Gaussian Naive Bayes

Salman Hanif - 13523056

Cara Kerja Algoritma

Algoritma Gaussian Naive Bayes ini didasari pada teorema bayes dengan adanya asumsi independent dari fitur-fitur data. Maksudnya satu fitur dianggap tidak ada relasi dengan fitur lain atau fitur tidak saling memengaruhi, meskipun pada kenyataannya belum tentu seperti itu. Algoritma ini juga didasari kepercayaan pada distribusi normal (Gaussian) pada fitur-fiturnya.

Teorema Bayes :

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}$$

P(y|X) = Probabilitas posterior, prob y (classification) diberikan data X.

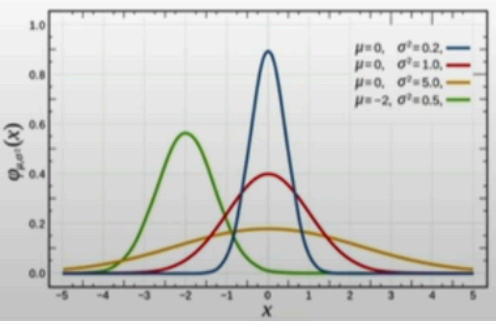
Sehingga dengan berbagai modifikasi, nantinya nilai y diambil dari rumus:

$$y = argmax_y \log(P(x_1|y)) + \log(P(x_2|y)) + ... + \log(P(x_n|y)) + \log(P(y))$$

log(P(y)) : Probabilitas Prior

Distribusi Gaussian :

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \cdot exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$



dengan μ adalah rata-rata serta σ adalah varians.

Fit (Training):

- Untuk setiap kelas, hitung probabilitas prior P(y) : jumlah data di kelas tersebut dibagi total sampel.
- Untuk setiap kelas dan setiap fitur, hitung rata-rata (μy) dan varians (σy2) dari nilai fitur tersebut.

Prediction :

- Untuk sebuah data point baru X yang ingin diklasifikasikan, hitung nilai P(X | y) dan P(y) untuk setiap kelas.
- Probabilitas P(X | y) dihitung dengan mengalikan probabilitas setiap fitur, yang didapat dari fungsi kepadatan probabilitas Gaussian menggunakan rata-rata dan varians yang dihitung pada tahap pelatihan.
- Kelas dengan nilai P(X | y) + P(y) tertinggi akan menjadi hasil prediksi.

Hasil Evaluasi model dari hasil scratch dan dari library. Jelaskan perbedaan

Dalam evaluasi, digunakan metode Hold-out dan K-Fold pada library scikit-learn dan Logistic Regression buatan saya. dan performa model saya di atas scikit learn secara akurasi.

Tetapi, hal ini dikarenakan sepertinya dataset saya agak bermasalah (dikarenakan semua fitur dimasukkan tanpa melihat korelasi) dan klasifikasi binary selalu mengarah ke 1 dibanding 0. Sepertinya hal ini diakibatkan adanya fitur kategorikal yang di one-hot encoding sehingga memunculkan varians 0.

Penanganan pada varians 0 inilah yang kemungkinan dimiliki oleh library Scikit-Learn sehingga performanya tidak terganggu.

Improvement yang bisa dilakukan

Improvement yang bisa saya lakukan adalah mengembangkan penanganan pada data hasil one-hot encoding yang memunculkan varians 0 sehingga data tidak memunculkan performa buruk ketika fitur kategorikal ikut serta.