

Logistic Regression

1. Cara kerja algoritma

Logistic Regression adalah salah satu algoritma machine learning yang digunakan untuk klasifikasi, terutama biner, meskipun dapat diperluas untuk klasifikasi multi kelas.

Berbeda dengan KNN yang berbasis instance, Logistic Regression adalah model berbasis probabilistik yang memprediksi probabilitas suatu data termasuk dalam kelas tertentu.

Logistic Regression bekerja dengan memodelkan hubungan antara fitur input dan probabilitas dari kelas output menggunakan fungsi logistik (sigmoid). Fungsi sigmoid mengubah output linear menjadi probabilitas antara 0 dan 1.

1. Menentukan fungsi sigmoid

Fungsi sigmoid atau fungsi logistik digunakan untuk mengkonversi output model menjadi probabilitas, Fungsi sigmoid didefinisikan sebagai $\sigma(z) = 1/(1 + e^{-z})$ di mana z adalah output dari model linear.

2. Model linear

Logistic regression memodelkan hubungan antara fitur dan kelas dengan fungsi linear: $z = w^T x + b$ di mana w adalah vektor bobot, x adalah vektor fitur, dan b adalah bias.

3. Menghitung probabilitas

Setelah menghitung z , probabilitas kelas positif (misalnya, kelas 1) dihitung dengan fungsi sigmoid

4. Menentukan kelas

Kelas diprediksi dengan membandingkan probabilitas p dengan threshold (biasanya 0.5)

- Jika $p \geq 0.5$, prediksi kelas adalah 1
- Jika $p < 0.5$, prediksi kelas adalah 0

5. Menghitung loss function

Logistic regression menghitung loss function cross-entropy (log loss) untuk mengukur perbedaan antara probabilitas yang diprediksi dan label yang sebenarnya.

6. Optimasi dengan gradient descent

Gradient descent digunakan untuk mengoptimalkan bobot dan bias dengan meminimalkan loss function. Gradien dari loss function dihitung dan digunakan untuk memperbarui bobot dan bias.

7. Evaluasi

Setelah training, model dievaluasi menggunakan data uji untuk mengukur akurasi dan kinerja. Jika diperlukan, parameter seperti learning rate, jumlah iterasi, atau regularisasi dapat disesuaikan

2. Perbandingan model buatan sendiri dan library

Berdasarkan hasil evaluasi, kedua logistic regression model memiliki nilai Recall dan F1-Score yang sama, yaitu 0.89 dan 0.88. Ini menunjukkan bahwa kedua model memiliki performa yang setara dalam hal metrik yang diukur

3. Improvement yang bisa dilakukan

- Optimasi Hyperparameter
Gunakan teknik seperti Grid Search atau Random Search untuk menemukan kombinasi hyperparameter yang optimal.
- Learning Rate Schedule
Mengimplementasikan penjadwalan learning rate yang menurun seiring waktu untuk membantu konvergensi yang lebih baik.
- Early Stopping
Implementasikan early stopping untuk menghentikan pelatihan ketika performa pada data validasi mulai menurun.
- Regularizations
Eksperimen dengan nilai `reg_lambda` yang berbeda untuk menemukan keseimbangan yang tepat antara bias dan varians.