

Gaussian Naive Bayes Classifier

1. Asumsi Dasar

Gaussian Naive Bayes (GNB) merupakan model klasifikasi generatif yang didasarkan pada Teorema Bayes dan asumsi **independensi bersyarat** antar fitur. Untuk setiap kelas $c \in C$, fitur $x_j \mid y = c$ diasumsikan mengikuti distribusi Gaussian:

$$p(x_j \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_{cj}^2}} \exp\left(-\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2}\right)$$

Dengan:

- μ_{cj} : rata-rata fitur ke- j untuk kelas c
- σ_{cj}^2 : variansi fitur ke- j untuk kelas c

2. Estimasi Parameter

Diberikan data pelatihan $\{(x_i, y_i)\}_{i=1}^n$, estimasi parameter dilakukan sebagai berikut:

- Prior kelas:

$$P(y = c) = \frac{N_c}{n}$$

- Estimasi rata-rata dan variansi untuk fitur j pada kelas c :

$$\mu_{cj} = \frac{1}{N_c} \sum_{i:y_i=c} x_{ij}, \quad \sigma_{cj}^2 = \frac{1}{N_c} \sum_{i:y_i=c} (x_{ij} - \mu_{cj})^2 + \epsilon$$

dengan ϵ adalah smoothing varians untuk stabilitas numerik.

3. Inferensi: Log Posterior

Untuk prediksi, dihitung log posterior untuk setiap kelas:

$$\log p(y = c \mid x) = \log P(y = c) + \sum_{j=1}^d \log p(x_j \mid y = c)$$

Karena $p(x_j \mid y = c)$ berbentuk Gaussian, maka log-nya:

$$\log p(x_j \mid y = c) = -\frac{1}{2} \log(2\pi\sigma_{cj}^2) - \frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2}$$

Log posterior kemudian dinormalisasi menggunakan log-sum-exp untuk memperoleh probabilitas:

$$\log p(y = c \mid x) = \log p(x \mid y = c) + \log P(y = c) - \log \sum_{c'} p(x \mid y = c') P(y = c')$$

4. Prediksi Kelas

Prediksi akhir dilakukan dengan memilih kelas dengan log posterior maksimum:

$$\hat{y} = \arg \max_c \log p(y = c \mid x)$$