Dasar-Dasar Machine Learning

1. Supervised Learning

Pada supervised learning, model belajar dari data yang sudah berlabel. Tugas klasifikasi memprediksi kategori, misalnya menandai surel sebagai spam atau bukan spam. Tugas regresi memprediksi nilai kontinu, misalnya harga rumah berdasarkan luas dan lokasi.

2. Pembagian Data

Data biasanya dibagi menjadi data latih 70 persen, data validasi 15 persen, dan data uji 15 persen. Data validasi dipakai untuk menyetel hiperparameter, sedangkan data uji hanya dipakai satu kali di akhir untuk menilai performa akhir. Data uji tidak boleh dipakai saat tuning karena akan menyebabkan kebocoran informasi.

3. Overfitting dan Underfitting

Overfitting terjadi ketika model sangat baik pada data latih tetapi buruk pada data baru. Cirinya adalah selisih besar antara akurasi latih, misalnya 99 persen, dan akurasi validasi, misalnya 72 persen. Model seolah menghafal data latih, termasuk derau di dalamnya.

Underfitting terjadi ketika model terlalu sederhana sehingga akurasi latih maupun validasi sama-sama rendah. Cara mengatasi overfitting antara lain regularisasi, early stopping, dropout pada jaringan saraf, dan menambah jumlah data.

4. Cross-Validation

Cross-validation menilai model secara lebih stabil daripada satu kali pembagian data. Pada k-fold cross-validation, data dibagi menjadi k lipatan dan setiap lipatan bergantian menjadi data validasi. Nilai k yang lazim adalah 5 atau 10.

Untuk kelas yang tidak seimbang, gunakan stratified k-fold agar proporsi kelas pada tiap lipatan tetap terjaga.

5. Metrik Klasifikasi

Akurasi menghitung proporsi prediksi yang benar, tetapi menyesatkan bila kelas tidak seimbang. Precision mengukur seberapa banyak prediksi positif yang benar, sedangkan recall mengukur seberapa banyak data positif yang berhasil ditemukan.

F1-score adalah rata-rata harmonik dari precision dan recall sehingga nilainya rendah bila salah satunya rendah. Untuk deteksi penyakit, recall diutamakan karena melewatkan pasien jauh lebih berbahaya daripada alarm palsu. Untuk filter spam, precision diutamakan agar surel penting tidak terbuang.

Tabel 1. Ringkasan metrik klasifikasi

Metrik	Rumus	Dipakai ketika
Akurasi	(TP+TN)/(TP+TN+FP+FN)	Kelas seimbang
Precision	TP/(TP+FP)	Biaya false positive tinggi
Recall	TP/(TP+FN)	Biaya false negative tinggi
F1-score	2*P*R/(P+R)	Kelas tidak seimbang

Sumber: ringkasan penulis.

6. Gradient Descent

Gradient descent adalah algoritma optimasi iteratif yang memperbarui parameter ke arah berlawanan dengan gradien fungsi kerugian. Learning rate yang terlalu besar membuat pelatihan divergen, sedangkan yang terlalu kecil membuat pelatihan sangat lambat.

Pada mini-batch gradient descent, gradien dihitung dari sebagian kecil data, biasanya berukuran 32 atau 64 sampel per langkah.

7. Decision Tree dan Random Forest

Decision tree membagi data berdasarkan fitur dengan kriteria Gini impurity atau entropi. Pohon yang terlalu dalam cenderung overfitting. Random forest menggabungkan banyak pohon yang dilatih pada sampel acak berbeda (bagging) lalu mengambil suara terbanyak, sehingga variansi prediksi menurun.

8. Regularisasi

Regularisasi L1, atau Lasso, mendorong sebagian bobot menjadi tepat nol sehingga berfungsi sebagai seleksi fitur. Regularisasi L2, atau Ridge, mengecilkan bobot tanpa menjadikannya nol. Kekuatan regularisasi diatur oleh hiperparameter lambda.