Naive Bayes Algorithm: Simple Example

The Naive Bayes classifier uses Bayes' theorem assuming conditional independence of features.

Given a new observation with categorical features, we compute:

$$P(y \mid \mathbf{x}) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Where $y \in \{\mathrm{M}, \mathrm{H}\}$, and $\mathbf{x} = (Color, Legs, Height, Smelly)$

Dataset

Sl. No.	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

Frequency Table

Class counts:

$$P(M) = \frac{4}{8} = 0.5, \quad P(H) = \frac{4}{8} = 0.5$$

Conditional Frequencies

Feature	$P(\cdot \mid M)$	$P(\cdot \mid H)$
Color=White	$\frac{2}{4} = 0.5$	$\frac{3}{4} = 0.75$
Color=Green	$\frac{2}{4} = 0.5$	$\frac{1}{4} = 0.25$
Legs=2	$\frac{1}{4} = 0.25$	$\frac{4}{4} = 1.0$
Legs=3	$\frac{3}{4} = 0.75$	$\frac{0}{4} = 0.0$
Height=Short	$\frac{3}{4} = 0.75$	$\frac{1}{4} = 0.25$
Height=Tall	$\frac{1}{4} = 0.25$	$\frac{3}{4} = 0.75$
Smelly=Yes	$\frac{3}{4} = 0.75$	$\frac{1}{4} = 0.25$
Smelly=No	$\frac{1}{4} = 0.25$	$\frac{3}{4} = 0.75$

Classification Task

Given a new data point:

- Color = White
- Legs = 2
- Height = Short
- Smelly = Yes

We compute the posterior for each class.

Class M:

$$P(M \mid \mathbf{x}) \propto P(M) \cdot P(White \mid M) \cdot P(2 \text{ legs} \mid M) \cdot P(Short \mid M) \cdot P(Yes \mid M)$$

$$= 0.5 \cdot 0.5 \cdot 0.25 \cdot 0.75 \cdot 0.75 = 0.5 \cdot 0.0703 = 0.03516$$

Class H:

$$P(H \mid \mathbf{x}) \propto P(H) \cdot P(White \mid H) \cdot P(2 \text{ legs} \mid H) \cdot P(Short \mid H) \cdot P(Yes \mid H)$$
$$= 0.5 \cdot 0.75 \cdot 1.0 \cdot 0.25 \cdot 0.25 = 0.5 \cdot 0.046875 = 0.046875$$

Normalize

$$P(M \mid \mathbf{x}) = \frac{0.03516}{0.03516 + 0.046875} \approx \frac{0.03516}{0.08203} \approx 0.4286$$
$$P(H \mid \mathbf{x}) \approx 1 - 0.4286 = 0.5714$$

Conclusion

The predicted class is **H (Species H)**, with posterior probability ≈ 0.57.

Naive Bayes: Full Calculation Example

Let's consider a very small dataset of email messages, classified as either "Spam" (S) or "Not Spam" (NS). We will analyze the presence (1) or absence (0) of three key words: "money," "free," and "Nigeria."

Dataset: Email Spam Classification

Our dataset consists of 5 emails, each labeled as Spam (S) or Not Spam (NS).

Email ID	money (Feature 1)	free (Feature 2)	Nigeria (Feature 3)	Class (Label)
1	1	1	0	S
2	0	1	1	S
3	1	0	0	NS
4	0	0	0	NS
5	0	1	0	NS

Goal: Given a new email with the features `money=1, free=0, Nigeria=0`, classify it as Spam (S) or Not Spam (NS) using Naive Bayes.

I. Bernoulli Naive Bayes

Bernoulli Naive Bayes is suitable when features are binary (0/1) and represent presence/absence. It models the probability of a feature being present or absent given the class.

Equations:

The Naive Bayes classifier predicts the class $\hat{y}$ for a new data point $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ by maximizing the posterior probability:

$$\hat{y} = \arg\max_{y \in Y} P(y|x_1, \ldots, x_n)$$

Using Bayes' Theorem:

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|y)P(y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant for all classes, we only need to maximize the numerator:

$$\hat{y} = \arg\max_{y \in Y} P(x_1, \ldots, x_n|y)P(y)$$

The "Naive" assumption of conditional independence of features given the class:

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

Substituting this into the equation:

$$\hat{y} = \arg\max_{y \in Y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

For **Bernoulli Naive Bayes**, for each feature x_i and class y, we need to calculate:

- $P(x_i = 1|y)$: Probability that feature i is present given class y.
- $P(x_i = 0|y)$: Probability that feature i is absent given class y.

If $x_i = 1$: we use $P(x_i = 1|y)$
If $x_i = 0$: we use $P(x_i = 0|y) = 1 - P(x_i = 1|y)$

To avoid zero probabilities, we use **Laplace Smoothing (add-one smoothing)**. For a feature x_i and class y:

$$P(x_i = k|y) = \frac{\text{count}(x_i = k, y) + \alpha}{\text{count}(y) + \alpha \cdot K_i}$$

where $k \in \{0, 1\}$, $\alpha = 1$ for Laplace smoothing, and $K_i = 2$ (number of possible values for a binary feature).

Calculations for Bernoulli Naive Bayes:

Step 1: Calculate Prior Probabilities, $P(y)$

- Total Emails = 5
- Number of Spam (S) emails = 2
- Number of Not Spam (NS) emails = 3

$$P(S) = \frac{\text{count}(S)}{\text{total emails}} = \frac{2}{5} = 0.4$$
$$P(NS) = \frac{\text{count}(NS)}{\text{total emails}} = \frac{3}{5} = 0.6$$

Step 2: Calculate Conditional Probabilities, $P(x_i|y)$ **with Laplace Smoothing** ($\alpha = 1$)
For Class = Spam (S): Total Spam emails = 2. Denominator for conditional probabilities: $\text{count}(S) + \alpha \cdot 2 = 2 + 1 \cdot 2 = 4$.

- **Feature: money**
  - $\text{count}(\text{money} = 1, S) = 1$
  - $\text{count}(\text{money} = 0, S) = 1$

$$P(\text{money} = 1|S) = \frac{1+1}{2+2} = \frac{2}{4} = 0.5$$
$$P(\text{money} = 0|S) = \frac{1+1}{2+2} = \frac{2}{4} = 0.5$$

- **Feature: free**
  - $\text{count}(\text{free} = 1, S) = 2$
  - $\text{count}(\text{free} = 0, S) = 0$

$$P(\text{free} = 1|S) = \frac{2+1}{2+2} = \frac{3}{4} = 0.75$$
$$P(\text{free} = 0|S) = \frac{0+1}{2+2} = \frac{1}{4} = 0.25$$

- **Feature: Nigeria**
  - $\text{count}(\text{Nigeria} = 1, S) = 1$
  - $\text{count}(\text{Nigeria} = 0, S) = 1$

$$P(\text{Nigeria} = 1|S) = \frac{1+1}{2+2} = \frac{2}{4} = 0.5$$
$$P(\text{Nigeria} = 0|S) = \frac{1+1}{2+2} = \frac{2}{4} = 0.5$$

For Class = Not Spam (NS): Total Not Spam emails = 3. Denominator for conditional probabilities: $\text{count}(NS) + \alpha \cdot 2 = 3 + 1 \cdot 2 = 5$.

- **Feature: money**
  - $\text{count}(\text{money} = 1, NS) = 1$
  - $\text{count}(\text{money} = 0, NS) = 2$

$$P(\text{money} = 1|NS) = \frac{1+1}{3+2} = \frac{2}{5} = 0.4$$
$$P(\text{money} = 0|NS) = \frac{2+1}{3+2} = \frac{3}{5} = 0.6$$

- **Feature: free**

- $\text{count}(\text{free} = 1, NS) = 1$
- $\text{count}(\text{free} = 0, NS) = 2$

$$P(\text{free} = 1|NS) = \frac{1+1}{3+2} = \frac{2}{5} = 0.4$$

$$P(\text{free} = 0|NS) = \frac{2+1}{3+2} = \frac{3}{5} = 0.6$$

- **Feature: Nigeria**
  - $\text{count}(\text{Nigeria} = 1, NS) = 0$
  - $\text{count}(\text{Nigeria} = 0, NS) = 3$

$$P(\text{Nigeria} = 1|NS) = \frac{0+1}{3+2} = \frac{1}{5} = 0.2$$

$$P(\text{Nigeria} = 0|NS) = \frac{3+1}{3+2} = \frac{4}{5} = 0.8$$

Step 3: Classify New Email: `money=1, free=0, Nigeria=0`
For Class = Spam (S):

$$P(\text{new email}|S)P(S) = P(\text{money} = 1|S) \cdot P(\text{free} = 0|S) \cdot P(\text{Nigeria} = 0|S) \cdot P(S)$$

$$= 0.5 \cdot 0.25 \cdot 0.5 \cdot 0.4$$

$$= 0.0625 \cdot 0.4 = 0.025$$

For Class = Not Spam (NS):

$$P(\text{new email}|NS)P(NS) = P(\text{money} = 1|NS) \cdot P(\text{free} = 0|NS) \cdot P(\text{Nigeria} = 0|NS) \cdot P(NS)$$

$$= 0.4 \cdot 0.6 \cdot 0.8 \cdot 0.6$$

$$= 0.192 \cdot 0.6 = 0.1152$$

Step 4: Comparison Since $0.1152 > 0.025$, the new email is classified as **Not Spam (NS)**.

II. Multinomial Naive Bayes

Multinomial Naive Bayes is typically used for discrete features representing counts, such as word counts in a document. This model is more appropriate when a feature can appear multiple times.

Equations:

The core classification equation remains the same:

$$\hat{y} = \arg\max_{y \in Y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

For **Multinomial Naive Bayes**, x_i represents the count of feature i in a document. The conditional probability $P(x_i|y)$ is modeled using a multinomial distribution.

$$P(x_i|y) = \frac{\text{count}(x_i \text{ in documents of class } y) + \alpha}{\sum_{k=1}^{V}(\text{count}(x_k \text{ in documents of class } y) + \alpha)}$$

where V is the total number of unique features (vocabulary size), and $\alpha = 1$ for Laplace (add-one) smoothing. The denominator is effectively the total number of word occurrences in documents of class y plus $\alpha \cdot V$.

Let's re-interpret our features as counts. In this simple dataset, a '1' means the word appeared once, and '0' means it appeared zero times.

Calculations for Multinomial Naive Bayes:

Step 1: Calculate Prior Probabilities, $P(y)$ (Same as Bernoulli Naive Bayes)

$$P(S) = 0.4$$

$$P(NS) = 0.6$$

Step 2: Calculate Conditional Probabilities, $P(x_i|y)$ **with Laplace Smoothing** ($\alpha = 1$)
Vocabulary size $V = 3$ (money, free, Nigeria).
For Class = Spam (S):

- Total word occurrences in Spam documents:
  - Email 1 (S): money=1, free=1, Nigeria=0 → 2 words
  - Email 2 (S): money=0, free=1, Nigeria=1 → 2 words
  - Total count of all words in Spam emails = 2 + 2 = 4
- Denominator for conditional probabilities: Total word occurrences in S + $\alpha \cdot V = 4 + 1 \cdot 3 = 7$.
- **Feature: money**
- count(money in S) = 1

$$P(\text{money}|S) = \frac{1+1}{4+3} = \frac{2}{7} \approx 0.2857$$

- **Feature: free**
- count(free in S) = 2

$$P(\text{free}|S) = \frac{2+1}{4+3} = \frac{3}{7} \approx 0.4286$$

- **Feature: Nigeria**
- count(Nigeria in S) = 1

$$P(\text{Nigeria}|S) = \frac{1+1}{4+3} = \frac{2}{7} \approx 0.2857$$

For Class = Not Spam (NS):

- Total word occurrences in Not Spam documents:
  - Email 3 (NS): money=1, free=0, Nigeria=0 → 1 word
  - Email 4 (NS): money=0, free=0, Nigeria=0 → 0 words
  - Email 5 (NS): money=0, free=1, Nigeria=0 → 1 word
  - Total count of all words in Not Spam emails = 1 + 0 + 1 = 2
- Denominator for conditional probabilities: Total word occurrences in NS + $\alpha \cdot V = 2 + 1 \cdot 3 = 5$.
- **Feature: money**
- count(money in NS) = 1

$$P(\text{money}|NS) = \frac{1+1}{2+3} = \frac{2}{5} = 0.4$$

- **Feature: free**
- count(free in NS) = 1

$$P(\text{free}|NS) = \frac{1+1}{2+3} = \frac{2}{5} = 0.4$$

- **Feature: Nigeria**
- count(Nigeria in NS) = 0

$$P(\text{Nigeria}|NS) = \frac{0+1}{2+3} = \frac{1}{5} = 0.2$$

Step 3: Classify New Email: `money=1, free=0, Nigeria=0` For Multinomial Naive Bayes, x_i in the product refers to the *count* of feature i in the new email. So, for the new email:

- `money` appears 1 time ($x_{\text{money}} = 1$)
- `free` appears 0 times ($x_{\text{free}} = 0$)
- `Nigeria` appears 0 times ($x_{\text{Nigeria}} = 0$)

For Class = Spam (S):

$$\begin{aligned} P(\text{new email}|S)P(S) &= P(\text{money}|S)^{x_{\text{money}}} \cdot P(\text{free}|S)^{x_{\text{free}}} \cdot P(\text{Nigeria}|S)^{x_{\text{Nigeria}}} \cdot P(S) \\ &= (0.2857)^1 \cdot (0.4286)^0 \cdot (0.2857)^0 \cdot 0.4 \\ &= 0.2857 \cdot 1 \cdot 1 \cdot 0.4 \approx 0.11428 \end{aligned}$$

For Class = Not Spam (NS):

$$\begin{aligned} P(\text{new email}|NS)P(NS) &= P(\text{money}|NS)^{x_{\text{money}}} \cdot P(\text{free}|NS)^{x_{\text{free}}} \cdot P(\text{Nigeria}|NS)^{x_{\text{Nigeria}}} \cdot P(NS) \\ &= (0.4)^1 \cdot (0.4)^0 \cdot (0.2)^0 \cdot 0.6 \\ &= 0.4 \cdot 1 \cdot 1 \cdot 0.6 = 0.24 \end{aligned}$$

Step 4: Comparison Since $0.24 > 0.11428$, the new email is classified as **Not Spam (NS)**.

III. Gaussian Naive Bayes

Gaussian Naive Bayes (GNB) is used when features are continuous and assumed to follow a normal (Gaussian) distribution for each class. For each feature x_i and class y, we estimate the mean $\mu_{i,y}$ and variance $\sigma^2_{i,y}$ from the training data:

$$\mu_{i,y} = \frac{1}{N_y} \sum_{j:y_j=y} x_i^{(j)}, \qquad \sigma^2_{i,y} = \frac{1}{N_y} \sum_{j:y_j=y} \left(x_i^{(j)} - \mu_{i,y}\right)^2$$

Given a new sample $\mathbf{x} = (x_1, \ldots, x_n)$, the class likelihood is computed as:

$$P(\mathbf{x} \mid y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\,\sigma^2_{i,y}}} \exp\Big(-\frac{(x_i - \mu_{i,y})^2}{2\sigma^2_{i,y}}\Big)$$

The posterior probability is then:

$$P(y \mid \mathbf{x}) \propto P(y)\, P(\mathbf{x} \mid y)$$

Application of GNB in Conceptual Example

Although our features ("money", "free", "Nigeria") are binary, one can imagine instead that each feature represents the **frequency count** of the word in an email—making them continuous.

For illustrative purposes:

- Suppose for Spam (S), we estimate

$$\mu_{\text{money},S} = 0.8,\ \sigma_{\text{money},S} = 0.4$$

- For Not-Spam (NS),

$$\mu_{\text{money},NS} = 0.3,\ \sigma_{\text{money},NS} = 0.2$$

Given a new email where "money" occurs once ($x_{\text{money}} = 1$), the likelihood term for this feature under each class is:

$$P(x_{\text{money}} = 1 \mid S) = \frac{1}{\sqrt{2\pi \cdot 0.4^2}} \exp\left(-\frac{(1-0.8)^2}{2 \cdot 0.4^2}\right)$$

$$P(x_{\text{money}} = 1 \mid NS) = \frac{1}{\sqrt{2\pi \cdot 0.2^2}} \exp\left(-\frac{(1-0.3)^2}{2 \cdot 0.2^2}\right)$$

You would similarly compute likelihoods for the "free" and "Nigeria" features (using their μ and σ), multiply them together with the class prior $P(y)$, and then select the class with the highest posterior probability.

Discussion:

- **Differences in Output:** Both Bernoulli and Multinomial Naive Bayes gave the same classification for this specific example, but the intermediate probabilities differed. This highlights the different underlying generative models for the features. Bernoulli is presence/absence, while Multinomial is counts.
- **The "Naive" Assumption:** Emphasize how we completely ignored the potential correlations between "money" and "free" (e.g., "free money" is a common spam phrase). This is the "naive" part. Discuss scenarios where this assumption might severely limit performance.
- **Log-Probabilities for Stability:** In real-world implementations, especially with many features, the product of probabilities can become extremely small, leading to underflow. Graduate students should be aware that calculations are typically done in the log-domain to avoid this:

$$\log\left(P(y)\prod_{i=1}^{n}P(x_i|y)\right) = \log P(y) + \sum_{i=1}^{n}\log P(x_i|y)$$

- **Laplace Smoothing Impact:** Discuss the role of α (hyperparameter) in Laplace smoothing. A larger α smooths the probabilities more aggressively, reducing the impact of rare events but potentially biasing the model if α is too large.
- **Choosing the Right Model:** When would you choose Bernoulli vs. Multinomial Naive Bayes for text classification?
  - **Bernoulli:** When you care about the *presence or absence* of words, not their frequency. Good for shorter texts or when rare words are highly indicative.
  - **Multinomial:** When *word frequency* matters. Generally performs better on longer documents where word counts provide more information.
- **Gaussian Naive Bayes:** Briefly mention that for continuous features, Gaussian Naive Bayes assumes features are normally distributed given the class. The parameters (μ, σ^2) are estimated from the training data.

Prepared By:
Md. Atikuzzaman
Lecturer
Department of Computer Science and Engineering
Green University of Bangladesh
Email: atik@cse.green.edu.bd