Na”ıve Bayes: Theory and Practice

September 9, 2025

1 Introduction

Na”ıve Bayes (NB) is a family of simple yet effective probabilistic classifiers. Under the *conditional independence* assumption, the posterior is

$$p(y \mid \mathbf{x}) \propto p(y) \prod_{j=1}^{d} p(x_j \mid y), \tag{1}$$

where y is the class and $\mathbf{x} = (x_1, \ldots, x_d)$ are features. Despite the strong assumption, NB can work surprisingly well in many domains, especially with high-dimensional sparse inputs.

2 Theory and Formulas

Consider the Gaussian NB model for continuous features. For class $c \in \{1, \ldots, C\}$, assume

$$x_j \mid y = c \;\sim\; \mathcal{N}(\mu_{c,j},\, \sigma^2_{c,j}), \quad j = 1, \ldots, d. \tag{2}$$

Then the class-conditional density factorizes: $p(\mathbf{x} \mid y = c) = \prod_j \mathcal{N}(x_j\,;\, \mu_{c,j}, \sigma^2_{c,j})$. With prior $p(y = c)$, the (unnormalized) log-posterior is

$$\begin{aligned} \log p(y = c \mid \mathbf{x}) &\propto \log p(y = c) + \sum_{j=1}^{d} \log \mathcal{N}(x_j\,;\, \mu_{c,j}, \sigma^2_{c,j}) && (3) \\ &\propto \log p(y = c) - \sum_{j=1}^{d} \left[\tfrac{1}{2} \log(2\pi\sigma^2_{c,j}) + \tfrac{(x_j - \mu_{c,j})^2}{2\sigma^2_{c,j}} \right]. && (4) \end{aligned}$$

The predicted class is $\hat{y} = \arg\max_c \log p(y = c \mid \mathbf{x})$. Estimation is straightforward via sample means and variances within each class.

Notes. NB variants include Gaussian NB for continuous features and Multinomial/Bernoulli NB for count/binary features with Laplace (additive) smoothing. Calibration may be needed if probabilities are used downstream.

3 Applications and Tips

- **When it works:** high-dimensional sparse text features (bag-of-words), simple sensor data, baseline models.

- **Preprocessing:** standardize continuous features for Gaussian NB; for text, TF-IDF or raw counts for Multinomial NB.
- **Class priors:** either empirical (class frequencies) or domain-informed.
- **Independence assumption:** correlations between features may harm performance; use as a baseline and compare.
- **Evaluation:** compare with logistic regression/SVMs; use cross-validation.

4 Python Practice

The script below generates figures for Gaussian NB decision boundaries and simple diagnostics. Run it in the chapter folder; it saves images under `figures/`.

Listing 1: Generate Naive Bayes figures

```
# Terminal
python gen_naive_bayes_figures.py
```

5 Result

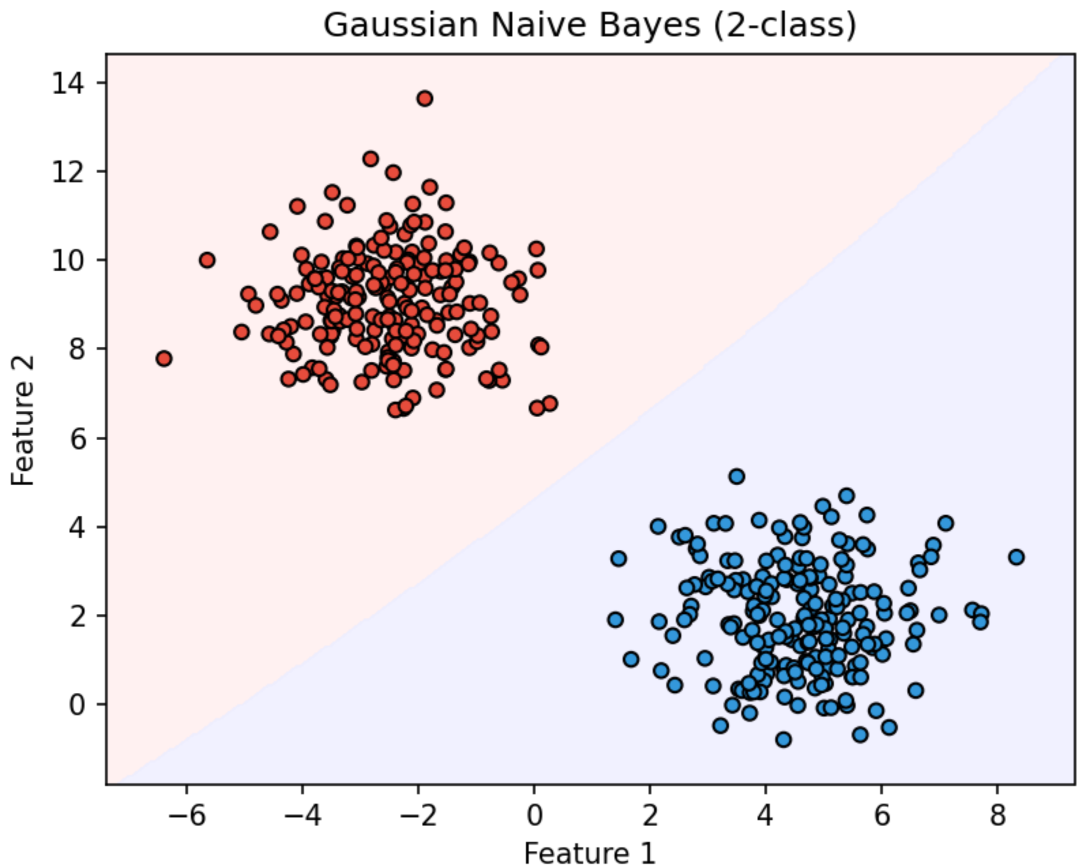

Figure 1: Gaussian NB decision boundary (2-class).

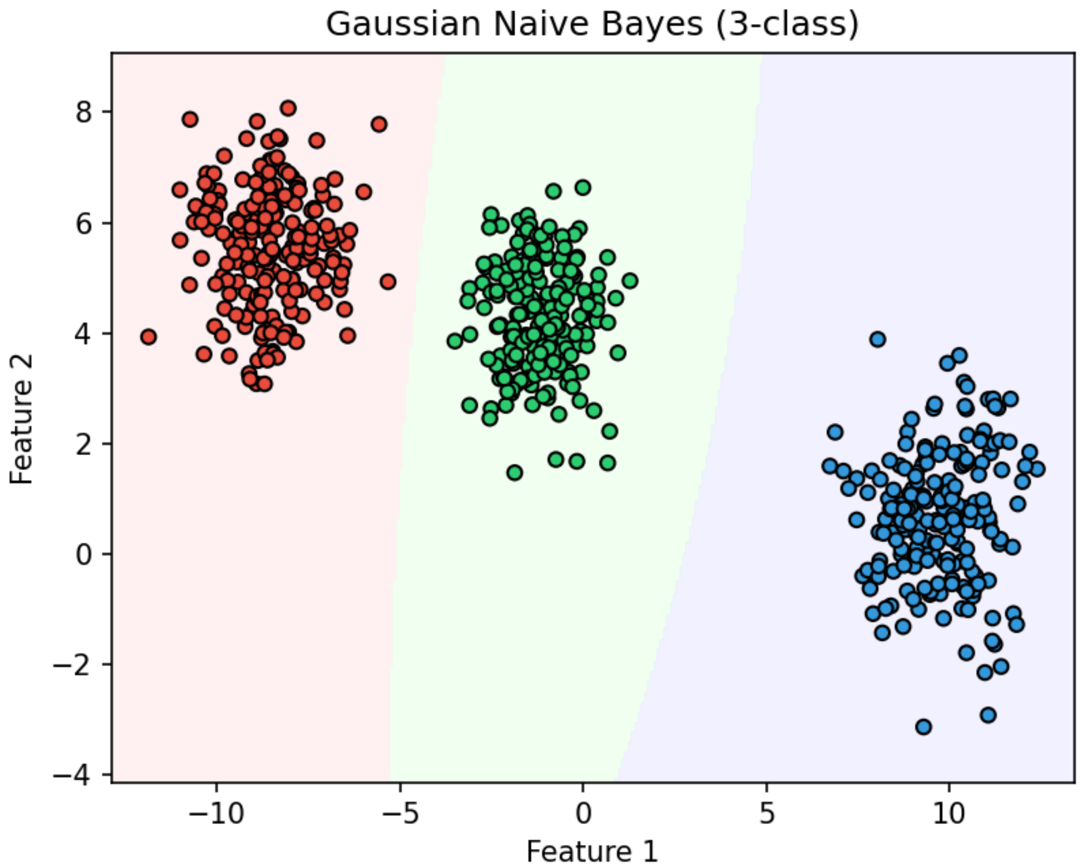

Figure 2: Gaussian NB decision regions (3-class).

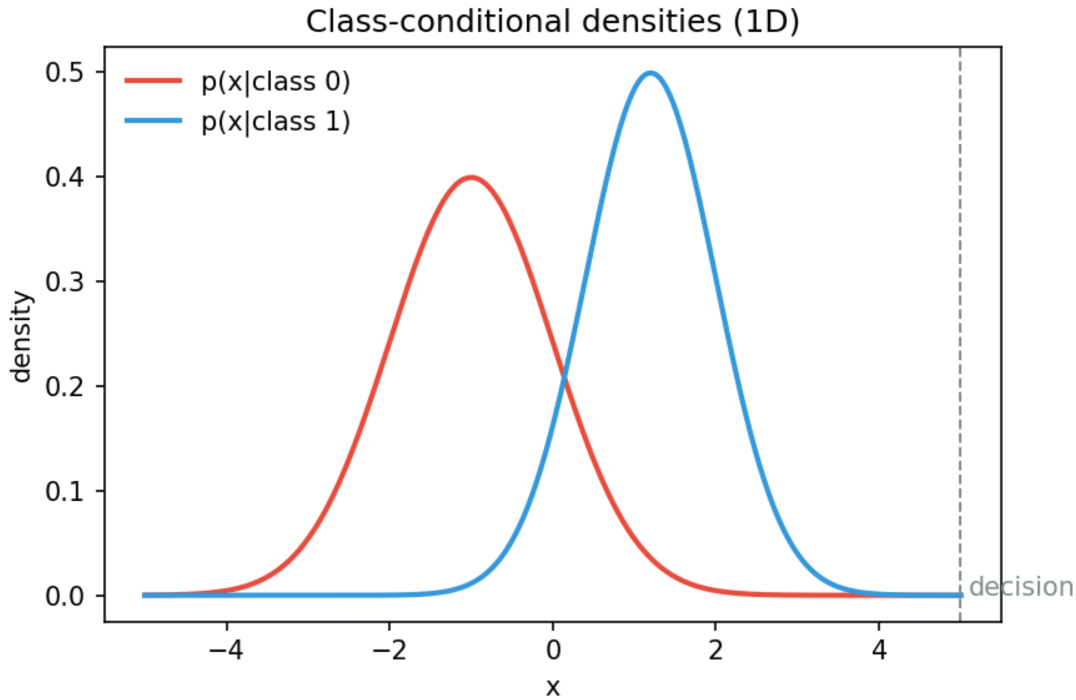

Figure 3: Class-conditional densities in 1D and a decision threshold.

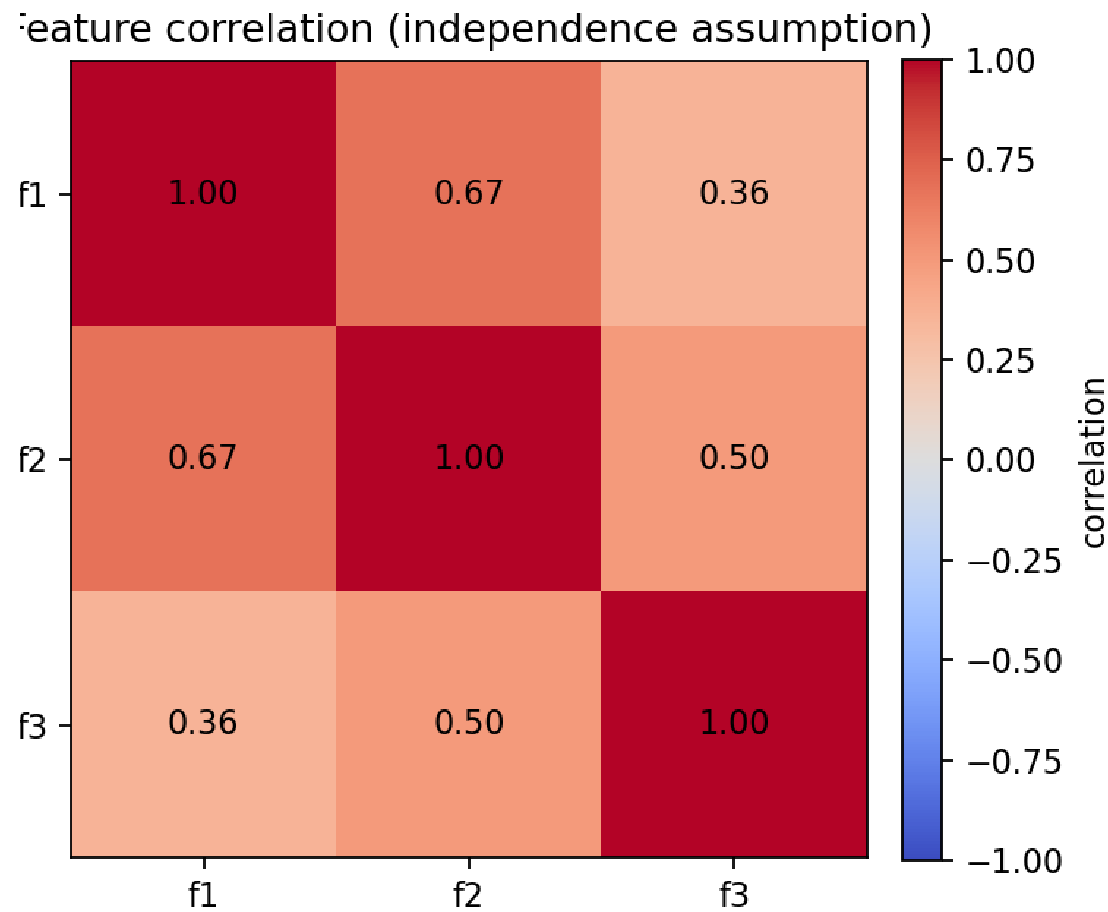

Figure 4: Feature correlation heatmap (independence assumption illustration).

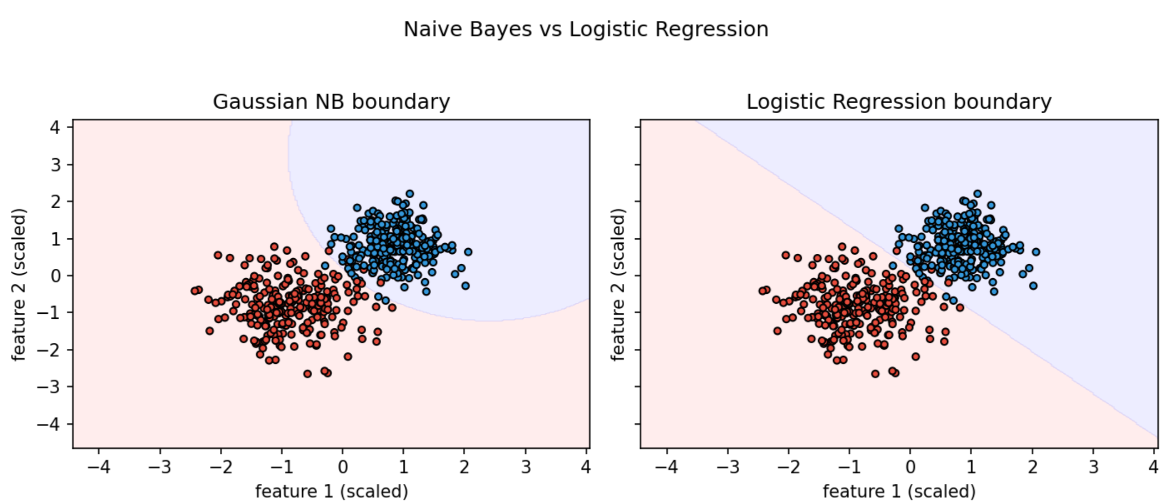

Figure 5: Decision boundary comparison: Gaussian NB vs Logistic Regression.

6 Summary

Na”ıve Bayes offers a fast, interpretable baseline. Its core idea is simple—combine class priors with per-feature likelihoods under conditional independence. While the assumption is often violated, NB remains competitive on certain problems and serves as a strong baseline against more flexible discriminative models.