a precious metal, while the secondary tasks provide the concentrations of common metals.
- Example : Web pages categorization
  1. Classify documents into categories
  2. The classification of each category is a task
  3. The tasks of predicting different categories may be latently related.

Q.21 Explain deep learning. What are the challenges in deep learning ? ☞ **[SPPU : May-18, (End Sem), Marks 7]**

OR Explain deep learning. ☞ **[SPPU : May-16, (End Sem), Marks 8]**

OR Write a note on deep learning ☞ **[SPPU : Dec.-16, (End Sem), Marks 8]**

Ans. : • Deep Learning is a new area of machine learning research, which has been introduced with the objective of moving machine learning closer to one of its original goals.

- Deep learning is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text.
- 'Deep learning' means using a neural network with several layers of nodes between input and output. It is generally better than other methods on image, speech and certain other types of data because the series of layers between input and output do feature identification and processing in a series of stages, just as our brains seem to.

- Deep learning emphasizes the network architecture of today's most successful machine learning approaches. These methods are based on "deep" multi-layer neural networks with many hidden layers.

Challenges in Deep learning :

- They need to find and process massive datasets for training
- One of the reasons deep learning works so well is the large number of interconnected neurons, or free parameters, that allow for capturing subtle nuances and variations in data
- Due to the sheer number of layers, nodes, and connections, it is difficult to understand how deep learning networks arrive at insights
- Deep-learning networks are highly susceptible to the butterfly effect-small variations in the input data can lead to drastically different results, making them inherently unstable

Q.22 Write a note on deep learning and its applications. ☞ **[SPPU : Dec.-17, (End Sem), Marks 7]**

Ans. : Refer Q.21 of Chapter 6.

Application :

1. Colorization of black and white images.
2. Adding sounds to silent movies.
3. Automatic machine translation.
4. Object classification in photographs.
5. Automatic handwriting generation.
6. Character text generation.
7. Image caption generation.
8. Automatic game playing.

END...

SOLVED MODEL QUESTION PAPER (In Sem)
Machine Learning
T.E. (IT) Semester - V [As Per 2019 Pattern]

Time : 1 Hour] **[Maximum Marks : 30**

N.B. : i) Attempt Q.1 or Q.2, Q.3 or Q.4.

ii) Neat diagrams must be drawn wherever necessary.

iii) Figures to the right side indicate full marks.

iv) Assume suitable data, if necessary.

Q.1 a) *What is machine learning ? Explain types of machine learning.* **(Refer Q.1 of Chapter - 1)** **[5]**

b) *Differentiate between supervised and unsupervised learning.* **(Refer Q.9 of Chapter - 1)** **[4]**

c) *What is the need if dimensionality reduction ? Explain subset selection.* **(Refer Q.22 of Chapter - 1)** **[6]**

OR

Q.2 a) *Explain predictive and descriptive task.* **(Refer Q.6 of Chapter - 1)** **[5]**

b) *Explain qualitative data types.* **(Refer Q.12 of Chapter - 1)** **[3]**

c) *What is feature selection ? What is role of feature selection in ML ? Explain feature selection algorithm.* **(Refer Q.19 of Chapter - 1)[7]**

Q.3 a) *Explain any 5 binary classification performance evaluation parameters (excluding accuracy and error rate).* **(Refer Q.3 of Chapter - 2)** **[5]**

b) *Write short note on F-Measure.* **(Refer Q.12 of Chapter - 2)[4]**

c) *Explain kernel methods for non-linearity.* **(Refer Q.20 of Chapter - 2)** **[6]**

OR

Q.4 a) *State formulae for calculating accuracy, true positive rate, true negative rate, false positive rate and false negative rate for binary classification tasks.* **(Refer Q.4 of Chapter - 2)** **[7]**

b) *Explain various multiclass classification techniques.* **(Refer Q.14 of Chapter - 2)** **[8]**

SOLVED MODEL QUESTION PAPER (End Sem)
Machine Learning
T.E. (IT) Semester - V [As Per 2019 Pattern]

Time : $2\frac{1}{2}$ Hours] **[Maximum Marks : 70**

N.B. : i) Attempt Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.

ii) Neat diagrams must be drawn wherever necessary.

iii) Figures to the right side indicate full marks.

iv) Assume suitable data, if necessary.

Q.1 a) *Consider following data for 5 students.*

Each Xi (i = 1 to 5) represents the score of i^{th} student in standard X and corresponding Yi (i = 1 to 5) represents the score of i^{th} student in standard XII.

i) What linear regression equation best predicts standard XII^{th} score ?

ii) Find regression line that fits best for given sample data.

iii) How to interpret regression equation ?

iv) If a student's score is 80 in std X, then what is his expected score in XII standard ? **(Refer Q.6 of Chapter - 3)**

Student	Score in X standard (Xi)	Score in XII standard (Yi)
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

[8]

b) *When is it suitable to use linear regression over classification ?* **(Refer Q.3 of Chapter - 3)** [5]

c) *Explain gradient descent algorithm. Also explain its limitation.* **(Refer Q.14 of Chapter - 3)** [5]

OR

Q.2 a) *Consider following data :*

i) Find values of β_0 *and* β_1 *w.r.t. linear regression model which best fits given data.*

ii) Interpret and explain equation of regression line.

iii) If new person rates "Bahubali -Part I" as 3 then predict the rating of same person for "Bahubali -Part II". **(Refer Q.7 of Chapter - 3)**

Person	X_i = Rating for movie "Bahubali Part - I" by i^{th} person	Y_i = Rating for movie "Bahubali Part - II" by i^{th} person
1^{st}	4	3
2^{nd}	2	4
3^{rd}	3	2
4^{th}	5	5
5^{th}	1	3
6^{th}	3	1

[8]

b) *Define and explain Squared Error (SE) and Mean Squared Error (MSE) w.r.t Regression.* **(Refer Q.8 of Chapter - 3)** [5]

c) *What is a polynomial regression ? How it can be represented in a form of a matrix.* **(Refer Q.21 of Chapter - 3)** [5]

Q.3 a) *Write the Grow Tree algorithm to generate feature tree ? Explain the role of best split in this algorithm.* **(Refer Q.7 of Chapter - 4)** [10]

b) *Write and explain Naïve Bayes classification algorithm.* **(Refer Q.13 of Chapter - 4)** [7]

OR

Q.4 a) *What is a probabilistic model ? Give an example of it.* **(Refer Q.10 of Chapter - 4)** [5]

b) *Write short on regression tree with its algorithm.* **(Refer Q.3 of Chapter - 4)** [5]

c) *What is decision tree ? Explain.* **(Refer Q.2 of Chapter - 4)** [7]

Q.5 a) *Explain kNN algorithm with its advantages and disadvantages.* **(Refer Q.5 of Chapter - 5)** [10]

b) *Write a note on clustering trees.* **(Refer Q.9 of Chapter - 5)** [8]

OR

Q.6 a) *Discuss various distance measures.* **(Refer Q.1 of Chapter - 5)** [10]

b) *What is Apriori algorithm ? Explain use and limitation of Apriori algorithm.* **(Refer Q.24 of Chapter - 5)** [8]

Q.7 a) *What is the function of a summation junction of a neuron ? What is threshold activation function ?* **(Refer Q.12 of Chapter - 6)** **[5]**

b) *Which factors are influencing backpropagation training ?*
(Refer Q.15 of Chapter - 6) **[4]**

c) *Explain McCulloch Pitts neuron with diagram.*
(Refer Q.8 of Chapter - 6) **[8]**

OR

Q.8 a) *Write short note on recurrent neural network.*
(Refer Q.18 of Chapter - 6) **[4]**

b) *Briefly discuss multilayer perceptron NN.*
(Refer Q.13 of Chapter - 6) **[5]**

c) *What is artificial neural network ? List its characteristics.*
(Refer Q.4 of Chapter - 6) **[8]**

END...

Notes

DECODE®

A Guide For Engineering Students

MACHINE LEARNING

(For END SEM Exam - 70 Marks)

SUBJECT CODE : 314443

T.E. (Information Technology) Semester - V

Published by :

TECHNICAL PUBLICATIONS
SINCE 1993 An Up-Thrust for Knowledge

Amit Residency, Office No.1, 412, Shaniwar Peth,
Pune - 411030, M.S. INDIA Ph.: +91-020-24495496/97
Email : info@technicalpublications.in
Website : www.technicalpublications.in

Printer :
Yogiraj Printers & Binders, Sr.No. 10/1A, Ghule Industrial Estate, Nanded Village Road,
Tal. - Haveli, Dist. - Pune - 411041.

ISBN 978-93-5585-114-7

SPPU 19

JUNE - 2022 [END SEM] [5870] - 1142 **Solved Paper**

Course 2019

Time : 2 $\frac{1}{2}$ Hours] **[Maximum Marks : 70**

Instructions to the candidates :

1) Answers Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.

2) Neat diagrams must be drawn wherever necessary.

3) Figures to the right side indicate full marks.

4) Assume suitable data if necessary.

Q.1 a) *Why do we need optimization in regression ? Explain gradient descent optimization technique in detail.* **[9]**

Ans. : • Optimization is the process where we train the model iteratively that results in a maximum and minimum function evaluation. It is one of the most important phenomena in machine learning to get better results.

- We compare the results in every iteration by changing the hyperparameters in each step until we reach the optimum results. We create an accurate model with less error rate.
- There are different ways using which we can optimize a model. Two important optimization algorithms : Gradient descent and Stochastic gradient descent algorithms.
- Gradient descent is an iterative optimization algorithm to find the minimum of a function. Also Refer Q.14 of Chapter - 3.

b) *What do you meant by least square method ? Explain least square method in the context of linear regression.*
(Refer Q.5 of Chapter - 3) **[8]**

OR

Q.2 a) *What is overfitting and underfitting ? Explain the reasons of overfitting and underfitting.* **(Refer Q.22 of Chapter - 3)** **[9]**

b) *What is multivariate and univariate regression state the difference with examples.* **(Refer Q.2 and Q.16 of Chapter - 3)** **[8]**

Ans. :

Sr. No.	Multivariate regression	Univariate regression
1.	For analysis, it uses two or more variable	For analysis, it uses only one variable.
2.	It uses Principal Component Analysis or logistic regression, linear regression, cluster analysis etc.	The most common univariate analysis is checking the central tendency, the range, the maximum and minimum values, and standard deviation of a variable.
3.	In terms of visualization, scatter plots and histograms are used.	Visualizations such as histogram, distribution, frequency tables, bar chart and box plots are commonly used.

Q.3 a) *Explain ID-3 decision tree algorithm in detail.* **[9]**

Ans. :

- ID3 stands for Iterative Dichotomiser 3. This algorithm used to generate a decision tree. ID3 uses Entropy function and Information gain as metrics.
- The ID3 follows the Occam's razor principle. It attempts to create the smallest possible decision tree.
- The calculation for information gain is the most difficult part of this algorithm.
- ID3 performs a search whereby the search states are decision trees and the operator involves adding a node to an existing tree. It uses information gain to measure the attribute to put in each node, and performs a greedy search using this measure of worth.
- The algorithm goes as follows : Given a set of examples (S), categorised in categories c_i, then :

1. Choose the root node to be the attribute, A, which scores the highest for information gain relative to S.
2. For each value v that A can possibly take, draw a branch from the node.

3. For each branch from A corresponding to value v, calculate S_v. Then :

i. If S_v is empty, choose the category c default which contains the most examples from S, and put this as the leaf node category which ends that branch.

ii. If S_v contains only examples from a category c, then put c as the leaf node category which ends that branch.

iii. Otherwise, remove A from the set of attributes which can be put into nodes. Then put a new node in the decision tree, where the new attribute being tested in the node is the one which scores highest for information gain relative to S_v.

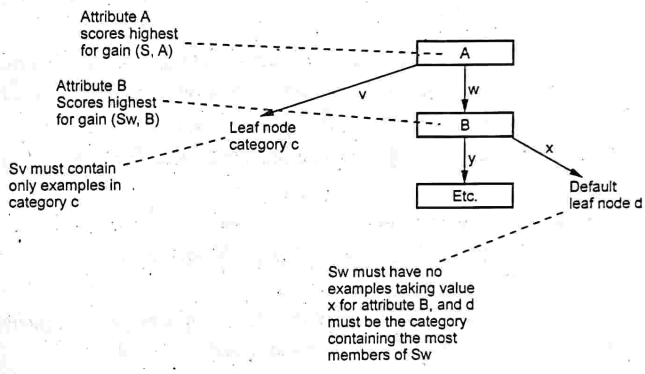


Fig. 1

- The ID3 algorithm is a classic data mining algorithm for classifying instances. The input is a set of training data for building a decision tree.
- By applying the ID3 algorithm, a decision tree is created. To create the decision tree, we have to choose a target attribute.
- Take all unused attributes and calculates their entropies. Chooses attribute that has the lowest entropy is minimum or when information gain is maximum. Makes a node containing that attribute.

Capabilities and Limitations of ID3

1. Hypothesis space is a complete space of all discrete valued functions.
2. Cannot determine how many alternative trees are consistent with training data.
3. ID3 in its pure form performs no backtracking.
4. ID3 uses all training examples at each step to make statistically based decisions regarding how to refine its current hypothesis.

b) *Explain the measures of impurity (information gain, gini index, entropy).* **[6]**

Ans. :

- Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
- Gini index, entropy and twoing rule are some of the frequently used impurity measures.
- Gini Index for a given node t :

$$GINI(t) = \sum_j p(j \mid t)(1 - p(j \mid t)) = \sum_j p(j \mid t)^2$$

- Maximum of $1-1/n_c$ (number of classes) when records are equally distributed among all classes = maximal impurity.
- Entropy measures the impurity of a collection. Information gain is defined in terms of entropy.
- Information gain tells us how important a given attribute of the feature vectors is.
- Information gain of attribute A is the reduction in entropy caused by partitioning the set of examples S.

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in values} \frac{|S_v|}{|S|} Entropy(S_v)$$

where values (A) is the set of all possible values for attributes A and S_v is the subset of S for which attribute A has value v.

c) *What are the advantages and limitation of tree-based model.* **(Refer Q.4 of Chapter - 4)** **[2]**

OR

Q.4 a) *Explain in brief the Bayesian network for learning and inference.* **[9]**

Ans. :

- Bayesian belief networks represent the full joint distribution over the variables more compactly with a smaller number of parameters.
- It take advantage of conditional and marginal independences among random variables.
- A and B are independent then P(A, B) = P(A)P(B).
- A and B are conditionally independent given C.

  P(A, B | C) = P(A | C)P(B | C)

  P(A | C, B) = P(A | C)
- **Example : Alarm system example.**
- Assume your house has an alarm system against burglary. You live in the seismically active area and the alarm system can get occasionally set off by an earthquake.
- You have two neighbors, Mary and John, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.
- We want to represent the probability distribution of events : Burglary, Earthquake, Alarm, Mary calls and John calls

Causal relations :

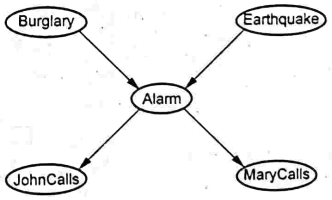


Fig. 2

Directed acyclic graph :

- Nodes = Random variables
  Burglary, Earthquake, Alarm, Mary calls and John calls
- Links = Direct (causal) dependencies between variables.

The chance of Alarm is influenced by Earthquake, The chance of John calling is affected by the Alarm.

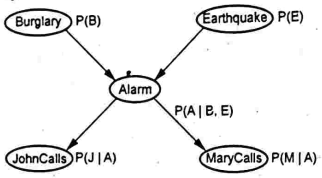


Fig. 3

Local conditional distributions : Relate variables and their parents

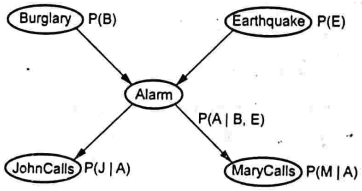


Fig. 4

Bayesian belief network :

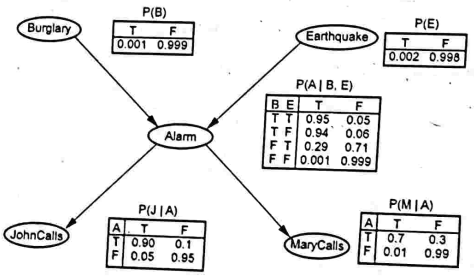


Fig. 5

b) *Explain Naïve Bayes algorithm with example.*
(Refer Q.13 of Chapter - 4) **[6]**

c) *Enlist any four applications of Naïve Bayes classifier.* **[2]**

Ans. : Applications of Naïve Bayes classifier :

1. It is used for credit scoring.
2. It is used in medical data classification.
3. It can be used in real-time predictions.
4. It is used in text classification such as spam filtering and sentiment analysis.

Q.5 a) *Following is a dataset for weight of teens and whether they like pizza.* **[6]**

Weight	*Like pizza ?*
78	*YES*
54	*NO*
69	*YES*
73	*YES*
59	*NO*
48	*NO*
82	*NO*
65	*YES*

Using K-Nearest Neighbours (KNN) classification algorithm determine if a teen weighing 63 kgs likely to like pizza ? (Use K = 3)
(Refer similar example of Q.6 of Chapter - 5)

b) *Explain agglomerative hierarchical clustering.*
(Refer Q.17 of Chapter - 5) **[6]**

c) *Explain association rule mining. Explain the various approaches to improve the efficiency of Apriori algorithm.* **[6]**

Ans. :

- Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository.

- An example of an association rule would be "If a customer buys a dozen eggs, he is 80 % likely to also purchase milk."
- It is an important data mining model studied extensively by the database and data mining community. Assume all data are categorical. No good algorithm for numeric data.
- Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.
- Association rule mining can be viewed as a two-step process :
  1. Find all frequent item sets : By definition, each of these item sets will occur at least as frequently as a predetermined minimum support count, min sup.
  2. Generate strong association rules from the frequent item sets : By definition, these rules must satisfy minimum support and minimum confidence.
- An association rule is commonly understood to be an expression of the form : $X \Rightarrow Y$ where X and Y are sets of items such that $X \cap Y = \varnothing$. The association rule $X \Rightarrow Y$ means that transactions containing items from set X tend to contain items from set Y.
- The term association rule was coined by R. Agrawal in the early 90s in relation to a so called market basket analysis. In this analysis, transaction data recorded by point-of-sale systems in supermarkets are analyzed in order to understand the purchase behavior of groups of customers and used to increase sales and for cross-selling, store design, discount plans and promotions.
- This idea of association rules was later generalized to any data in the tabular, attribute-value form. So data describing properties (values of attributes) of some examples can be analyzed in order to find associations between con-junctions of attribute-value pairs (categories) called antecedent (Ant) and succedent (Suc).
- The two basic characteristics of an association rule are support and confidence. Let a be the number of examples that fulfill both Ant and

Suc, b the number of examples that fulfill Ant but do not fulfill Suc, c the number of examples that fulfill Suc but do not fulfill Ant and d the number of examples that fulfill neither Ant nor Suc. Then Support is then defined as,

$$\text{Sup} = P(\text{Ant} \wedge \text{Suc}) = \frac{a}{a+b+c+d}$$

and confidence is defined as,

$$\text{conf} = P(\text{Suc} \mid \text{Ant}) = \frac{a}{a+b}$$

- Various Approaches to improve the efficiency of Apriori algorithm are as follows :
  1. **Hash-Based Technique :** This method uses a hash-based structure called a hash table for generating the K-item sets and its corresponding count. It uses a hash function for generating the table.
  2. **Transaction Reduction :** This method reduces the number of transactions scanning in iterations. The transactions which do not contain frequent items are marked or removed.
  3. **Partitioning :** This method requires only two database scans to mine the frequent item sets. It says that for any item set to be potentially frequent in the database, it should be frequent in at least one of the partitions of the database.
  4. **Sampling :** This method picks a random sample S from Database D and then searches for frequent item set in S. It may be possible to lose a global frequent item set. This can be reduced by lowering the min_sup.
  5. **Dynamic Itemset Counting :** This technique can add new candidate item sets at any marked start point of the database during the scanning of the database.

OR

Q.6 a) *Explain the any two algorithms used for clustering.*
(Refer Q.10 and Q.13 of Chapter - 5) **[6]**

b) *Explain following terms : a) Centroid b) Medoid c) Dendrogram.* **[6]**

Ans. :

a) Cluster centroid : The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Each cluster has a well defined centroid.

b) Mediod : A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set.

c) Dendrogram : Hierarchical clustering arranges items in a hierarchy with a treelike structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a **dendrogram.**

c) *Explain association rule mining. Comment on role of support and confidence in associations of rule mining ?*
(Refer Q.5 (c) of Same question paper and also refer Q.23 of Chapter - 5) **[6]**

Q.7 a) *Explain what is deep learning and its different architectures ? State the various applications of deep learning ?* **[9]**

Ans. : Deep learning architectures are RNN, LSTM, GRU, CNN, DBN and DSN.

- **Recurrent Neural Networks :** RNN is one of the fundamental network architectures from which other deep learning architectures are built. RNNs can use their internal state (memory) to process variable-length sequences of inputs.

- **Long Short-Term Memory :** LSTM has feedback connections. This means that it can process not only single data points (such as images) but also entire sequences of data.
- **Gated Recurrent Unit :** It's a type of LSTM. GRUs are used for smaller and less frequent datasets, where they show better performance.
- **Convolutional Neural Networks :** This architecture is commonly used for image processing, image recognition, video analysis, and NLP. It can take in an input image, assign importance to various aspects/objects in the image, and be able to differentiate one from the others. CNNs consist of an input and an output layer, as well as multiple hidden layers. The CNN's hidden layers typically consist of a series of convolutional layers.
- **Deep Belief Network :** DBN is a multilayer network in which each pair of connected layers is a Restricted Boltzmann Machine. Also Refer Q.21 and Q.22 Chapter - 6.

b) *Explain how the learning parameters are updated for multilayer perceptron ?* **(Refer Q.19 of Chapter - 6)** **[9]**

OR

Q.8 a) *Explain the following activation functions.*
i) Sigmoid ii) Tanh iii) ReLU. **[9]**

Ans. : i) Sigmoid :

- A sigmoid function produces a curve wi an "S" shape. The example sigmoid function shown on the left is a special case of the logistic function, which models the growth of some set.

$$\text{sig}(t) = \frac{1}{1+e^{-t}}$$

- In general, a sigmcid function is real-valued and differentiable, having a non-negative or non-positive first derivative, one local minimum, and one local maximum.

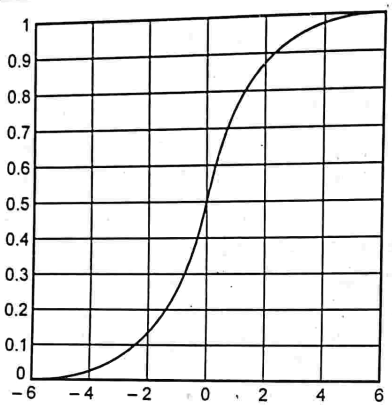


Fig. 6

- The logistic sigmoid function is related to the hyperbolic tangent as follows :

$$1 - 2\,\text{sig}(x) = 1 - 2\frac{1}{1+e^{-x}} = -\tanh\frac{x}{2}$$

- Sigmoid functions are often used in artificial neural networks to introduce nonlinearity in the model.
- A neural network element computes a linear combination of its input signals and applies a sigmoid function to the result.
- A reason for its popularity in neural networks is because the sigmoid function satisfies a property between the derivative and itself such that it is computationally easy to perform.

$$\frac{d}{dt}\text{sig}(t) = \text{sig}(t)(1-\text{sig}(t))$$

- Derivatives of the sigmoid function are usually employed in learning algorithms.

ii) Tanh

- Tanh help to solve non zero centered problem of sigmoid function. Tanh squashes a real-valued number to the range [– 1, 1]. It's non-linear too.
- Tanh is a hyperbolic tangent function.
- In general binary classification problems, the tanh function is used for the hidden layer and the sigmod function is used for the output layer.
- The function maps a real-valued number to the range [– 1, 1] according to the following equation :

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

iii) ReLU

- The ReLU function is actually a function that takes the maximum value. The equation for ReLU (Rectified Linear Unit) is f(x) = max(0, x). It gives an output x if x is positive and 0 otherwise.
- Advantages of the ReLU Activation Function
  a) It is computationally effective as it involves simpler mathematical operations than sigmoid and tanh.
  b) Although it looks like a linear function, it adds non-linearity to the network, making it able to learn complex patterns.
  c) It doesn't suffer from the vanishing gradient problem.
  d) It is unbounded at the positive side. Hence removing the problem of gradient saturation.
  e) It provides sparsity to the network, which as a result lessens the space and time complexity.

b) *What is the difference between biological neuron and artificial neuron ? Explain the simulation of AND gate using McCulloch Pitts Neuron ?* **(Refer Q.3 and Q.8 of Chapter - 6)** **[9]**

DECEMBER - 2022 [END SEM] - [5926] - 121 **Solved Paper**

Course 2019

Time : 2 $\frac{1}{2}$ Hours] **[Maximum Marks : 70**

Q.1 a) *What do you mean by coefficient of regression ? Explain SSE, MSE and MAE in context of regression.* $[CO_2, L_3]$.
(Refer Q.11 of Chapter - 3) **[5]**

b) *What is multiple regression ? How it is different from simple linear regression.* $[CO_2, L_1]$. **[5]**

Ans. :

- Multiple regression is an extension of linear regression models that allow predictions of systems with multiple independent variables. Multiple regression is specifically designed to create regressions on models with a single dependent variable and multiple independent variables.
- Multiple regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term.
- Multiple regression requires two or more predictor variables, and this is why it is called multiple regression.
- Difference between Simple and Multiple Regression

Simple regression	Multiple regression
One dependent variable Y predicted from one independent variable X.	One dependent variable Y predicted from a set of independent variables $(X_1, X_2 \ldots X_k)$.
One regression coefficient.	One regression coefficient for each independent variable.

r^2 : Proportion of variation in dependent variable Y predictable from X.	R^2 : Proportion of variation in dependent variable Y predictable by set of independent variables (X's).

c) *Consider the following data*

The values of x and their corresponding values of y are shown in the table below

i) Find values of β_0 *and* β_1 *w.r.t. linear regression model which best fits given data.*

ii) Interpret and explain equation of regression line.

iii) Estimate the value of y for x = 90. **[7]**

	x	y
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

$[CO_2, L_3]$ **(Refer Q.6 of Chapter - 3)**

OR

Q.2 a) *Explain under fit, over fit and just fit models for regression* $[CO_2, L_1]$. **(Refer Q.22 of Chapter - 3)** **[5]**

b) *Explain bias-variance dilemma.* $[CO_2, L_2]$
(Refer Q.23 of Chapter - 3) **[5]**

c) *What is univariate and multivariate regression ? Explain any three measures of evaluation of performance of regression model.* $[CO_2, L_2]$ **(Refer Q.2(b) of June-2022 and Q.9 of Chapter - 3)** **[7]**

Q.3 a) *For the given data set apply Naïve Bayes classifier and predict the class for weather = Sunny and car = Working.* **[10]**

Whether	Car	Class
Sunny	Working	Go-out
Rainy	Broken	Go-out
Sunny	Working	Go-out
Sunny	Working	Go-out
Sunny	Working	Go-out
Rainy	Broken	Stay-home
Rainy	Broken	Stay-home
Sunny	Working	Stay-home
Sunny	Broken	Stay-home
Rainy	Broken	Stay-home

$[CO_4, L_3]$

Ans. : Each input has only two values and the output class variable has two values. We can convert each variable to binary as follows :

- Weather : Sunny = 1, rainy = 0
- Car : Working = 1, broken = 0
- Class : Go-out = 1, stay-home = 0

Weather	Car	Class
1	1	1
0	0	1
1	1	1
1	1	1
1	1	1
0	0	0
0	0	0
1	1	0
1	0	0
0	0	0

- There are two types of quantities that need to be calculated from the dataset for the naïve bayes model : Class probabilities and conditional probabilities.
- Calculate the class probabilities : Dataset is a two class probabilities of classes 0 and classes 1.

$$P(\text{class} = 1) = \frac{\text{count}(\text{class} = 1)}{\text{count}(\text{class} = 0) + \text{count}(\text{class} = 1)}$$

$$P(\text{class} = 0) = \frac{\text{count}(\text{class} = 0)}{\text{count}(\text{class} = 0) + \text{count}(\text{class} = 1)}$$

$$P(\text{class} = 1) = \frac{5}{5+5}$$

$$P(\text{class} = 0) = \frac{5}{5+5}$$

- Probability of class 0 and class 1 is 0.5.
- Calculate the conditional probabilities : The conditional probabilities are the probability of each input value given each class value. The conditional probabilities for the dataset can be calculated as follows :

$$P(\text{weather} = \text{sunny}|\text{class}) = \text{go-out}) = \frac{\text{count}(\text{weather} = \text{sunny} \wedge \text{class} = \text{go-out})}{\text{count}(\text{class} = \text{go-out})}$$

$$P(\text{weather} = \text{rainy}|\text{class}) = \text{go - out}) = \frac{\text{count}(\text{weather} = \text{rainy} \wedge \text{class} = \text{go-out})}{\text{count}(\text{class} = \text{go - out})}$$

$$P(\text{weather} = \text{sunny}|\text{class} = \text{stay - home}) = \frac{\text{count}(\text{weather} = \text{sunny} \wedge \text{class} = \text{stay - home})}{\text{count}(\text{class} = \text{stay - home})}$$

$$P(\text{weather} = \text{rainy}|\text{class} = \text{stay - home}) = \frac{\text{count}(\text{weather} = \text{rainy} \wedge \text{class} = \text{stay - home})}{\text{count}(\text{class} = \text{stay - home})}$$

P (weather = sunny|class = go-out) = 0.8

P (weather = rainy|class = go-out) = 0.2

P (weather = sunny|class = stay-home) = 0.4

P (weather = rainy|class = stay-home) = 0.6

Similarly,

P (car = working |class = go - out)

$$= \frac{\text{count (car = working} \wedge \text{ class = go - out)}}{\text{count (class = go - out)}}$$

P (car = broken|class) = go - out)

$$= \frac{\text{count (car = broken} \wedge \text{ class = go - out)}}{\text{count (class = go - out)}}$$

P (car = working |class = stay - home)

$$= \frac{\text{count (car = working} \wedge \text{ class = stay - home)}}{\text{count (class = stay - home)}}$$

P (car = broken |class = stay - home)

$$= \frac{\text{count (car = broken} \wedge \text{ class = stay - home)}}{\text{count (class = stay - home)}}$$

P(car = working|class = go-out) = 0.8

P(car = broken|class = go-out) = 0.2

P(car = working|class = stay-home) = 0.2

P(car = broken|class = stay-home) = 0.8

Make predictions with naïve bayes :

Let's take the first record from our dataset and use our learned model to predict which class it belongs.

First instance : weather = sunny, car = working go-out

= P(weather = **sunny**|class = **go-out**)

× P(car = **working**|class = **go-out**) × P(class = **go-out**)

= 0.8 × 0.8 × 0.5 = 0.32

stay - home = P(weather = **sunny**|class = **stay-home**)

× P(car = **working**|class = **stay-home**) × P(class = **stay-home**)

= 0.4 × 0.2 × 0.5 = 0.04

- We can see that probability 0.32 > 0.04, therefore prediction is **go-out** for this instance, which is correctly classified.

Weather	Car	Class	Go-out	Stay-home	Prediction
sunny	working	go-out	0.32	0.04	go-out
rainy	broken	go-out	0.02	0.24	stay-home
sunny	working	go-out	0.32	0.04	go-out
sunny	working	go-out	0.32	0.04	go-out
sunny	working	go-out	0.32	0.04	go-out
rainy	broken	stay-home	0.02	0.24	stay-home
rainy	broken	stay-home	0.02	0.24	stay-home
sunny	working	stay-home	0.32	0.04	go-out
sunny	broken	stay-home	0.08	0.16	stay-home
rainy	broken	stay-home	0.02	0.24	stay-home

$$\textbf{Accuracy} = \frac{\text{The number of correct classification}}{\text{The total number of classification}} = \frac{8}{10} = 0.80$$

We obtained a 80 % of accuracy using naïve bayes algorithm.

b) *What is decision tree ? Explain ID-3 algorithm of decision tree in detail.* $[CO_4, L_2]$ *?*

(Refer Q.2 of Chapter - 4 and Q.3(a) of June - 2022) **[8]**

OR

Q.4 a) *For the following data calculate weighted average entropy for all features.*

Length = [3, 4, 5] [2+, 0 –] [1+, 3 –] [2+, 2 –]
Gills = [Yes, No] [0+, 4 –] [5+, 1 –]
Beak = [Yes, No] [5+, 3 –] [0 +, 2 –]
Teeth = [many, few] [3+, 4 –] [2+, 1–] $[CO_4, L_3]$ **[10]**

Ans. : Calculate impurity of the first split.

We have 3 segments

1) Pure and so has entropy 0

2) Entropy $= -(1/4)\log_2(1/4) - (3/4)\log_2(3/4) = 0.5 + 0.31 = 0.81$

3) Entropy is 1

The total entropy is then the weighted average of these, which is

$$= \frac{2}{10} \times 0 + \frac{4}{10} \times 0.81 + \frac{4}{10} \times 1 = 0.724$$

Similar calculates for the other three features give the following entropies :

$$\text{Gills} = \frac{4}{10} \times 0 + \frac{6}{10}\left[(-5/6)\log_2(5/6) - (1/6)\log_2(1/6)\right]$$

Gills = 0.39

$$\text{Beak} = \frac{8}{10} \times \left[-(5/8)\log_2(5/8) - (3/8)\log_2(3/8)\right] + \frac{2}{10} \times 0$$

Beak = 0.76

$$\text{Teeth} = \frac{7}{10} \times \left[-(3/7)\log_2(3/7) - (4/7)\log_2(4/7)\right]$$

$$+ \frac{3}{10}\left[-(2/3)\log_2(2/3) - (1/3)\log_2(1/3)\right]$$

Teeth = 0.97

"Gills" is an excellent feature to split on; "Teeth" is poor.

b) *Define and explain following terms*

i) Bayesian network.

ii) Advantages and disadvantages of naïve bayes classifier. $[CO_4, L_2]$ **[8]**

Ans. : i) Bayesian network : Refer Q.4(a) of June-2022.

ii) Advantages and disadvantages of Naive Bayes Classifier :

Advantages :

- Simple to implement.
- If the conditional Independence assumption holds, it could give great results.
- It is not sensitive to noisy features.
- No overfitting.
- Suitable for Large Datasets.

Disadvantages :

- It will assume that all the attributes are independent, which rarely happens in real life. It will limit the application of this algorithm in real-world situations.
- It will estimate things wrong sometimes.
- The Naive Bayes algorithm has trouble with the 'zero-frequency problem'. It happens when we assign zero probability for categorical variables in the training dataset that is not available.

Q.5 a) *Find all association rules using apriori algorithm in the following database with minimum support = 2 and minimum confidence = 65 %.* **[10]**

Transactions	*Data items*
T_1	*Pen, pencil, notebook*
T_2	*Pencil, file*
T_3	*Pen, pencil, notebook, file*
T_4	*Pen, notebook*
T_5	*Pencil, scale, file*
T_6	*Pencil, scale*
T_7	*Pen, pencil, scale*

$[CO_5, L_3]$

Ans. : Consider a database, D, consisting of 7 transactions. Suppose min. support count required is 2 (i.e. min_sup = 2/7 = 28 %)

- Let minimum confidence required is 65 %.
- We have to first find out the frequent itemset using Apriori algorithm. Then, association rules will be generated using min. support & min. confidence.

Step 1 : Generating 1-itemset frequent pattern

C_1

Itemset	Support count
{Pen}	4
{Pencil}	6
{Notebook}	3
{File}	3
{Scale}	3

Scan D for count of each candidate

L_1

Compare candidate support count with minimum support count

Itemset	Support count
{Pen}	4
{Pencil}	6
{Notebook}	3
{File}	3
{Scale}	3

- The set of frequent 1-itemsets, L_1, consists of the candidate 1-itemsets satisfying minimum support. In the first iteration of the algorithm, each item is a member of the set of candidate.

Step 2 : Generating 2-itemset frequent pattern

Generate C_2 candidate from L_1

Itemset
{Pen, Pencil}
{Pen, Notebook}
{Pen, File}
{Pen, Scale}
{Pencil, Notebook}
{Pencil, File}
{Pencil, Scale}
{Notebook, File}
{Scale, File}

C_2

Scan D for count of each candidate

Itemset	Support count
{Pen, Pencil}	3
{Pen, Notebook}	3
{Pen, File}	1
{Pen, Scale}	1
{Pencil, Notebook}	2
{Pencil, File}	3
{Pencil, Scale}	3
{Notebook, File}	1
{Scale, File}	1

C_2

Compare candidate support count with minimum support count

Itemset	Support count
{Pen, Pencil}	3
{Pen, Notebook}	3
{Pencil, Notebook}	2
{Pencil, File}	3
{Pencil, Scale}	3

L_2

Step 3 : Generating 3-itemset frequent pattern

Scan D for count of each candidate

Itemset
{Pen, Pencil, Notebook}
{Pen, Pencil, File}
{Pencil, Scale, File}
{Pen, Pencil, Scale}
{Pencil, Notebook, File}

C_3

Scan D for count of each candidate

Itemset	Support count
{Pen, Pencil, Notebook}	2
{Pen, Pencil, File}	1
{Pencil, Scale, File}	1
{Pen, Pencil, Scale}	1
{Pencil, Notebook, File}	1
{Pencil, Notebook, File}	1

C_3

Compare candidate support count with min support count

Itemset	Support count
{Pen, Pencil, Notebook}	2

L_3

- **The generation of the set of candidate 3-itemsets, C_3, involves use of the Apriori property.**

Step 4 : Generating 4-itemset frequent pattern

- The algorithm uses L_3 Join L_3 to generate a candidate set of 4-itemsets, C_4. There is no frequent. Thus, $C_4 = \phi$ and algorithm terminates, having found all of the frequent items. This completes our Apriori algorithm.
- These frequent itemsets will be used to generate strong association rules (where strong association rules satisfy both minimum support and minimum confidence).

Step 5 : Generating association rules from frequent itemsets

- Procedure :
- For each frequent itemset "l",generate all nonempty subsets of l.

- For every nonempty subset sof l, output the rule "s ⇒ (l-s)"if support_count(l) / support_count(s) >= min_confwhere min_confis minimum confidence threshold.
- "We had L = {{Pen}, {Pen}, {Pencial}, {Scale}, {File}, {Pen, Pencial}, {Pen, Notebook}, {Pencial, Notebook}, {Pencial, File}, {Pencial, Scale}, {Pen, Pencial, Notebook}, {Pen, Pencial, File}, {Pencial, Scale, File}, {Pen, Pencial, Scale}, {Pencial, Notebook, File}.
- Let minimum confidence threshold is, say 65 %.
- The resulting association rules are shown below, each listed with its confidence.

– R1 : T1 ^ T3 → T7

- Confidence = 3/2 = 150 %
- R1 is selected.

b) *What is use of K-means algorithm ? Explain centroid and medoid ? Explain different types of distances measures.* $[CO_5, L_2]$
(Refer Q.10 of Chapter - 5 and Q.6(b) of June-2022) **[8]**

OR

Q.6 a) *Explain following terms : i) Rule ii) Support iii) Lift iv) Confidence.* $[CO_5, L_2]$ **(Refer Q.21 and Q.23 of Chapter - 5)** **[8]**

b) *Apply KNN on the following data and classify the new sample (3, 7) to the respective class.* **[10]**

X	Y	Class
7	7	Pass
7	4	Pass
3	4	Fail
1	4	Fail
4	3	Fail
6	7	Pass
3	7	?

What will be the effect on output if K = 3 and K = 5 ? $[CO_5, L_3]$

Ans. : The distance formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Here $x_2 = 3$ and $y_2 = 7$ for all distance calculation

$$d_1 = \sqrt{(3-7)^2 + (7-7)^2} = 4$$

$$d_2 = \sqrt{(3-7)^2 + (7-4)^2} = 5$$

$$d_3 = \sqrt{(3-3)^2 + (7-4)^2} = 3$$

$$d_4 = \sqrt{(3-1)^2 + (7-4)^2} = 3.6$$

$$d_5 = \sqrt{(3-4)^2 + (7-3)^2} = 4.12$$

$$d_6 = \sqrt{(3-6)^2 + (7-7)^2} = 3$$

Rewritting :

x	y	Class	Distance
7	7	Pass	4
7	4	Pass	5
3	4	Fail	3
1	4	Fail	3.6
4	3	Fail	4.12
6	7	Pass	3
3	7	?	

Sort the table based on distance and assign rank

X	Y	Class	Distance	(K) Rank
3	4	Fail	3	1
6	7	Pass	3	2
1	4	Fail	3.6	3
7	7	Pass	4	4
4	3	Fail	4.12	5
7	4	Pass	5	6

Here K = 3, so class lable is fail and K = 5 class label is also fail. First two class label is fail, so next class label will be fail.

Q.7 a) *With the help of suitable diagram explain Biological Neuron.* $[CO_6, L_3]$ **(Refer Q.2 of Chapter - 6)** **[6]**

b) *What is the use of activation function in neural network ? Explain any two activation functions in detail.* $[CO_6, L_2]$
(Refer Q.10 of Chapter - 6) **[6]**

c) *What is deep learning ? Explain different applications of deep learning.* $[CO_6, L_1]$. **(Refer Q.21 and Q.22 of Chapter - 6)** **[5]**

OR

Q.8 a) *What is perceptron ? Explain multilayer perceptron in detail.* $[CO_6, L_3]$ **(Refer Q.1 and Q.13 of Chapter - 6)** **[6]**

b) *Write a note on following activation functions.*
i) Sigmoid ii) Tanh iii) ReLU.
$[CO_6, L_2]$ **(Refer Q.8(a) of June - 2022)** **[6]**

c) *What is ANN ? Explain McCulloch Pitts Neuron.* $[CO_6, L_2]$
(Refer Q.4 and Q.8 of Chapter - 6) **[5]**

JUNE - 2023 [END SEM] [6003] - 703 **Solved Paper**

Course 2019

Time : 2 $\frac{1}{2}$ Hours] **[Maximum Marks : 70**

Q.1 a) *State and explain need of regression analysis.* $[CO_2, L_1]$ **[4]**

Ans. : Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.

- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately.
- So for such case we need Regression analysis which is a statistical method and used in machine learning and data science. Below are some other reasons for using Regression analysis :
  - i) Regression estimates the relationship between the target and the independent variable.
  - ii) It is used to find the trends in data.
  - iii) It helps to predict real/continuous values.

b) *How gradient descent does helps to optimize linear regression model ?* $[CO_2, L_1]$ **(Refer Q.1(a) June-2022)** **[6]**

c) *What are the different ways to prevent overfitting.* $[CO_2, L_2]$
(Refer Q.22 of Chapter - 3) **[8]**

OR

Q.2 a) *What are different cost functions to access the performance of linear Regression model ? In the given dataset the outliers represent anomalies. Which cost function will be more suitable and why ?* **[5]**

$[CO_3, L_3]$

Ans. : Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution and arise due to inconsistent data entry or erroneous observations.

- Outlier can be of two types : Univariate and Multivariate.
- Methods is used to detect an outliers : Hypothesis Testing, Z-score method, DBSCAN Clustering, Isolation Forest, Linear Regression Models (PCA, LMS) and Standard Deviation. Also Refer Q.8 of Chapter - 3.

b) *Define of multivariate regression and state advantages and disadvantages of multivariate regression.* $[CO_2, L_2]$ **[5]**

Ans. : For multivariate regression : Refer Q.16 of Chapter - 3

Advantages :

- The multivariate regression method helps us to find a relationship between multiple variables or features.
- It also defines the correlation between independent variables and dependent variables.

Disadvantages :

- Multivariate regression technique requires high-level mathematical calculations. It is complex.
- The output of the multivariate regression model is difficult to analyse.
- The loss can use errors in the output.
- Multivariate regression yields better results when used with larger datasets rather than small ones.

c) *Consider the following data :*

Sr. No.	Prize in Rs.	Amount Demanded
1	10	40
2	11	38
3	16	48
4	18	40
5	20	60

i) Find values β_0 *and* β_1 *w.r.t. linear regression model which best fits given data.*

ii) Interpret and explain equation of regression line.

iii) Estimate the likely demand when the price is Rs. 15.

$[CO_2, L_3]$ **[8]**

Ans. :

X_i	Y_i	X^2	XY	Y^2	$(X_i - \overline{X})$	$(Y_i - \overline{Y})$
10	40	100	400	1600	– 5	– 5.2
11	38	121	418	1444	– 4	– 7.2
16	48	256	768	2304	1	2.8
18	40	324	720	1600	3	– 5.2
20	60	400	1200	3600	5	14.8
$\Sigma X = 75$	$\Sigma Y = 226$	$\Sigma X^2 = 1201$	$\Sigma XY = 3506$	$\Sigma Y^2 = 10548$		

$\overline{X} = 75/5$, $\overline{Y} = 226/5 = 45.2$

$$S_X = \sqrt{(\Sigma X - \overline{X})^2 / (n-1)} = \sqrt{(75-15)^2 / 4} = 30$$

$$S_Y = \sqrt{(\Sigma Y - \overline{Y})^2 / (n-1)} = \sqrt{(226-45.2)^2 / 4} = 90.4$$

We also need to compute the square of the deviation scores :

X_i	Y_i	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X}) * (Y_i - \overline{Y})$
10	40	25	27.04	676
11	38	16	51.84	829.44
16	48	1	7.84	7.84
18	40	9	27.04	243.36
20	60	25	219.04	5476
$\Sigma X = 75$	$\Sigma Y = 226$	$(X_i - \overline{X})^2 = 76$	$(Y_i - \overline{Y})^2 = 332.8$	$\Sigma(X_i - \overline{X}) * (Y_i - \overline{Y})$ = 7232.64

The regression equation is a linear equation of the form

$$\hat{Y} = \beta_0 + \beta_1 X$$

First, we solve regression coefficient (β_1)

$$\beta_1 = \Sigma[X_i - \overline{X} * (Y_i - \overline{Y}) / \Sigma(X_i - \overline{X})^2 = 7232.64/76 = 95.166$$

Once we know the value of the regression coefficient (β_1), we can solve for the regression slope (β_0)

$$\beta_0 = \overline{Y} - \beta_1 \overline{X} = 45.02 - 95.166 \times 15 = -1382.47$$

Q.3 a) *Consider following data. Which feature will be selected as a root node ? Use information gain. Played football is dependent feature.* $[CO_4, L_3]$ **(Refer Q.16 of Chapter - 4)** **[10]**

Outlook	*Temperature*	*Humidity*	*Wind*	*Played football (yes/no)*
Sunny	*Hot*	*High*	*Weak*	*No*
Sunny	*Hot*	*High*	*Strong*	*No*
Overcast	*Hot*	*High*	*Weak*	*Yes*
Rain	*Mild*	*High*	*Weak*	*Yes*

Rain	*Cool*	*Normal*	*Weak*	*Yes*
Rain	*Cool*	*Normal*	*Strong*	*No*
Overcast	*Cool*	*Normal*	*Strong*	*Yes*
Sunny	*Mild*	*High*	*Weak*	*No*
Sunny	*Cool*	*Normal*	*Weak*	*Yes*
Rain	*Mild*	*Normal*	*Weak*	*Yes*
Sunny	*Mild*	*Normal*	*Strong*	*Yes*
Overcast	*Mild*	*High*	*Strong*	*Yes*
Overcast	*Hot*	*Normal*	*Weak*	*Yes*
Rain	*Mild*	*High*	*Strong*	*No*

b) *Define and Explain following terms*

i) Bayesian network **(Refer Q.4(a) of June-2022)**

ii) Advantages and disadvantages of Naive Bayes classifier $[CO_4, L_2]$ **(Refer Q.4(b) of Dec.-2022)** **[7]**

OR

Q.4 a) *For the given data set apply Naive Bayes Classifier and predict the Class for Type of Type of family structure = Single Parent, Age group = Young and Income status = Low* $[CO_4, L_3]$

(Refer similar Q.15 of Chapter - 4) **[10]**

Type of family structure	*Age group*	*Income status*	*will they buy a car?*
Nuclear	*Young*	*Low*	*Yes*
Extended	*old*	*Low*	*No*
Childless	*Middle-aged*	*Low*	*No*
Childless	*Young*	*Medium*	*Yes*
Single Parent	*Middle-aged*	*Medium*	*Yes*
Childless	*Young*	*Low*	*No*
Nuclear	*Old*	*High*	*Yes*

Nuclear	Middle-aged	Medium	Yes
Extended	Middle-aged	High	Yes
Single Parent	Old	Low	No

b) *Define and explain following terms*
i) Minority class ii) Gini index iii) Entropy iv) Information gain $[CO_4, L_2]$ **[7]**

Ans. : i) Minority Class : A classification data set with skewed class proportions is called imbalanced. Classes that make up a smaller proportion are minority classes.

ii) Gini Index : Refer Q.3(b) of June-2022.

iii) Entropy : Refer Q.3(b) of June-2022.

iv) Information Gain : Refer Q.3(b) of June-2022.

Q.5 a) *Find all association rules in the following database in the following database with minimum support = 2 and minimum confidence = 75 %* $[CO_5, L_3]$
(Refer similar Q.26 of Chapter - 5) **[10]**

Transactions	Data Items
1	Bread, Milk, Diaper
2	Bread, Milk, Diaper, Coke
3	Diaper, Beer, Eggs
4	Bread, Milk, Coke

b) *State and explain with appropriate example different types of linkage use in clustering.* $[CO_5, L_2]$ **(Refer Q.11 of Chapter - 5)** **[8]**

OR

Q.6 a) *Explain following terms*
i) Rule ii) Support iii) Lift iv) Confidence $[CO_5, L_2]$
(Refer Q.23 of Chapter - 5) **[8]**

b) *Apply KNN on the following data. Find class of person whose height is 170 cm and weight is 57 kg. Consider value K = 5 and use Euclidian distance formula.* $[CO_5, L_3]$
(Refer similar Q.7 of Chapter - 5) **[10]**

Height (cm)	Weight (kg)	Class
167	51	Under weight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Under weight
169	58	Normal
173	57	Normal
170	55	Normal

Q.7 a) *With the help of suitable diagram explain biological neuron.* $[CO_6, L_3]$ **(Refer Q.2 of Chapter - 6)** **[5]**

b) *Explain the architecture of feed forward neural network. State its limitations.* $[CO_6, L_2]$ **[7]**

Ans. : Feed Forward Neural Network is an artificial neural network in which the connections between nodes does not form a cycle. The feed forward model is the simplest form of neural network as information is only processed in one direction. While the data may pass through multiple hidden nodes, it always moves in one direction and never backwards.

- They are called feed forward because information only travels forward in the network (no loops), first through the input nodes, then through the hidden nodes (if present) and finally through the output nodes.
- Feed-forward networks tends to be simple networks that associates inputs with outputs. It can be used in pattern recognition. This type of organization is represented as bottom-up or top-down.

- Fig. 1 shows basic structure of a Feed Forward (FF) Neural Network.

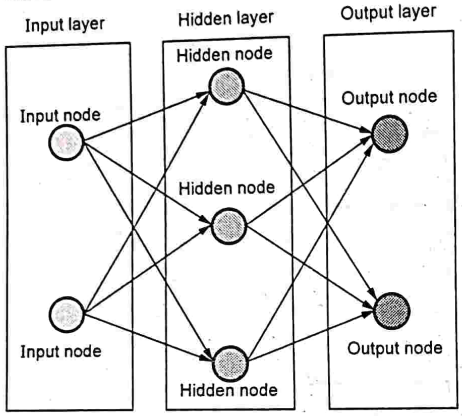


Fig. 1

- Input layer contains one or more input nodes. For example, suppose we want to predict whether it will rain tomorrow and base our decision on two variables, humidity and wind speed. In that case, our first input would be the value for humidity, and the second input would be the value for wind speed.
- **Hidden layer :** This layer contains an activation function. The hidden layers are positioned between the input and the output layer. The number of hidden layers depends on the type of model. Hidden layers have several neurons that impose transformations on the input before transferring. The weights in the network are constantly updated to make it easily predictable.
- Output layer contains one or more output nodes.
- Feed foward neural networks are primarily used for supervised learning in cases where the data to be learned is neither sequential nor time-dependent.

- Feed-forward networks have the following characteristics :
  1. Perceptron's are arranged in layers, with the first layer taking in inputs and the last layer producing outputs. The middle layers have no connection with the external world, and hence are called hidden layers.
  2. Each perceptron in one layer is connected to every perceptron on the next layer. Hence information is constantly "fed forward" from one layer to the next and this explains why these networks are called feed-forward networks.
  3. There is no connection among perceptron's in the same layer.
- Limitations :
  1. Loss of neighborhood information.
  2. More parameters to optimize
  3. These networks require large amounts of data in order to function properly, as well as a high level of computational power.

c) *What is deep learning ? Explain different applications of deep learning.* $[CO_6, L_1]$ **(Refer Q.7(c) of Dec.-2022)** **[5]**

OR

Q.8 a) *What is perceptron ? Explain multilayer perceptron in detail.* $[CO_6, L_3]$ **(Refer Q.8(a) of Dec.-2022)** **[5]**

b) *Explain why we use non-linearity function ? State and explain three types of neurons that add non-linearity in their computations.* $[CO_6, 12]$ **(Refer Q.7(b) of Dec.-2022)** **[7]**

c) *What is ANN ? Explain McCulloch Pitts Neuron* $[CO_6, L_2]$ **(Refer Q.8(c) of Dec.-2022)** **[5]**

END...