

Seleksi Asisten Lab AI '23

Supervised Learning: Gaussian Naive Bayes

1. Cara Kerja Algoritma Gaussian Naive Bayes.

Konsep inti: Gaussian Naive Bayes (GNB) adalah klasifikasi probabilistik yang didasarkan pada Teorema Bayes. Asumsi utamanya yang "naif" adalah bahwa semua fitur bersifat independen satu sama lain. GNB cocok untuk fitur kontinu karena ia mengasumsikan bahwa data dari setiap kelas terdistribusi secara Gaussian (Normal).

Probabilitas posterior dihitung sebagai berikut:

$$P(y|x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i|y)$$

Di mana $P(x_i|y)$ dihitung menggunakan *Probability Density Function* (PDF) dari distribusi Gaussian. Kelas dengan probabilitas posterior tertinggi akan menjadi hasil prediksi.

2. Analisis Perbandingan Eksperimen.

Table 1: Perbandingan kinerja model Gaussian Naive Bayes

Metric	From Scratch	Scikit-learn
Accuration	0.6558	0.6558
Precision	0.43	0.43
Recall	0.87	0.87
Time (s)	~0.07	~0.01

Kedua implementasi memberikan hasil yang identik. Model GNB memiliki akurasi yang lebih rendah dibandingkan model lain (KNN, Logistic Regression), namun menunjukkan nilai Recall yang sangat tinggi (0.87). Artinya, model ini sangat baik dalam mengidentifikasi pelanggan yang akan *churn*, meskipun dengan mengorbankan *precision* (banyak *false positive*).

3. *Improvement* yang bisa dilakukan untuk mencapai hasil yang lebih baik.

Kinerja model GNB sangat dipengaruhi oleh asumsi dasarnya. Berikut adalah beberapa area perbaikan:

- Transformasi Fitur
GNB mengasumsikan fitur kontinu terdistribusi secara Gaussian. Jika distribusi asli fitur *skewed*, kinerjanya akan menurun. Melakukan transformasi pada fitur numerik untuk membuatnya lebih mendekati distribusi normal mungkin dapat meningkatkan performa model.
- Menggunakan Varian Naive Bayes yang Tepat
Dataset yang digunakan memiliki campuran fitur numerik dan kategorikal (yang sudah di-*one-hot encode*). GNB ideal untuk fitur numerik. Pendekatan yang lebih canggih adalah menggunakan model Naive Bayes yang berbeda untuk setiap jenis fitur (misalnya, *CategoricalNB* untuk fitur asli yang kategorikal dan *GaussianNB* untuk fitur numerik), kemudian menggabungkan hasilnya.