Logistic Regression

Kelompok 1

DAFTAR ISI

PENDAHULUAN

Logistic Regression berawal dari tahun 1844 ketika Pierre Verhulst memperkenalkan kurva sigmoid untuk memodelkan pertumbuhan populasi. Pada awal 1900-an, fungsi ini mulai digunakan dalam biologi dan statistik untuk mempelajari proses pertumbuhan. Lalu, pada 1958, David Cox memformalkan metode Logistic Regression sebagai teknik untuk memprediksi probabilitas kejadian biner. Sejak era komputer pada 1970–1980-an, metode ini menjadi populer karena mudah dihitung, dan memasuki era Machine Learning di 2000-an, Logistic Regression menjadi algoritma dasar untuk klasifikasi biner.

APA ITU LOGISTIC REGRESSION?

Logistic Regression adalah algoritma Machine Learning yang digunakan untuk klasifikasi, yaitu memprediksi apakah suatu data termasuk ke dalam kelas 0 atau 1. Berbeda dengan regresi linear yang menghasilkan angka kontinu, Logistic Regression menghasilkan probabilitas (nilai antara 0 sampai 1) menggunakan fungsi bernama sigmoid. Nilai probabilitas ini kemudian dikonversi menjadi kelas; misalnya, jika probabilitas ≥ 0.5 maka diprediksi sebagai kelas 1, dan jika < 0.5 maka kelas 0.

DIGUNAKAN DIMANA?

Logistic Regression banyak digunakan dalam kasus seperti:

- Deteksi penyakit (sehat / sakit)
- Analisis default pinjaman (lancar / macet)
- Deteksi spam (spam / bukan spam)
- Prediksi kelulusan (lulus / tidak lulus)

Meskipun namanya ada kata “regression”, model ini sebenarnya dipakai untuk klasifikasi biner, bukan prediksi angka. Kata “regression” digunakan karena proses pembelajarannya tetap menggunakan pendekatan regresi untuk menghitung parameter model, namun hasil akhirnya berupa kategori.

CARA KERJA

1. Menerima input fitur, yaitu variabel yang digunakan untuk membuat prediksi.
2. Menggabungkan fitur dalam sebuah persamaan linear menggunakan bobot (parameter) dan bias.
3. Mengonversi nilai linear ke probabilitas menggunakan fungsi sigmoid agar output berada di rentang 0 sampai 1.
4. Menghitung kesalahan prediksi menggunakan cost function (log loss) untuk mengetahui seberapa jauh probabilitas yang dihasilkan dari nilai sebenarnya.
5. Melakukan proses training dengan mengubah bobot secara bertahap melalui algoritma optimasi seperti gradient descent agar cost semakin kecil.
6. Menghasilkan probabilitas kelas setelah model terlatih, sebagai dasar pengambilan keputusan.
7. Menerapkan threshold (biasanya 0.5) untuk menentukan apakah data masuk ke kelas 0 atau kelas 1 berdasarkan probabilitas tersebut.

HIPOTESA FUNCTION

Hypothesis function adalah rumus yang digunakan model untuk menghasilkan prediksi probabilitas apakah suatu data termasuk kelas 1.

Rumusnya:

$$h_\theta(x) = \sigma(\theta^T x)$$

1. $h_\theta(x)$
   - probabilitas prediksi model (0–1)
2. $\theta \rightarrow$
   - parameter/weight yang dipelajari.
3. x
   - fitur input.
4. $\theta^T x$
   - perkalian linear antara parameter dan fitur.
5. $\sigma(\cdot)$
   -fungsi sigmoid untuk mengubah nilai linear menjadi probabilitas.

Fungsi sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

1. $\sigma(z)$

- fungsi sigmoid yang menghasilkan nilai 0–1.

2. $z \rightarrow$

- input (biasanya hasil θTx\theta^T xθTx).

3. e^{-z}

- bagian eksponensial yang menentukan bentuk kurva sigmoid.

4. : $1 + e^{-z}$

- memastikan hasil selalu positif dan terbatas.

5. Hasil Akhir

- nilai probabilitas antara 0 dan 1.

COST FUNCTION

Cost function adalah rumus yang digunakan untuk mengukur seberapa jauh prediksi model berbeda dari nilai yang sebenarnya.
Dalam Logistic Regression, output model adalah probabilitas (antara 0 sampai 1), sehingga cost function yang dipakai harus bisa menangani probabilitas.
Pada Logistic Regression, cost function yang digunakan adalah Log Loss atau Binary Cross-Entropy.

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

1. J(θ)
→ total cost (error) model.

2. m
Jumlah seluruh data training.

3. $\sum_{i=1}^{m}$
Menjumlahkan error untuk setiap data satu per satu.

4. $y^{(i)}$
Label asli data ke-i (0 = tidak, 1 = ya).

5. $h_\theta(x^{(i)})$
Prediksi model (probabilitas), hasil dari fungsi sigmoid.

6. $y^{(i)} \log(h_\theta(x^{(i)}))$
Bagian error jika label sebenarnya 1.

7. $(1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$
Bagian error jika label sebenarnya 0.

8. Tanda minus di depan
Supaya nilai cost selalu positif.

9.Dibagi 1/m
Supaya error jadi rata-rata, bukan total.

CONTOH KASUS

Sebuah perusahaan e-commerce ingin meningkatkan efektivitas iklan berbayar di media sosial dengan menargetkan pengguna yang memiliki kemungkinan terbesar untuk melakukan pembelian. Mereka memiliki data pengguna yang mencakup usia, estimasi gaji tahunan, serta informasi apakah pengguna tersebut melakukan pembelian atau tidak. Dengan menggunakan dataset Social Network Ads dari Kaggle, perusahaan meminta Anda untuk membangun model Logistic Regression untuk memprediksi apakah seorang pengguna akan membeli produk berdasarkan usia dan estimasi gajinya.

Berdasarkan hasil pengujian menggunakan Logistic Regression, baik dari sklearn maupun implementasi manual, model mampu memprediksi kemungkinan pengguna melakukan pembelian dengan akurasi sekitar 86%. Nilai precision untuk kelas pembeli (1) sebesar 0.90, menunjukkan bahwa prediksi “akan membeli” cukup akurat. Namun recall-nya 0.68, artinya masih ada pengguna yang sebenarnya membeli tetapi tidak terdeteksi oleh model. Pada implementasi manual, nilai cost function turun dari 0.69 menjadi sekitar 0.36, menandakan proses pembelajaran berlangsung dengan baik. Secara keseluruhan, model ini sudah cukup efektif membantu perusahaan dalam menargetkan iklan ke pengguna yang berpotensi membeli berdasarkan usia dan estimasi gaji, meskipun performa bisa ditingkatkan melalui penambahan fitur atau model yang lebih kompleks.

MARI DISKUSI

KAPAN HARUS DIGUNAKAN

Logistic Regression adalah metode klasifikasi yang ideal ketika masalah yang dihadapi melibatkan dua kemungkinan hasil (klasifikasi biner), seperti memprediksi apakah seorang nasabah akan membeli produk (Ya/Tidak), apakah email adalah spam (Spam/Bukan Spam), atau apakah pinjaman akan gagal bayar (Gagal/Berhasil). Keunggulan utamanya adalah ia secara alami memodelkan probabilitas hasil, bukan hanya label kelas. Artinya, model tidak hanya mengatakan "Ya", tetapi juga memberikan tingkat kepercayaan, misalnya "Probabilitas Ya adalah 85%."

Selain itu, Regresi Logistik sangat diutamakan ketika interpretasi dari model adalah hal yang penting. Karena ia adalah model linear dalam peluang, koefisien model dapat menjelaskan dengan jelas bagaimana perubahan pada setiap variabel independen (fitur) akan memengaruhi peluang hasil akhir. Model ini juga sangat efisien secara komputasi dan sering digunakan sebagai model baseline yang cepat dan kuat untuk membandingkan kinerja model-model klasifikasi yang lebih kompleks. Regresi Logistik bekerja dengan baik ketika batas pemisah antar kelas relatif sederhana.

KAPAN TIDAK DIGUNAKAN

Logistic Regression tidak digunakan saat memecahkan masalah regresi, yaitu memprediksi nilai numerik kontinu seperti memprediksi harga saham, suhu besok, atau berat badan. Model ini secara fundamental dirancang untuk masalah klasifikasi (memprediksi kategori). Selain itu, model ini juga kurang efektif ketika data memiliki hubungan yang sangat non-linear atau batas keputusan yang kompleks. Karena Logistic Regression mencari batas pemisah linear (atau batas yang dapat diubah menjadi linear melalui peluang), ia akan berkinerja buruk pada data di mana kelas-kelasnya terjalin secara rumit.

Logistic Regression juga cenderung menghadapi tantangan dalam masalah klasifikasi multikelas dengan banyak kategori dibandingkan dengan model klasifikasi yang lebih canggih. Walaupun bisa diperluas (Multinomial Logistic Regression), model berbasis pohon keputusan seringkali lebih efisien untuk membedakan banyak kelas secara bersamaan. Terakhir, model ini sensitif terhadap isu multikolinearitas, yaitu ketika variabel independen sangat berkorelasi satu sama lain, yang dapat menyebabkan koefisien menjadi tidak stabil dan merusak interpretasi model yang seharusnya menjadi keunggulannya.

KESIMPULAN

Secara keseluruhan, Logistic Regression merupakan metode klasifikasi yang sederhana namun efektif untuk kasus prediksi pembelian pada dataset Social Network Ads. Dari hasil pengujian ini, dengan menggunakan metode Logistic Regression menjadikan pengujian kemampuan yang sangat baik dalam mengenali pengguna yang tidak membeli, serta precision yang tinggi untuk kelas pembeli. Hal ini menunjukkan bahwa ketika model memprediksi seseorang akan membeli, prediksinya cukup dapat diandalkan.
Namun model masih melewatkan sebagian pembeli nyata, terlihat dari nilai recall kelas 1 yang lebih rendah. Ini menunjukkan bahwa batas keputusan linear Logistic Regression belum sepenuhnya mampu menangkap pola pembeli yang mungkin non-linear. Meskipun begitu, model ini tetap memberikan pemahaman yang jelas mengenai pengaruh usia dan estimasi gaji terhadap peluang pembelian, dan dapat dijadikan baseline yang solid sebelum mencoba model lain yang lebih kompleks.

HASIL AKURASI

```
Akurasi: 0.8625
              precision    recall  f1-score   support

           0       0.85      0.96      0.90        52
           1       0.90      0.68      0.78        28

    accuracy                           0.86        80
   macro avg       0.88      0.82      0.84        80
weighted avg       0.87      0.86      0.86        80
```

Dengan hanya menggunakan Logistic Regression yang merupakan model paling sederhana dan paling interpretable kita sudah berhasil mencapai akurasi 86,25% dan weighted F1-score 86%. Ini menunjukkan bahwa model kita sudah sangat baik dan reliable. Pencapaian ini bahkan sudah cukup untuk diimplementasikan di Social Network, apalagi jika kita mengutamakan interpretabilitas dan kecepatan prediksi.