# Gaussian Naive Bayes

Evelyn Yosiana / 13522083

1. Konsep gaussian naive bayes pada intinya yaitu probabilitas (tepatnya likelihood probability). Model ini cocok digunakan untuk data numerik berdistribusi normal (sesuai dengan namanya, gaussian).

   **Proses train:** untuk proses train, dicari mean, variansi, dan prior (probabilitas awal) untuk tiap label target

   **Proses predict:** untuk proses ini, tiap data yang akan diprediksi (X_test) dihitung nilai pdf[label, fitur] (posterior probability). Misal terdapat 2 label (0 dan 1) dan 2 fitur (fitur_1 dan fitur_2), maka pdf yang dicari yaitu pdf0,fitur_1, pdf0,fitur_2, pdf1_fitur_1, dan pdf1,fitur_2. Persamaan untuk mencari pdf yaitu

   $$pdf_{n,m} = exp(-(x - mean_n)^2 / 2\, var_y) / sqrt(2 .\, phi .\, var_y)$$

   dengan n merupakan label yang diprediksi dan m merupakan banyak fitur.
   Setelah itu dicari log posterior (likelihood probability) dari tiap label dengan persamaan sebagai berikut.

   $$log\, posterior_n = log(prior_n) + log\,(pdf_{n,1}) + \ldots + log\,(pdf_{n,m})$$

   dengan n merupakan label yang diprediksi dan m merupakan banyak fitur.
   Setelah didapatkan log posterior dari seluruh label, label dengan log posterior tertinggi akan menjadi label hasil. Nilai log posterior sendiri menunjukkan seberapa kuat suatu data berasal dari kelas tertentu. Proses tersebut diulangi untuk seluruh data di X_test.

   **Parameter:**
   - classes : seluruh nilai prediksi (distinct).
   - mean: rata-rata awal.
   - var: variansi awal.
   - priors: probabilitas awal.

4. Pada model ini, hasil perbandingan model yang saya buat dan model dari library kurang sama. Hal ini dapat disebabkan oleh beberapa faktor, antara lain:
   - **Penggunaan epsilon**: dalam kode yang saya buat, terdapat penambahan epsilon (angka yang sangat kecil) untuk menghindari error (misalnya log(0)) pada variansi awal, sedangkan kode dalam library sudah dioptimasi menggunakan laplace smoothing yang lebih komprehensif untuk perhitungan yang lebih presisi.

- **Perhitungan prior**: dalam kode saya, prior dihitung secara manual sedangkan dalam library, prior dihitung dengan cara yang sudah dioptimasi untuk data yang imbalanced.

4. Improvement yang dapat saya lakukan antara lain:
    - **Hyperparameter tuning** untuk mendapatkan kombinasi parameter yang optimal (salah satunya dengan menggunakan metode grid search atau library optuna).
    - Menggunakan **confusion metrics** untuk model evaluation.

Lampiran

Contoh perhitungan manual:

Evelyn Yoriana / 13522083

GAUSSIAN NAIVE BAYES

contoh data

$x_{train} = \begin{bmatrix} 1 & 3 \\ 4 & 2 \\ 5 & 2 \end{bmatrix}$ $y_{train} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ learning rate = 0,01

$mean_0 = \left[\frac{1+5}{2}, \frac{3+2}{2}\right] = [3, 2,5]$

$var_0 = \frac{(1-3)^2+(5-3)^2}{2-1}, \frac{(3-2,5)^2+(2-2,5)^2}{2-1} = [8, 0,5]$

$prior_0 = \frac{2}{3} \approx 0,666$

$mean_1 = \left[\frac{4}{1}, \frac{2}{1}\right] = [4,2]$

$var_1 = [0,0]$ tapi di codenya ditambah $1e^{-9}$ biar gak 0

$prior_1 = \frac{1}{3} \approx 0,333$

Rumus!

$pdf_y(x) = \frac{\exp \frac{-(x-mean_y)^2}{2.var_y}}{\sqrt{2\pi\, var_y}}$

predict [3,7] :

$pdf_{0,1} = \frac{\exp\left(\frac{-(3-3)^2}{2.8}\right)}{\sqrt{2\pi 8}} \approx 0,1413$

$pdf_{0,2} = \frac{\exp\left(\frac{-(7-2,5)^2}{2.0,5}\right)}{\sqrt{2\pi.0,5}} \approx 5,9$

$pdf_{1,1} = \frac{\exp\left(\frac{-(3-4)^2}{2.1e^{-9}}\right)}{\sqrt{2\pi.1e^{-9}}} \approx 2,3 \times 10^4$

$pdf_{1,2} = \frac{\exp\left(\frac{-(7-2)^2}{2.1e^{-9}}\right)}{\sqrt{2\pi\ 1e^{-9}}} \approx 1,4 \times 10^{10}$

$\log posterior_0 = \log(prior_0) + \log(pdf_{0,1}) + \log(pdf_{0,2})$
$\approx -23,8881$

$\log posterior_1 = \log(prior_1) + \log(pdf_{1,1}) + \log(pdf_{1,2})$
$\approx 32,4609$

$\log posterior_1 > \log posterior_0$, so prediction = 1 //