AML

LAB ASSIGNMENT - 2

PROBLEM STATEMENT - To study feature engg and implement the dimensionality reduction techniques (PCA & TSNE)

OBJECTIVES -

1. To understand feature engg & learn feature selection techniques
2. To understand the concept of dimensionality reduction.

THEORY -

→ Feature Engg

- Process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Some feature engg methods are -

1. Binning
2. Log transform
3. Scaling
4. Feature selection.

→ Feature selection techniques

- Process where you automatically or manually select those features which contribute to your prediction value or o/p in which you are interested in. This method can deal w/

1. Multicollinearity

(2) High date dimension
(3) High training time

Feature selection methods
(1) Forward selection
(2) Backward elimination

→ Dimensionality reduction (PCA, LDA & t-SNE)

- Filter method

(1) <u>ANOVA</u> - Analysis of variance is a method used to analyze the diff among the group means in a sample

(2) <u>Chi square</u> - This is defined as where O_i is the observation & E_i is the expected value. This value of chi can be used to derive the p-value that gives us the probability of independence. If p value is high (> 0.05), we can say that the attr is not statistically significant to the target var.

(3) Pearson's correlation - This is defined as The value is the measure of strength of linear association b/w 2 variables, where $r = 1$ means a perfect +ve correlation and the value $r = 1$ means a perfect -ve correlation.

- Wrapper method

(1) RFE - It is a wrapper type feature selection algo. This means that a different ML algo is given & used in core of the method & wrapped by RFE and used to help select features

- Intrinsic method

(1) DT - A decision tree is a method which can be traversed based on the attr values and can give an intrinsic value at the leaf nodes.

- PCA

* Algo that uses the eigen values derived from the correlation matrix in order to reduce the dimension of the dataset. The reduced features are representative of dataset but doesn't hold any meaning on its own.

- t-SNE

* t-distributed stochastic gradient descent neighborhood embedding is ML algo that employees stochastic neighbor embedding to reduce the no of attr by projecting them on low dimension space.

CONCLUSION - feature Engg was studied & implemented the two dimensionality reducti techniques PCA & t-SNE.

FAQs

Q1) What are the various dimensionality reduction techniques?

Ans) The various dimensionality reduction techniques are -

1) Ratio of missing values
2) Low variance in the column values
3) High correlation b/w 2 columns
4) Principal component analysis (PCA)
5) Candidates and split columns in random forest
6) Backward feature elimination
7) Forward feature construction
8) Linear discriminant analysis (LDA)
9) Neural autoencoder
10) t- distributed stochastic neighborhood embedding (t-SNE)

Q2) Define feature engg?

Ans) Process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve performance of ML algos.

Q7) Difference b/w t-SNE & PCA ?

Ans)

PCA	t-SNE
(1) Linear dimensionality reduction technique	(1) Non linear dimensionality reduction technique
(2) Tries to preserve the global structure of data	(2) Tries to preserve local structure (cluster) of data
(3) Deterministic algo	(3) Non deterministic or randomised algo
(4) Works by rotating the vectors for preserving variance	(4) Works by minimising the distance b/w the point in a gaussian