Page No.
Date

Module 3

Q1) Naive Bayes classifier

=> (1) Naive Bayes is a probabilistic classification algorithm based on Bayes Theorem, commonly used for classification tasks

(2) It assumes that the features of the data are independent, which is the 'naive' assumption

(3) Steps to make prediction using naive bayes

(i) Bayes theorem formula

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

where,

$P(C|X)$: Posterior probability of class C given X

$P(X|C)$: Likelihood of X given class C

$P(C)$: Prior probability of class C

$P(X)$: Normalization factor

(ii) Step-by-step prediction

(a) Calculate the probability of each class $P(C)$

(b) For each feature in X, compute $P(X|C)$

(c) Use the formula to compute $P(C|X)$ for all possible classes.

(d) Choose the class with the highest posterior probability.

(iii) The 'Naive' assumption

(a) The algorithm assumes that all features are conditionally independent of each other given the class.

(b) This means:

$$P(X|C) = P(x_1|C) \cdot P(x_2|C) \cdot \ldots\ldots \cdot P(x_n|C)$$

(c) In practice, this assumption rarely holds, as features often interact.

Page No.
Date

④ Example:

ⓐ Scenario: Spam email classification
Features: keywords in a email (win, offer)
Classes: Spam (S) or Not spam (N)

ⓑ P(S): Probability that an email is spam
P(N): Probability that an email is not spam

ⓒ For a word like 'win', calculate
P(win|S) and P(win|N) based on data.

ⓓ Given an email with words win & offer, compute:
$P(S \mid win, offer) = P(win \mid S) \cdot P(offer \mid S) \cdot P(S)$
$P(N \mid win, offer) = P(win \mid N) \cdot P(offer \mid N) \cdot P(N)$

ⓔ Compare $P(S \mid win, offer)$ and $P(N \mid win, offer)$.

ⓕ Assign the label with the one having higher probability.

⑤ Real-world applications.

(i) Sentiment analysis
Classify customer reviews as 'positive' or 'negative'

(ii) Spam filtering
Classify emails as 'spam' or 'not spam' based on presence of certain keywords.

(iii) Disease prediction
Predict diseases such as 'diabetes' or 'cancer' or 'heart problem' by considering symptoms like 'sugar level', 'blood pressure', etc.

(iv) Movie recommendation
Predict genres a user might like based on previously watched movies.

Page No.
Date

Q2) Methods for estimating a classifier's accuracy.

=> (1) Holdout method
(i) In the holdout method, the largest dataset is randomly divided into three subsets:
(a) Training Set : Used to train the model
(b) Validation step : Helps to adjust the model settings to find the best version.
(c) Test set: Used to evaluate the model's performance on unseen data.
(ii) Commonly, 2/3 of the data is used for training and 1/3 for testing.
(iii) This method prevents the model from being too specific to the training data by testing it on new data
(iv) Advantages
(a) Simple to implement
(b) Efficient for large datasets.
(2) Random Subsampling
(i) An extension of the holdout method where the splitting process is repeated multiple times.
(ii) Each time, a random subset is chosen for training and testing.
(iii) The model's performance is averaged over all iterations.
(iv) Process :
(a) Split the data randomly into training and test sets.
(b) Train the model and evaluate using the test set.
(c) Compute the error metric such as mean squared error (MSE).

(3) Cross - validation
(i) Divides the dataset into k equal-sized subsets (folds)
(ii) Each subset is used as a test set while the remaining k-1 subsets form the training set.
(iii) The process is repeated k-times, and the results are averaged.
(iv) Advantages:
(a) Efficient use of the entire dataset for training and testing.
(b) Provides an unbiased estimate of the model's performance.
(v) Disadvantage
(a) Computationally expensive for large datasets.
(4) Bootstrapping.
(i) Bootstrapping is used to make estimations from the data by taking an average of the estimates from smaller data samples.
(ii) Process
(a) Resample the dataset to create new training sets.
(b) Train the model on each resampled data
(c) Combine predictions or results from all the samples.
(iii) Advantages:
(a) Effective for smaller datasets.
(5) Applications
(i) Holdout : Fraud detection
(ii) Random Subsampling : Iterative experiments.
(iii) Cross-validation : Machine learning competition
(iv) Bootstrapping : Medical research

Q8) Write a short note on Decision tree.

⇒ (1) A decision tree is a flowchart-like structure which is used to make decisions or predictions.
(2) Decision tree classifier has tree type structure which has root node, internal nodes, branches and leaf nodes.
(3) The root node represents the entire dataset and the initial decision to be made.
(4) The internal nodes represent decisions or tests on attributes. Each internal node has one or more branches.
(5) The branches represent the outcome of a decision or test, leading to another node.
(6) The leaf nodes represent the final decision or prediction. No further splits occur at this node.
(7) The process of creating a decision tree involves.
(i) Selecting the best attribute: Using a metric like entropy, the best attribute to split the data is selected.
(ii) Splitting the dataset: The dataset is splitted into subsets based on the selected attributes.
(iii) Repeating the process: The process is repeated recursively for each subset, creating a new internal node or leaf node until a termination condition is met.
(8) Advantages:
(i) Easy to understand and interpret.
(ii) Decision trees are versatile.
(iii) They do not require normalization.
(9) Applications of decision tree: Buisness decision

making, healthcare, finance, marketing.

(10) Decision tree representation for playing tennis.