Naive Bayes

What Is Statistical Modeling?

- Uses **all attributes simultaneously**.
- Assumes attributes contribute **equally** and **independently** to the class.
- Surprisingly effective despite its unrealistic assumptions.
- Leads to the classic **Naive Bayes classifier**.

Why Use Statistical Modeling?

- Fast and scalable to large datasets.
- Robust to irrelevant features.
- Works well with limited training data.
- Serves as a **strong baseline model**.

Foundations: Bayes' Rule

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

Where:

- H = hypothesis (e.g., class = yes)
- E = evidence (attribute values)
- $P(H)$ = prior probability
- $P(E \mid H)$ = likelihood

Naive Bayes Assumption

- Attributes are assumed to be **independent given the class**.

$$P(H|E_1, E_2, \ldots, E_k) \propto P(H) \prod_{i=1}^{k} P(E_i|H)$$

- This creates a simple, multiplicative probability model.

Naive Bayes Process

1. Count frequencies per class
2. Compute P(E|Class)
3. Compute P(Class)
4. Multiply probabilities for new instance
5. Select class with highest posterior

Example: Weather Dataset

- Predict **Play = yes/no**.
- Attributes used:
  - Outlook
  - Temperature
  - Humidity
  - Windy

Numerical Example (Simplified)

- For a new day:
  - Outlook = sunny
  - Temperature = cool
  - Humidity = high
  - Windy = true
- Compute: P(yes) * P(sunny|yes) * P(cool|yes) * P(high|yes) * P(true|yes)
- Compute same for P(no)
- Choose larger value

Zero-Probability Problem

- If a value never occurs with a class:

$$P(E \mid H) = 0$$

- This zeroes out the entire probability product.

Solution: Laplace Estimator

- Add **1** to all frequency counts:

$$P(E \mid H) = \frac{count(E, H) + 1}{count(H) + k}$$

- k = number of possible values.
- Prevents zero probabilities.

Numeric Attributes

- Naive Bayes uses a **Gaussian distribution**:

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

- Compute mean μ and standard deviation σ per class.

Numeric Naive Bayes

1. Compute mean
2. Compute standard deviation
3. Use Gaussian density as $P(E \mid Class)$

Missing Values

- If value missing in test instance:
  - Skip attribute and renormalize.
- If missing in training:
  - Estimate probabilities using available values only.

Example: Document Classification

- Two variants:
    - Bernoulli NB — presence/absence
    - Multinomial NB — word frequencies
- Used in:
    - Spam filtering
    - Topic classification
    - Sentiment analysis

Example: Multinomial

- For word counts:

$$P(E \mid H) = N! \prod_{i=1}^{k} \frac{P_i^{n_j}}{n_i!}$$

- Handles word frequency information.

Log Probabilities

- To avoid underflow:

$$\log(P) = \log(P(H)) + \sum_{i=1}^{k} \log(P(E_i \mid H))$$

- Prevents numeric issues
- Speeds computation

Strengths of Naive Bayes

- Extremely fast
- Works well with high-dimensional data
- Strong baseline model
- Interpretable probability model
- Handles missing values naturally

Weaknesses of Naive Bayes

- Assumes independence
- Sensitive to redundant features
- Gaussian assumption not always accurate
- Struggles with correlated variables

When To Use Naive Bayes

- Best for:
    - Text classification
    - Medical/sensor data
    - High-dimensional datasets
- Not ideal for:
    - Highly correlated attributes
    - Complex feature interactions

Summary

- Statistical modeling -> Naive Bayes
- Uses conditional independence
- Handles numeric and nominal attributes
- Laplace correction prevents zero errors
- Widely used due to speed and robustness