Experiment No: 3 Date: 23/01/2025

Title:- Implementation of Bayes Classifier for data Classification.

Aim: to implement a Bayes classifier to classify data into different categories using probabilistic reasoning.

Theory:
The Bayes classifier is a Supervisied learning Algorithm based on Bayes theorem, which provide a probabilistic approach to classify datapoint. It predict the Class of a given data instance calculate the posterior probability of each class and selecting the one with the highest probability.

Bayes Theorem: $$P(C_k/X) = \frac{P(X/C_k) \cdot P(C_k)}{P(X)}$$

$P(C_k/X)$: Posterior probability of class C_k given the data x.
$P(X/C_k)$: Likelihood of observing data X given the class C_k.
$P(C_k)$: Prior probability of class C_k (based on prior knowledge)
$P(X)$: Evidence or normalization factor, calculated as the sum of the probabilities of X over all classes.

The Naive Bayes classifier simplifies computation by assuming that the features are conditionally independent given the class level.

If the features are continuous, assuming that the data for each features follow a Gaussian (Normal) Distribution. The probability density function (PDF):

$$P(X/C_k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

x : feature value

μ : The mean (average) of the feature for class c_k

σ^2 : The variance of the feature for class C_k

$P(x/C_k)$: The probability of observing x given that the data belongs to class c_k.

for the classification.

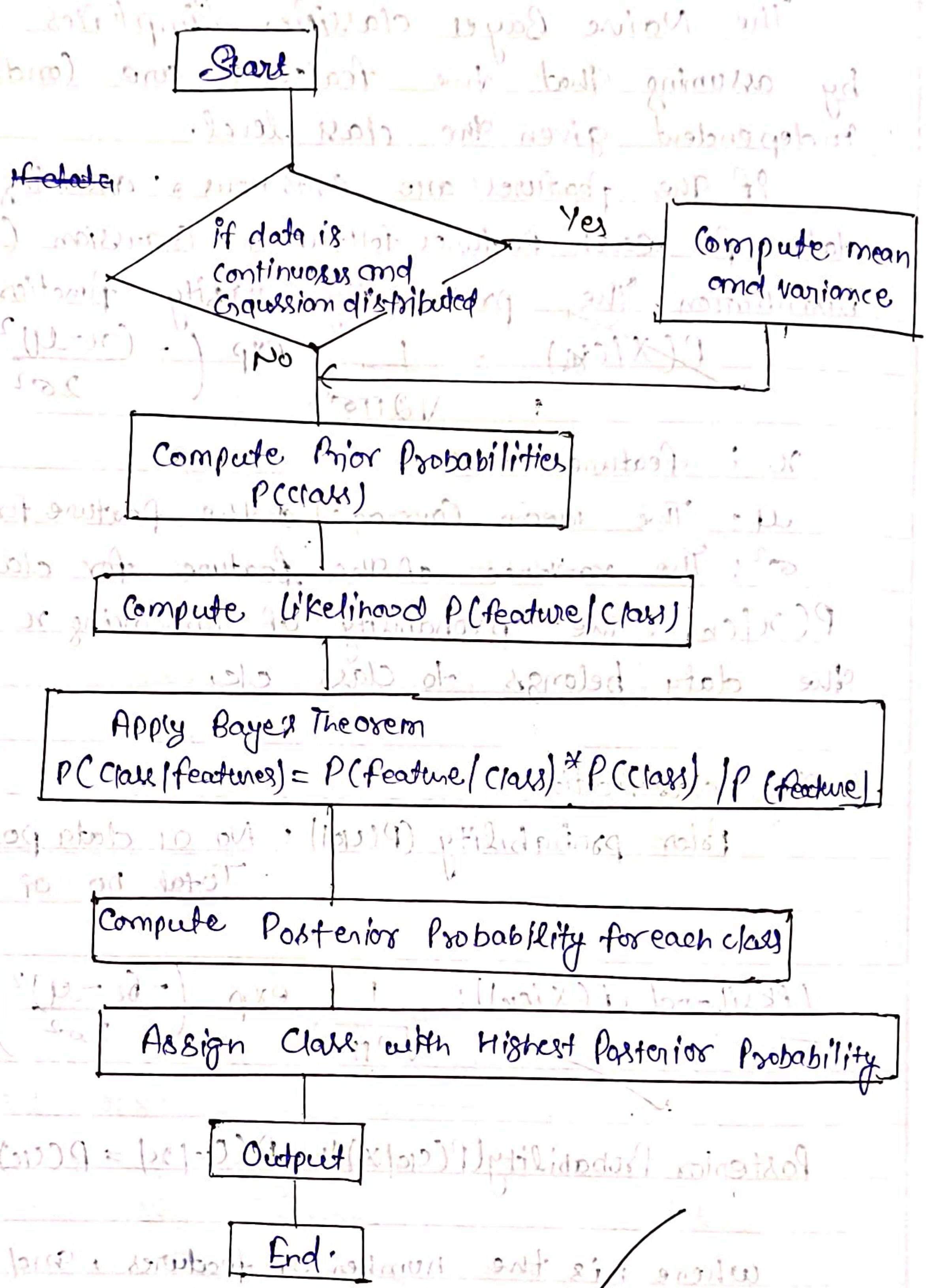
Prior probability $(P(C_k))$: $\frac{\text{No of data points in class } C_k}{\text{Total no of data points.}}$

Likelihood $(P(X/C_k))$: $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$

Posterior Probability $(P(C_k/X))$: $P(C_k/x) = P(C_k) \cdot \prod_{i=1}^{n} P(x_i/C_k)$

where n is the number of features, and x_i is the value of the ith features.

Flowchart:


Start.
if data is continuous and Gaussian distributed
Yes
Compute mean and variance
No
Compute Prior Probabilities P(class)
Compute likelihood P(feature/class)
Apply Bayes Theorem
P(class/features) = P(feature/class) * P(class) / P(feature)
Compute Posterior Probability for each class
Assign Class with Highest Posterior Probability
Output
End.

Class prediction: Select the class with the highest posterior probability $= \max_{\substack{C_k \\ k=1..K}} (P(C_k/X))$.

The Algorithm:

→ Dataset with n features and m data points
→ target classes $C_1, C_2, \dots C_K$.
→ for each class C_k, compute the mean μ and variance (σ^2) of each feature using the training data if the data is continuous gaussian distributed.
→ Compute the prior probability for each class.

$$P(C_k) = \frac{\text{No of data point in class } C_k}{\text{total no of data points}}$$

→ Liklihood Calculation.
for a given test point $X = \{x_1, x_2 \dots x_n\}$ calculate the liklihood of each feature using

$$P(x_i / C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-(x_i - \mu_k)^2}{2\sigma_k^2}\right)$$

→ Posterior Calculation:
Combine the prior and likelihood for all features to calculate the posterior probability for each class

$$P(C_k/X) = P(C_k) \cdot \prod_{i=1}^{n} P(x_i / C_k)$$

→ Class C_k with the highest posterior probability.

Result:

Dataset:

Person	height (feet)	weight (lbs)	footsize (inches)
male	6	180	12
male	5.92	190	11
male	5.58	170	12
male	5.92	165	10
female	5	100	6
female	5.5	150	8
female	5.42	130	7
female	5.75	150	9

Sample data (height: 6, weight: 180, footsize: 8)

fruit	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	500	650	800	1000

--- Mean and Variance for Male Data ---

	Mean	Variance
Height	5.855	0.035033
Weight	176.25	122.92
Foot Size	11.25	0.91667

--- Mean and Variance for Female Data ---

	Mean	Variance
Height	5.4175	0.097225
Weight	132.5	558.33
Foot Size	7.5	1.6667

--- Probabilities for Test Data ---

Male probability for Height: 1.578883

Female probability for Height: 0.223459

Male probability for Weight: 0.000006

Female probability for Weight: 0.016789

Male probability for Foot Size: 0.001311

Female probability for Foot Size: 0.286691

--- Prior Probabilities ---

Prior probability for Male: 0.50

Prior probability for Female: 0.50

--- Posterior Probabilities ---

Posterior probability for Male: 0.00011

Posterior probability for Female: 0.99998

Result: The test data is classified as Female

Discussion:

→ In the first dataset included height, weight and footsize measurement for male and female. These features were used to calculate probabilities and classify the given test data.

→ For each features the mean and variance were computed separately for male and female data. Using the gaussian probability density function, the likelihood P(feature/male) and P(feature/female) were calculated for the test data. [6, 130, 8]

→ The prior probabilities (P(male) & P(female)) were calculated based on the size of male and female data.

→ The posterior probabilities P(male/data) and P(female/data) were computed by multiplying the likelihoods of all features with their respective prior.

→ The Based on the posterior probabilities the test data was classified as female. The test data better matched the male data distribution than the female distribution, resulting in a female classification.

→ The bayesian classifier successfully combined prior probabilities and feature likelihood to reach an interpretable and probabilistic decision.

→ The bayesian classifier is effective in gender classification using continuous features. However increasing the dataset size and incorporating dependencies b/w features could improve performance.

--- Prior Probabilities ---

Banana: 0.4048

Orange: 0.1786

Other: 0.4167

--- Likelihood Probabilities ---

Banana (P(Features|Fruit)): 0.1212

Orange (P(Features|Fruit)): 0.0000

Other (P(Features|Fruit)): 0.0400

--- Posterior Probabilities ---

Banana (P(Fruit|Features)): 0.7463

Orange (P(Fruit|Features)): 0.0000

Other (P(Fruit|Features)): 0.2537

The given data (Long, Sweet, Yellow) corresponds to: Banana

→ For the second dataset with three fruit (Banana, Orange, Other and with feature long, sweet, and yellow. Each fruit was representing by its association with these features in terms of count.

→ The prior probability of each fruit was calculated based on the total no of features count for that fruit relative to the dataset.

→ The likelihood P(features/fruit) was computed for each fruit by analyzing how strongly the features matched the characteristics of each fruits. Banana showed the heighest one, reflecting its strong alignment with the feature.

→ Using the bayesian theorem the posterior probabilities were calculated by combining the liklihoods and prior. Banana emerged with the highest posterior probability, indicating that ~~the~~ features most likely correspond to a banana.

→ The experiment classified clear and interpretable result, with probabilities calculated at each step. This transperancy helps understand why a particular fruit ~~cuts~~ choosen.

→ The dataset assumed all features contribute equally to Classification. This could limit accuracy if certain features are more important than others.

→ The small data set constrained the generalizability of result. Larger dataset with more diverse example would improve the model's reliability.

Conclusion:

The Bayesian classifier effectively classified the test data as female, leveraging height, weight and foot size. By combining prior and likelihood probabilities the model provided an interpretable and accurate decision. However a larger data set and consideration of feature dependencies could enhance performance.

Using this, the test data was accurately classified as Banana based on features. The method's probabilistic approach effectively matched features with prior knowledge demonstrating its robustness for categorial classification tasks.

30.01.25