UNIT - 02

Classification..

- Supervised learning
  It is the process of categorizing / grouping the data based on predefined (labeled) data.
- Goal:
  Is to develop a model that can effectively assign a label & category

Phases :

1. First phase : A model is created based on the trained data
2. Second phase : applied the 1st phase to categories
   Ex: Granting loan, input : income, age, loan, Output : loan take & no / high & low

Methods

1. Specifying Boundaries [threshold value & Design
   - Executed by partitioning the potential database tuple into distinct region
   - Each region being linked to specific data.

2. Using probability Distribution

- Involves utilizing the probability density function (PDF)

3. Using posterior prob. $P(t_1, t_2)$
   - Denoted as $P(t_1 | c_2)$
     - $t_1 \rightarrow$ likelihood of a specific point / attribute
     - $c_2 \rightarrow$ class c_2

Issues in classification

1. Overfitting
2. Underfitting
3. Noise in data
4. Feature Selection.
5. Scalability.

Measuring Performance in classification

- Used to evaluate the accuracy of a classification algorithm in predicting the correct class for a given instance
- One of the approach is
  - * Confusion matrix.

"classified" model: Accuracy, Recall, Precision, F1-score, AUC

"Regression": RMSE, R^2

"Segmented": IOU, pixel Accuracy, Recall precision.

What data should be used to measure the performance

- As the models are tuned w.r.t 'training set' during the process, the output predicted
- Therefore, inorder to estimate the generalization error, → the model required to eval through unseen data.
- ∴ Use Test data.

- Used to describe the perform of a classification model.

Actual	Predicted Neg	Predic Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

1. True positive (TP) : Instance that is correctly classified as belonging to the positive class

2. True Negative (TN): Instance that is correctly classified as not belonging to the positive class

3 False Negative :- Instance is incorrectly classified as not belonging to the positive class

4 False Positive :- Instance is incorrectly classified as belonging to positive class

1. Accuracy :- Probability of correct prediction.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

2. Precision :- Proportion of true positive among the instances that the model predicted as positive.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

3. Recall :- $\frac{TP}{(TP + FN)}$

4. F Score :- $\frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$

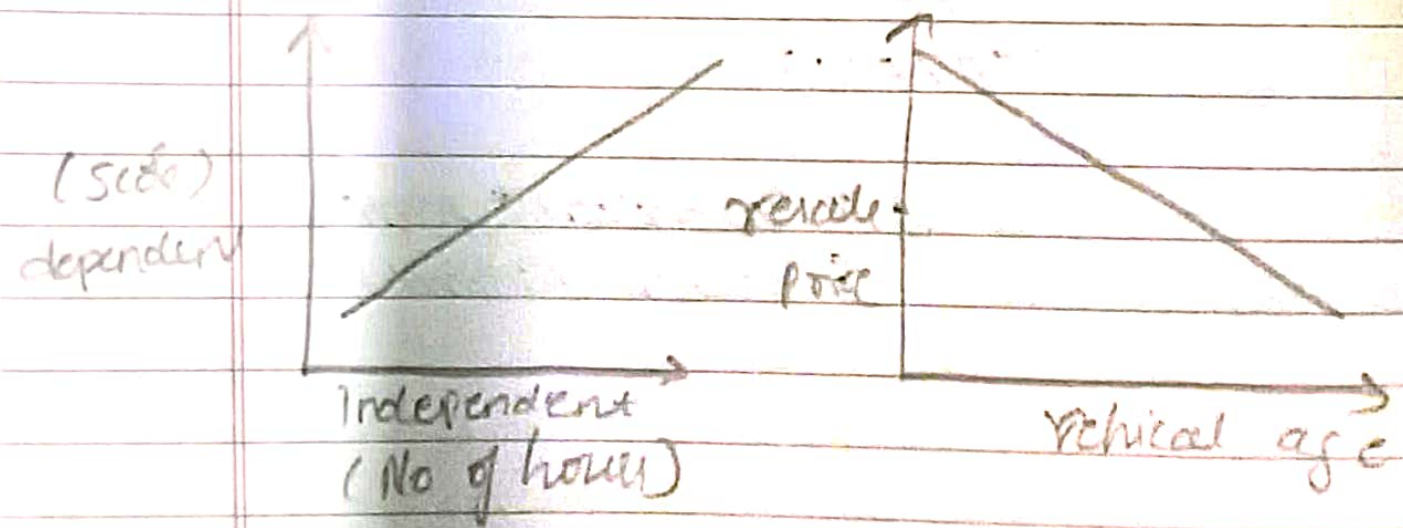
Statistical Based algorithms

classification models

1. Regression Analysis
2. Bayesian Classification.

Regression Analysis

- Regression is a statistical method that helps us to understand & predict the relat°ship btw variables.
  variable → terms ← Independent / Dependent
- Describe how one variables varies if another changes.
- Ex: Score based on hours, resale price based on vehical age



+ve linear regression

-ve linear regression

Algorithms and metrics

1. Data Mining Metrics
   1. Return on Investment (ROI)
   2. Business or Usefulness perspective
   3. Space and Time complexity
   4. Accuracy in classification
   5. Specific Metrics for DM task
   6. Effectiveness of the implementation

2. Statistical Perspective on DM:
   1. Point Estimate
   2. Bias of an estimator (esd)
   3. Squared error
   4. Mean square error (MSE)
   5. Root mean square
   6. Variance of estimator
   7.
   8. Confident interval
   9. Jackknife Estimate
   10. Maximum likelihood Estimate (MLE)

8. Metrics for text similarity in DM

1. Cosine Similarity $\cos\theta = \frac{A \cdot B}{|A||B|}$

2. Jaccard Sim $= \frac{A \cdot B}{\sqrt{A}\sqrt{B}}$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

3. Pearson correlation coefficient.

$$P_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

$\bar{x}$ ⇒ mean

σ ⇒ SD

4. Sorensen - Dice coefficient.

$$S(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

$|A|$ & $|B|$ ⟶ Size of sets

4 Dissimilarity measures

↓ ↑ Dissimilarity Distance ↑ ↓

- 1 → complete dissimilarity
- 0 → — Similarity

1. Euclidean Distance

$$d(x,y) = \sqrt{\Sigma (x_i - y_i)^2}$$

2 Manhattan Distance

$$d(x,y) = \Sigma |x_i - y_i|$$

5 Measuring performance in classification

1. Accuracy $= \frac{(TP + TN)}{(TP + TN + FP + FN)}$

2. Precision $= \frac{TP}{(TP + FP)}$

3. Recall $= \frac{TP}{(TP + FN)}$

4. F1 score $= 2 * \frac{(Pre \times recall)}{(Pre + recall)}$

6. Attribute Selection Measure (ASM)

1. Information Gain :

a. Cal Entropy

b Cal average entropy

c Gain

$$Gain(A) = Entro(D) - Entro_A(D)$$

2. Gain Ratio :

$$Gain\ Ratio = \frac{Gain(A)}{Split\ Info(D)}$$

$$Split\ Info(D) = -\sum \frac{|D_i|}{|D|} \log \frac{|D_i|}{|D|}$$

3. Gini Index :

$$Gini(D) = 1 - \sum P_i^2$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D_2|} Gini(D_2)$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$