朴素贝叶斯：理论与实践

2025 年 9 月 9 日

目录

1　引言

朴素贝叶斯（Naïve Bayes，NB）是在**条件独立**假设下建立的概率分类器族：

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

其中 y 为类别，$\mathbf{x} = (x_1, \ldots, x_d)$ 为特征。尽管独立性假设较强，NB 在高维稀疏特征（如文本）等任务中常表现稳健，且训练、预测效率较高。

2　原理与公式

以高斯朴素贝叶斯（Gaussian NB）为例，对于连续特征，假设对每个类别 $c \in \{1, \ldots, C\}$ 与每个特征 j 有：

$$x_j \mid y = c \ \sim\ \mathcal{N}(\mu_{c,j},\, \sigma_{c,j}^2). \tag{2}$$

则条件似然分解为 $p(\mathbf{x} \mid y = c) = \prod_j \mathcal{N}(x_j\,;\, \mu_{c,j}, \sigma_{c,j}^2)$。结合先验 $p(y = c)$，(未归一化的) 对数后验为：

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

预测类别为 $\hat{y} = \arg\max_c \log p(y = c \mid \mathbf{x})$。参数估计可由各类别内样本的均值与方差直接得到。

备注 变体包括：连续特征的 Gaussian NB；计数/二值特征的 Multinomial/Bernoulli NB（常配合拉普拉斯平滑）。若需使用概率值做后续决策，建议做概率校准。

3 应用场景与要点

- **适用场景**：高维稀疏文本（BOW/TF-IDF）、简单传感器数据、作为强基线。
- **预处理**：Gaussian NB 建议对连续特征做标准化；文本常用计数或 TF-IDF（Multinomial NB）。
- **类别先验**：可用经验频率或领域知识设定。
- **独立性假设**：特征强相关时性能可能下降；建议与逻辑回归/线性 SVM 等对比。
- **评估**：采用交叉验证对比不同模型与超参数。

4 Python 实战

在章节目录内运行下述脚本，图片将保存到本目录下的 `figures/`：

Listing 1: 生成朴素贝叶斯配图

```
# 在 4_Naive Bayes 目录中执行：
python gen_naive_bayes_figures.py
```

5 结果

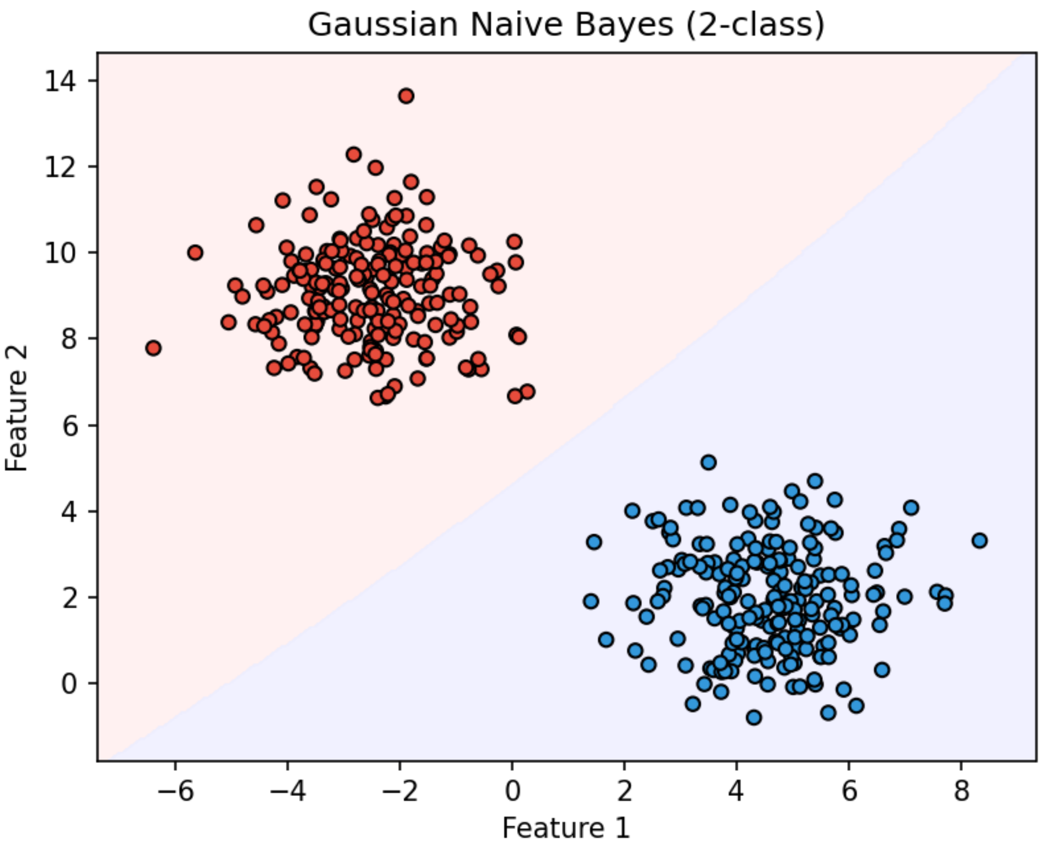

图 1: Gaussian NB 分类边界（两类）。

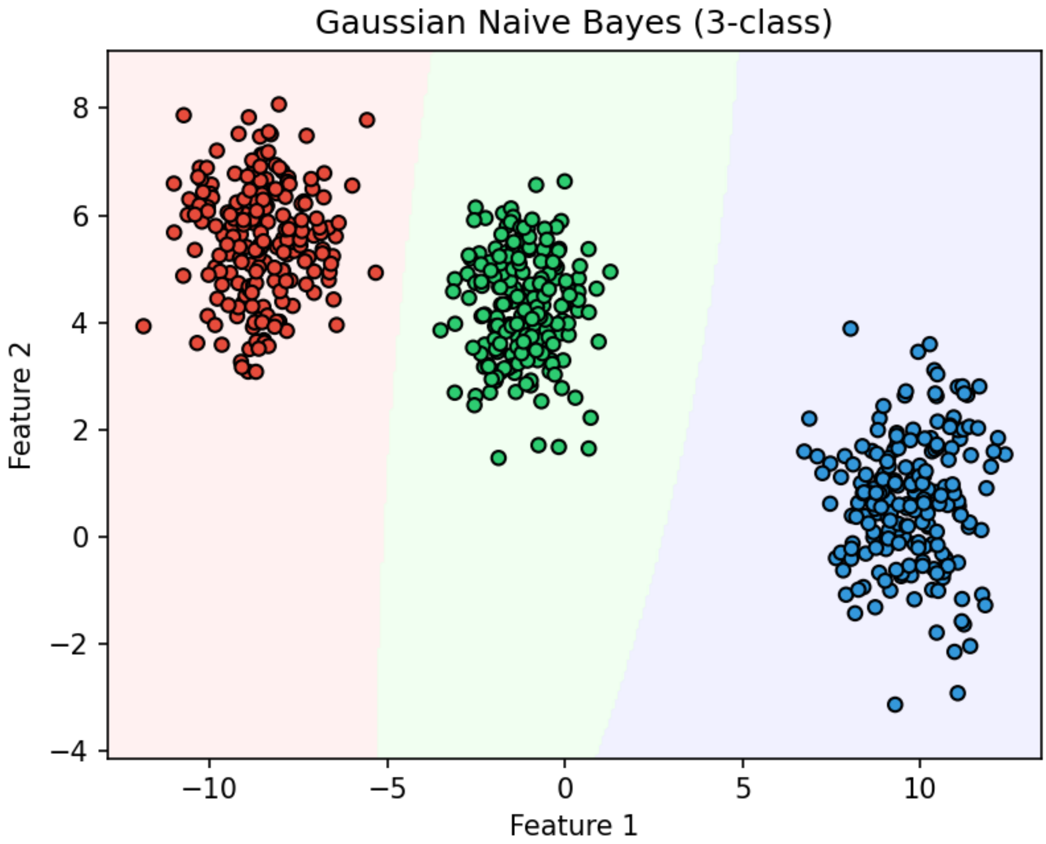

图 2: Gaussian NB 决策区域（三类）。

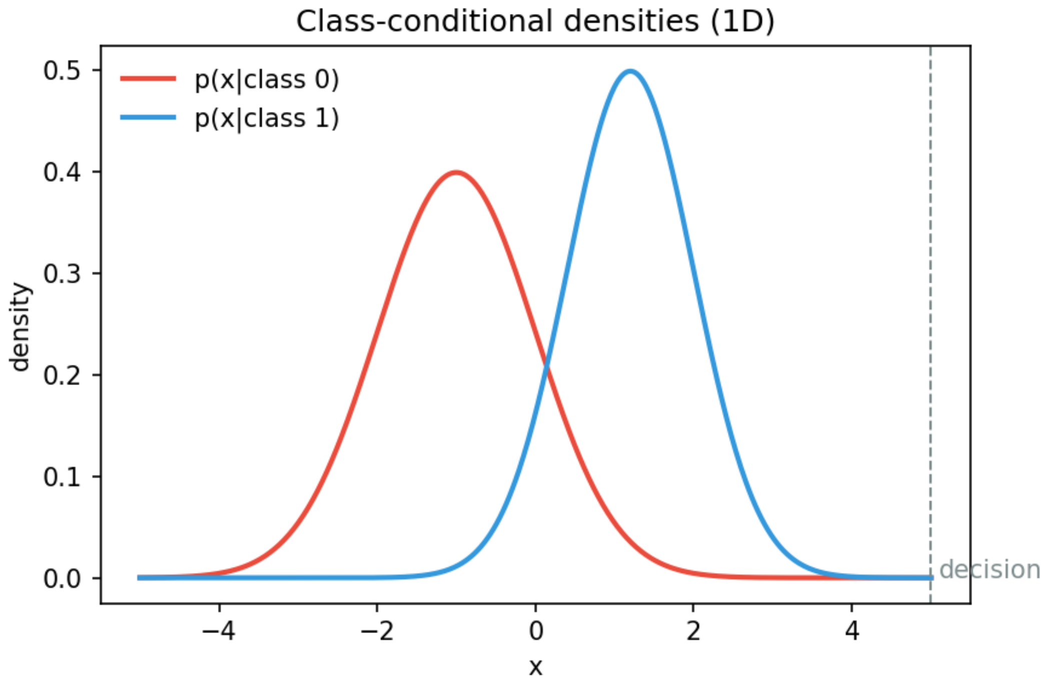

图 3: 一维类别条件密度与决策阈值。

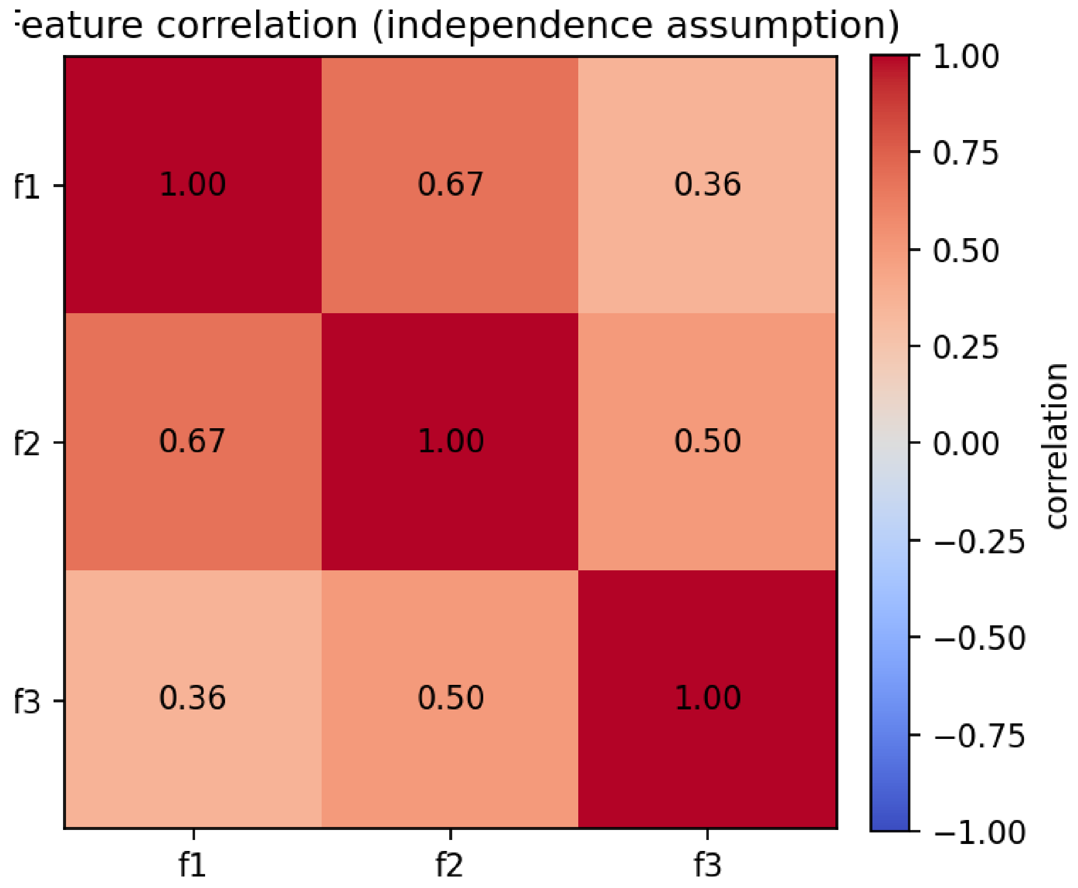

图 4: 特征相关性热力图（独立性假设示意）。

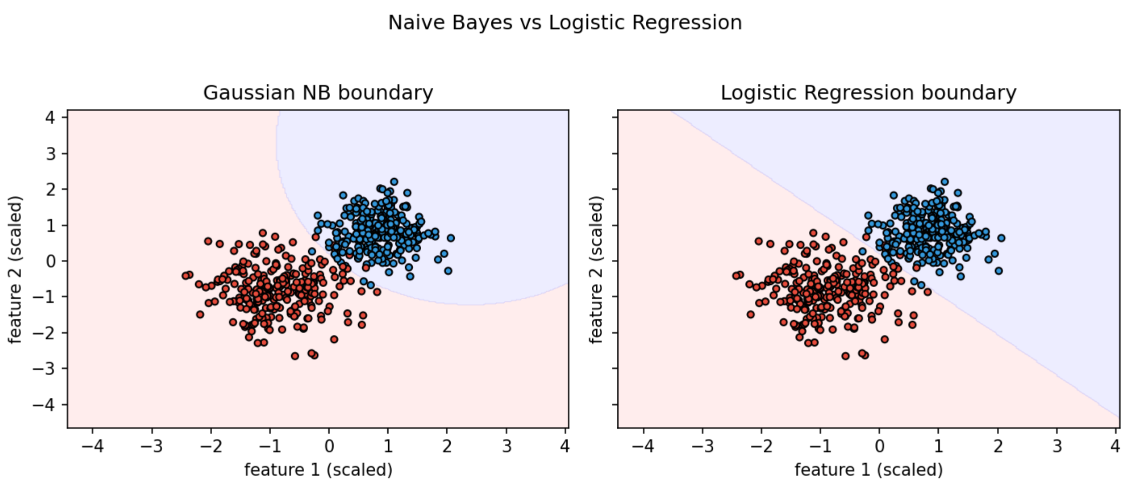

图 5: Gaussian NB 与逻辑回归的决策边界对比。

6 总结

朴素贝叶斯以简洁可解释、训练与预测高效为特点：核心是先验与逐特征似然的乘积（条件独立）。虽然假设并非总成立，它依然是可靠的基线模型，常用于与更强的判别式模型进行对比。