Naive Bayes' Classifier

Ipshita Bonhi Upoma

Department of Computer Science and Engineering

School of Data and Sciences

Introduction to Classification Problem

- A classification problem involves predicting the category or class of an object based on features.
- It is a type of supervised learning.
- The model is trained using labeled data to predict the class of unseen data.

- **Feature:** A feature is an individual measurable property or characteristic of the data. It describes aspects of the data that can help differentiate between classes. For example, in an email spam detection problem, features could include the frequency of certain words or the sender's email address.
- **Class:** The class is the label or category that the model aims to predict. It is the target variable in a classification problem. For example, in spam detection, the classes would be "spam" or "not spam."
- **Classifier:** A classifier is an algorithm that uses the features to predict the class label. It learns from labeled training data to establish a relationship between the features and their corresponding classes.

- The classifier predicts a class label based on the input features and the patterns it learned during training.
- The performance of the classifier heavily depends on the quality and relevance of the features.

What is Naive Bayes Classifier?

Definition: A probabilistic classifier based on Bayes' Theorem with a strong independence assumption between features.

- Assumes that each feature contributes independently to the outcome.
- Used for classification tasks.

Applications:

- Spam detection
- Sentiment analysis
- Document classification
- Medical diagnosis

Bayes' Theorem

Formula:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Explanation:

- $P(A)$: Prior probability of A.
- $P(B|A)$: Likelihood of B given A.
- $P(B)$: Evidence (normalization constant).
- $P(A|B)$: Posterior probability of A given B.

Use:

- Updates probabilities based on new evidence.
- Handles uncertainty in decision-making.

Example: Bayes' Theorem in Action

Problem: Compute the probability of a student passing (A) given they attended a review session (B).

Given:

- $P(A) = 0.4$
- $P(B|A) = 0.7$
- $P(B) = 0.5$

Solution:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.7 \cdot 0.4}{0.5} = 0.56$$

Conclusion: The probability that the student passes given they attended the review session is 56%.

Example: Probability of HIV

Problem: Calculate the probability of having HIV after a positive test result.

Given:

- HIV prevalence, $P(HIV) = 0.008$
- Test sensitivity, $P(\text{T}|\text{HIV}) = 0.95$
- Test specificity, $P(\neg\text{T}|\neg\text{HIV}) = 0.95$

Solution:

$$P(HIV|T) \propto P(T|HIV) \cdot P(HIV) = 0.95 \cdot 0.008 = 0.0076$$

$$P(\neg HIV|T) \propto P(T|\neg HIV) \cdot P(\neg HIV) = 0.05 \cdot 0.992 = 0.0496$$

Conclusion: Even with a positive result, the probability of having HIV is low due to the low prior probability.

Working of Naive Bayes

Key Idea:

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

where $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$.

Naive Assumption:

$$\begin{aligned} P(\mathbf{X}|C) &= \prod_{i=1}^{n} P(x_i|C) \\ &= P(x_1|C)P(x_2|C)\ldots P(x_n|C) \end{aligned}$$

Steps:

- Compute prior probabilities, $P(C)$.
- Compute likelihoods, $P(x_i|C)$.
- Combine them to find the posterior probability $P(C|X)$.

Mathematical Motivation Behind Naive Bayes

The Challenge: Joint Probability Distribution

- In classification, we need to compute:

$$P(A|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|A) \cdot P(A)}{P(x_1, x_2, \ldots, x_n)}.$$

- The core challenge is calculating $P(x_1, x_2, \ldots, x_n|A)$, the joint probability of the features given the class.
- For n features, this involves modeling a complex, high-dimensional distribution.

Why is it Hard?

- Requires a large amount of data to estimate joint probabilities accurately.
- Computationally expensive for high-dimensional feature spaces.
- Susceptible to overfitting when data is sparse.

The Naive Assumption Simplifies the Problem

The Naive Assumption:

$$P(x_1, x_2, \ldots, x_n|A) = \prod_{i=1}^{n} P(x_i|A).$$

- Assumes that all features $x_1, x_2, \ldots, x_n$ are conditionally independent given the class A.
- Breaks the joint probability into a product of simpler, one-dimensional probabilities.

Naive Bayes' Classification

Simplified Formula:

$$P(A|x_1, x_2, \ldots, x_n) \propto P(A) \cdot \prod_{i=1}^{n} P(x_i|A).$$

Why is the Naive Assumption Helpful?

1. Computational Efficiency:

- Joint probability $P(x_1, x_2, \ldots, x_n|A)$ requires 2^n parameters for binary features.
- With the naive assumption, only n parameters, $P(x_i|A)$ are needed.

2. Reduced Data Requirements:

- Estimating $P(x_1, x_2, \ldots, x_n|A)$ accurately requires a large dataset.
- By assuming independence, the number of parameters to estimate reduces significantly.
- Feasible to train the model even with limited data.

3. Robustness in High-Dimensional Spaces:

- Handles high-dimensional data well because it avoids explicitly modeling correlations between features.

Example: Why Independence Helps?

Problem: Classify emails as spam or not spam based on three features:

- x_1: Presence of the word "free".
- x_2: Presence of the word "win".
- x_3: Presence of the word "money".

Without Independence:

- Joint probability $P(x_1, x_2, x_3|\text{Spam})$ requires estimating $2^3 = 8$ probabilities.
- For n features, 2^n combinations need estimation.

With Independence:

$$P(x_1, x_2, x_3|\text{Spam}) = P(x_1|\text{Spam}) \cdot P(x_2|\text{Spam}) \cdot P(x_3|\text{Spam}).$$

- Only 3 probabilities, $P(x_1|\text{Spam})$, $P(x_2|\text{Spam})$, $P(x_3|\text{Spam})$ need estimation.

Impact of the Naive Assumption

Advantages:

- Scalability: Handles datasets with many features efficiently.
- Feasibility: Enables training on smaller datasets without overfitting.
- Simplicity: Reduces complexity in both training and inference.

Trade-offs:

- Independence assumption is not always valid.
- Performance may degrade if features are highly correlated.

Conclusion: The naive assumption makes Naive Bayes computationally efficient and practical, especially in high-dimensional or sparse data scenarios.

Example: Naive Bayes for Spam Detection

Problem: Classify an email as "Spam" or "Not Spam" based on the occurrence of the words "Free", "Money" and "Win."

Dataset: The dataset consists of 50 spam emails and 50 non-spam emails. The count of the presence of words "Free", "Win", and "Money" is given in the dataset.

Word	Spam	Non-spam
"Free"	30	5
"Win"	25	10
"Money"	20	5

Table: Keyword Frequencies in Emails

Example: Naive Bayes for Spam Detection

Prior Probabilities:

$$P(\text{S}) = \frac{\text{Total Spam}}{\text{Total Emails}} = \frac{50}{100} = 0.5$$

$$P(\neg\text{S}) = \frac{\text{Total Non-spam}}{\text{Total Emails}} = \frac{50}{100} = 0.5$$

Example: Naive Bayes for Spam Detection

Likelihood Probabilities:

$$P(\text{Free}|\text{S}) = \frac{\text{Spam emails with "Free"}}{\text{Total Spam}} = \frac{30}{50} = 0.6$$

$$P(\text{Win}|\text{S}) = \frac{\text{Spam emails with "Win"}}{\text{Total Spam}} = \frac{25}{50} = 0.5$$

$$P(\text{Money}|\text{S}) = \frac{\text{Spam emails with "Money"}}{\text{Total Spam}} = \frac{20}{50} = 0.4$$

Example: Naive Bayes for Spam Detection

Likelihood Probabilities:

$$P(\text{Free}|\neg\text{S}) = \frac{\text{Non-spam emails with "Free"}}{\text{Total Not Spam}} = \frac{5}{50} = 0.1$$

$$P(\text{Win}|\neg\text{S}) = \frac{\text{Non-spam emails with "Win"}}{\text{Total Non-spam}} = \frac{10}{50} = 0.2$$

$$P(\text{Money}|\neg\text{S}) = \frac{\text{Non-spam emails with "Money"}}{\text{Total Non-spam}} = \frac{5}{50} = 0.1$$

Example: Naive Bayes for Spam Detection

Posterior Calculation:

$$\begin{aligned} P(\text{Spam}|\text{Free, Win, Money}) &\propto P(\text{Free}|\text{S}) \cdot P(\text{Win}|\text{S}) \cdot P(\text{Money}|\text{S}) \cdot P(\text{S}) \\ &= 0.6 \cdot 0.5 \cdot 0.4 \cdot 0.5 \\ &= 0.06 \end{aligned}$$

$$\begin{aligned} P(\neg\text{S}|\text{Free, Win, Money}) &\propto P(\text{Free}|\neg\text{S}) \cdot P(\text{Win}|\neg\text{S}) \cdot P(\text{Money}|\neg\text{S}) \cdot P(\neg\text{S}) \\ &= 0.1 \cdot 0.2 \cdot 0.1 \cdot 0.5 \\ &= 0.001 \end{aligned}$$

Comparing the results, we can say that there is higher probability that the email is spam if all three words "Free", "Win" and "Money" are present in the email.

Example: Predicting Tennis Play

Dataset:

Sl. No.	Outlook	Temp.	Humidity	Windy	Play Tennis?
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Table: Weather Dataset for Tennis Prediction

Computing Joint Probabilities for Tennis Play

Problem: Predict whether to play tennis given:

- Outlook = Sunny
- Temperature = Hot
- Humidity = High
- Windy = False

Learning Phase: Computing Probabilities for Tennis Play

Step 1: Compute Class Priors

Class	Count	Prior Probability $P(\text{Class})$
Yes	9	$P(\text{Y}) = \frac{9}{14} = 0.64$
No	5	$P(\text{N}) = \frac{5}{14} = 0.35$

Table: Class Priors for the Tennis Dataset

Step 2: Compute Conditional Probabilities For each feature and class, calculate $P(\text{Feature}|\text{Class})$.

Conditional Probabilities: Outlook

Feature: Outlook

Outlook	Class: Yes	Class: No	Total
Sunny	2/9	3/5	5
Overcast	4/9	0/5	4
Rainy	3/9	2/5	5

Table: Conditional Probabilities for Outlook

Conditional Probabilities: Temperature

Feature: Temperature

Temperature	Class: Yes	Class: No	Total
Hot	2/9	2/5	4
Mild	4/9	2/5	6
Cool	3/9	1/5	4

Table: Conditional Probabilities for Temperature

Conditional Probabilities: Humidity and Windy

Feature: Humidity

Humidity	Class: Yes	Class: No	Total
High	3/9	4/5	7
Normal	6/9	1/5	7

Table: Conditional Probabilities for Humidity

Feature: Windy

Windy	Class: Yes	Class: No	Total
False	6/9	2/5	8
True	3/9	3/5	6

Table: Conditional Probabilities for Windy

Learning Phase Summary

Key Steps in the Learning Phase:

1. Compute prior probabilities for each class ($P(\text{Yes})$, $P(\text{No})$).
2. Compute conditional probabilities for each feature given the class ($P(\text{Feature}|\text{Class})$).
3. Store these probabilities as the learned parameters of the model.

Outcome:

- The model is now ready to predict outcomes for unseen data by combining these probabilities using Bayes' Rule.

Summarizing the Conditional Probabilities

Feature	$P(\textbf{Feature}\|\textbf{Y})$	$P(\textbf{Feature}\|\textbf{N})$
Outlook = S	0.22	0.6
Temperature = Hot	0.22	0.4
Humidity = High	0.33	0.8
Windy = F	0.66	0.4

Table: Likelihoods for Tennis Prediction

Final Decision

Step 4: Normalize and Compare

$$P(\text{Yes}|X) \propto P(\text{Yes}) \cdot P(\text{Sunny}|\text{Yes}) \cdot \ldots$$

$$P(\text{No}|X) \propto P(\text{No}) \cdot P(\text{Sunny}|\text{No}) \cdot \ldots$$

Classification Decision:

- Compute both posterior probabilities.
- Classify as the class with the higher posterior probability.

Result: Based on the computed values, the model predicts whether to play tennis.

Posterior Probabilities:

$$\begin{aligned} P(\text{Y}|\text{S, Hot, High, F}) &\propto P(\text{Y}) \cdot P(\text{S}|\text{Y}) \cdot P(\text{Hot}|\text{Y}) \cdot P(\text{High}|\text{Y}) \cdot P(\text{F}|\text{Y}) \\ &= (.64)(0.22)(0.22)(0.33)(0.66) \\ &= 0.007 \end{aligned}$$

$$\begin{aligned} P(\text{N}|\text{S, Hot, High, F}) &\propto P(\text{N}) \cdot P(\text{S}|\text{N}) \cdot P(\text{Hot}|\text{N}) \cdot P(\text{High}|\text{N}) \cdot P(\text{F}|\text{N}) \\ &= (.36)(0.6)(0.4)(0.8)(0.4) \\ &= 0.046 \end{aligned}$$

The class with the higher posterior is the prediction.

Playing Tennis has higher probability

Example: HIV Diagnosis

Problem: Calculate the probability that a person has HIV given a positive test result.

Given:

- $P(\text{HIV}) = 0.008$
- $P(\text{Positive}|\text{HIV}) = 0.95$
- $P(\text{Positive}|\neg\text{HIV}) = 0.05$
- $P(\neg\text{HIV}) = 0.992$

Step-by-Step: HIV Example

Using Bayes' Theorem:

$$P(\text{HIV}|\text{Positive}) = \frac{P(\text{Positive}|\text{HIV}) \cdot P(\text{HIV})}{P(\text{Positive})}$$

Where:

$$P(\text{Positive}) = P(\text{Positive}|\text{HIV}) \cdot P(\text{HIV}) + P(\text{Positive}|\neg\text{HIV}) \cdot P(\neg\text{HIV})$$

Calculation:

$$P(\text{Positive}) = (0.95 \cdot 0.008) + (0.05 \cdot 0.992) = 0.0486$$

$$P(\text{HIV}|\text{Positive}) = \frac{0.95 \cdot 0.008}{0.0486} \approx 0.156$$

Conclusion: Despite a positive result, the probability of having HIV is approximately 15.6% due to the low prevalence of the disease.

Example: HIV Diagnosis with Two Positive Results

Problem: Calculate the probability that a person has HIV after receiving two positive test results from independent tests.

Given:

- $P(\text{HIV}) = 0.008$: Prevalence of HIV in the population.
- $P(\text{Positive}|\text{HIV}) = 0.95$: Sensitivity (True Positive Rate).
- $P(\text{Negative}|\neg\text{HIV}) = 0.95$: Specificity (True Negative Rate).
- Tests are assumed to be conditionally independent given the status (HIV or not).

Step-by-Step Solution

Step 1: Define Key Probabilities

- $P(\neg \text{HIV}) = 1 - P(\text{HIV}) = 0.992$
- False Positive Rate:

  $P(\text{T}|\neg \text{HIV}) = 1 - P(\neg \text{T}|\neg \text{HIV}) = 0.05$

Step 2: Use Naive Bayes for Two Positive Tests

$$P(\text{HIV}|\text{T}_1, \text{T}_2) \propto P(\text{T}_1|\text{HIV}) \cdot P(\text{T}_2|\text{HIV}) \cdot P(\text{HIV})$$

$$P(\neg \text{HIV}|\text{T}_1, \text{T}_2) \propto P(\text{T}_1|\neg \text{HIV}) \cdot P(\text{T}_2|\neg \text{HIV}) \cdot P(\neg \text{HIV})$$

Step-by-Step Solution

$$P(\mathrm{T}_1) = P(\mathrm{T}_1|\mathrm{HIV})P(\mathrm{HIV}) + P(\neg\mathrm{T}_1|\neg\mathrm{HIV})P(\neg\mathrm{HIV})$$
$$P(\mathrm{T}_2) = P(\mathrm{T}_2|\mathrm{HIV})P(\mathrm{HIV}) + P(\neg\mathrm{T}_2|\neg\mathrm{HIV})P(\neg\mathrm{HIV})$$

Assuming, independence of the tests, computing the predictor prior probability (normalization factor),

$$P(\mathrm{T}_1, \mathrm{T}_2) = P(\mathrm{T}_1)P(\mathrm{T}_2)$$

Substitute the values to find the **posterior probabibility,**

$$P(\mathbf{HIV}|\mathbf{T}_1, \mathbf{T}_2) = \frac{P(\mathbf{HIV})P(\mathbf{T}_1|\mathbf{HIV})P(\mathbf{T}_2|\mathbf{HIV})}{P(\mathbf{T}_1, \mathbf{T}_2)}$$

Types of Naive Bayes Classifiers

- **Gaussian Naive Bayes:** - For continuous data. - Assumes features follow a Gaussian distribution:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- **Multinomial Naive Bayes:** - For count data (e.g., word counts). - Used in text classification.
- **Bernoulli Naive Bayes:** - For binary data (e.g., presence/absence of a word).

Advantages of Naive Bayes

- Simple and easy to implement.
- Computationally efficient:
  - Training: Linear in the number of features.
  - Testing: Fast due to precomputed probabilities.
- Handles categorical and continuous data.
- Works well with small datasets and high-dimensional data.

Limitations of Naive Bayes

- Relies on the assumption of feature independence.
- Fails when features are strongly correlated.
- Sensitive to zero probabilities (addressed with Laplace smoothing).
- Limited expressive power compared to more complex classifiers.

Questions?

Thank you for your attention!