HACKTIV8

Machine Learning Algorithms - Part 3

Kampus Merdeka
IBM SkillsBuild For AI & Cybersecurity

HACKTIV8

Agenda

- Mempelajari tentang Ensemble Learning
- Mempelajari tentang Hyperparameter Tuning
- Mempelajari tentang Cross Validation
- Hands-On Ensemble Learning & Hyper Parameter Tuning.

HACKTIV8

Ensemble Learning

Kampus Merdeka IBM SkillsBuild For AI & Cybersecurity

Ensemble Learning

Definisi

Ensemble learning adalah teknik Machine Learning yang melibatkan penggabungan beberapa model individu untuk membuat model yang lebih kuat. Model baru ini seringkali lebih kokoh dan lebih akurat. Berbagai model machine learning mungkin beroperasi pada sampel data populasi yang berbeda, teknik pemodelan yang berbeda mungkin digunakan, dan hipotesis yang berbeda mungkin digunakan.

Metode ensemble bertujuan untuk mengurangi varians dan bias dari model individu, membuat model yang digabungkan lebih generalisasi dan lebih baik dalam menangani berbagai jenis data. Proses pembuatan ensemble biasanya melibatkan pelatihan beberapa model pada dataset yang sama dan kemudian menggabungkan prediksi mereka dengan cara tertentu.

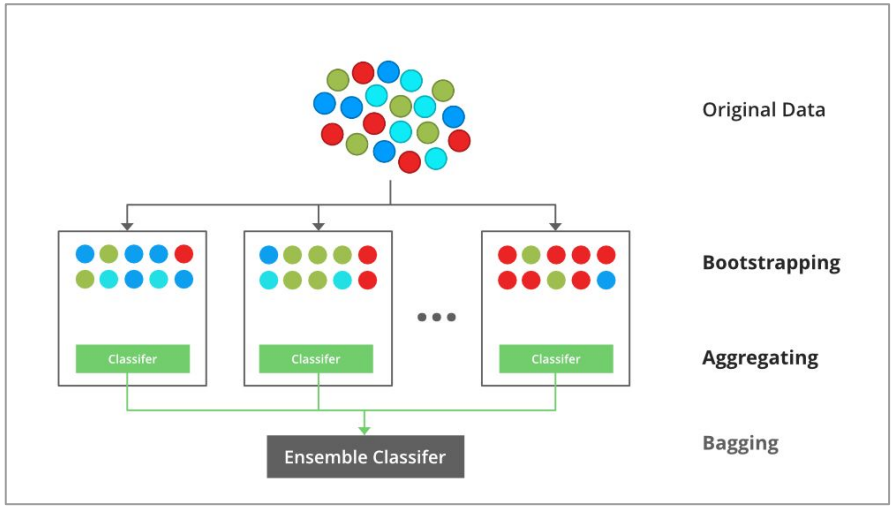
Ada 3 jenis Ensemble Learning:

- Bagging
- Boosting
- Stacking

Namun pada sesi ini kita hanya akan mempelajari Bagging dan Boosting.

Bagging

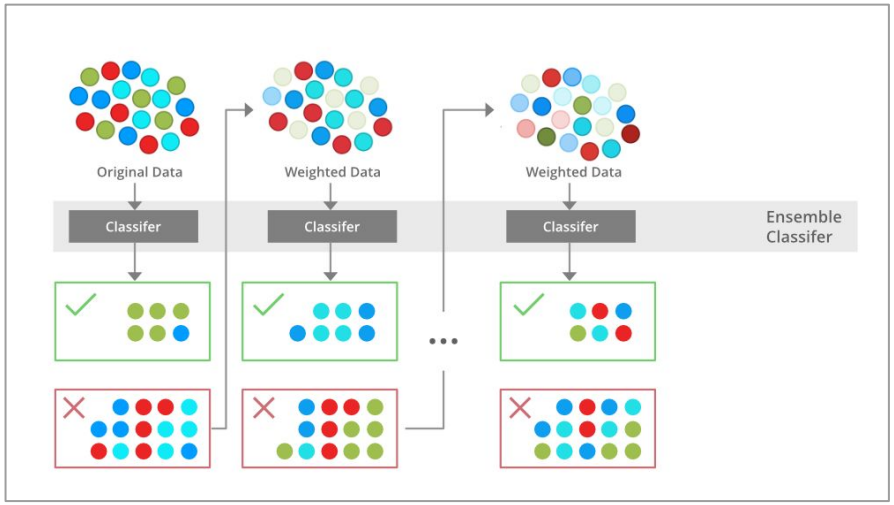
Definisi



- **Bagging** adalah singkatan dari Bootstrap AGGregating karena menggabungkan Bootstrapping dan Aggregasi untuk membentuk satu model ensemble.
- **Bootstrapping**: pengambilan sampel dengan penggantian (sampling with replacement).
- Dalam bagging, beberapa model dilatih secara independen pada subset yang berbeda dari data pelatihan, seringkali diperoleh melalui bootstrapping.
- Prediksi dari setiap model kemudian digabungkan, biasanya dengan cara averaging untuk tugas regresi atau dengan mayoritas voting untuk tugas klasifikasi.
- **Random Forest** adalah contoh populer dari algoritma bagging. Random Forest adalah model ensemble dimana setiap model independen adalah Decision Tree.

Boosting

Definisi



- Dalam boosting, **model-model dilatih secara berurutan**, dengan setiap model berfokus pada **memperbaiki kesalahan model sebelumnya**.
- Pada setiap langkah, algoritma memberikan bobot yang lebih tinggi pada titik data yang salah diklasifikasikan, memaksa model-model berikutnya untuk memperhatikan kasus-kasus tersebut..
- Prediksi dari semua model kemudian digabungkan menggunakan weighted averaging.
- Contoh algoritma boosting termasuk AdaBoost, Gradient Boosting, dan XGBoost.

Bagging vs Boosting

Perbedaan

Bagging	Boosting
Setiap model dibangun secara independen (pelatihan paralel)	Model-model baru dipengaruhi oleh kinerja model yang telah dibangun sebelumnya (sequential training)
Melatih beberapa model secara independen pada subset yang berbeda dari data pelatihan.	Melatih menggunakan data yang salah diklasifikasikan oleh model sebelumnya.
Prediksi dari masing-masing model digabungkan dengan cara rata-rata (untuk tugas regresi) atau dengan voting mayoritas (untuk tugas klasifikasi).	Prediksi dari masing-masing model digabungkan menggunakan rata-rata terbobot.
Contoh : Random Forest	Contoh : AdaBoost, XGBoost

HACKTIV8

Hyperparameter Tuning

Kampus Merdeka IBM SkillsBuild For AI & Cybersecurity

Hyperparameter

Definisi

Sejauh ini, Anda telah mempelajari berbagai algoritma dalam Machine Learning. Setiap algoritma memiliki mekanisme kerja yang berbeda dan asumsi yang perlu dipenuhi agar algoritma dapat berfungsi secara optimal.

Setiap algoritma memiliki hyperparameter mereka sendiri. **Hyperparameter** adalah variabel konfigurasi eksternal yang digunakan untuk mengelola model Machine Learning selama fase pelatihan. Hyperparameter biasanya dipilih berdasarkan masalah yang sedang dipelajari atau melalui eksperimen dan pencarian grid untuk menemukan nilai yang optimal.

Perbedaan utama antara hyperparameter dan parameter adalah bahwa hyperparameter mengontrol proses pelatihan model sedangkan parameter adalah bagian dari model itu sendiri yang ditemukan selama pelatihan, seperti bias, bobot dan lain-lain.
Contoh, algoritma Random Forest:
- Hyperparameter: Number of trees, maximum depth of trees, minimum samples split.
- Parameter: Decision rules in each tree, feature subsets

Hyperparameter Tuning

Definisi

Hyperparameter dapat memiliki dampak besar pada kinerja algoritma pembelajaran. Pengaturan hyperparameter seringkali berbeda untuk setiap kumpulan data. Oleh karena itu, mereka harus dioptimalkan untuk setiap kumpulan data.

Proses menemukan hyperparameter terbaik untuk suatu kumpulan data tertentu disebut Hyperparameter Optimization atau Hyperparameter Tuning. Tantangan dalam melakukan ini adalah:

1. Tidak ada formula pasti untuk menemukan hyperparameter. Diperlukan eksperimen dan evaluasi berulang untuk menemukan kombinasi yang optimal.
2. Harus mencoba berbagai kombinasi hyperparameter dan mengevaluasi kinerja model.

Selama melakukan Proses ini biasanya data akan dibagi menjadi beberapa bagian salah satunya yaitu dengan Cross Validation agar mendapatkan hasil evaluasi yang tepat dan akurat.

Cross Validation

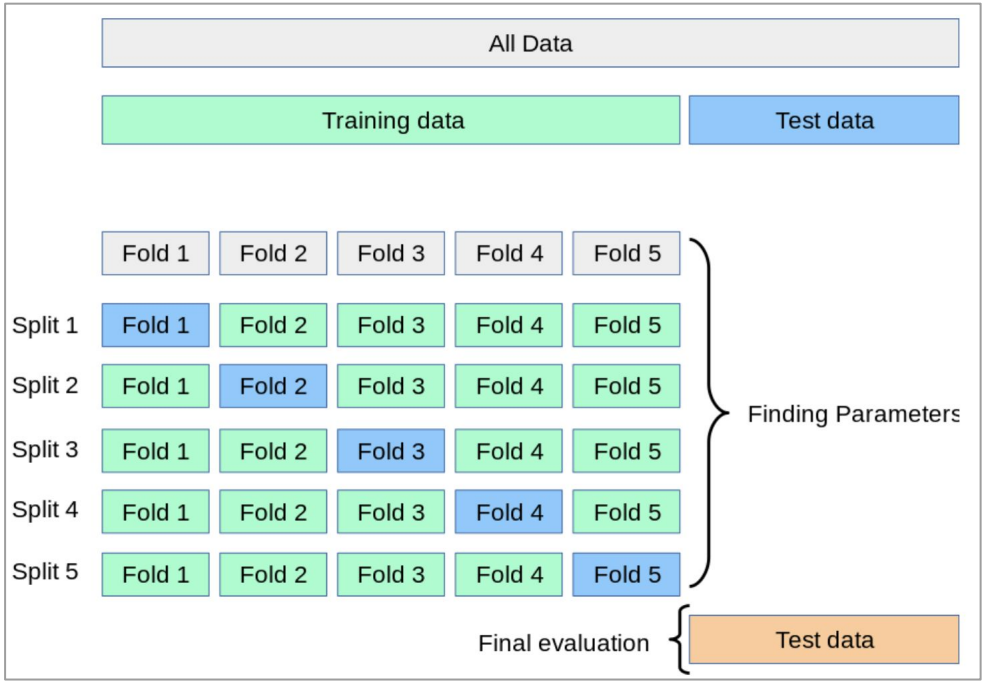
Definisi

Cross-validation adalah teknik untuk memperkirakan kinerja model pada data yang belum terlihat. Manfaat dari Cross-validation adalah:

1. Cross-validation secara sistematis membuat dan mengevaluasi beberapa model pada beberapa subset dari kumpulan data.
2. Memberikan pengukuran kinerja seperti mean dan standar deviasi.
   - mean: seberapa baik prosedur tersebut berperforma secara rata-rata.
   - standar deviasi: seberapa besar variasi yang diharapkan dari prosedur tersebut dalam praktiknya.
3. Mean dan Standard Deviation dapat digunakan untuk memberikan interval kepercayaan pada kinerja yang diharapkan di set pengujian.

Cross Validation

Ilustrasi



ilustrasi dari Cross Validation (dikenal juga sebagai K-Fold Cross Validation)

Tuning Method

Algorithms (1)

Sebelum mencoba berbagai kombinasi hyperparameter, ada beberapa hal yang perlu dipertimbangkan:
1. Jumlah hyperparameter dari algoritma Machine Learning.
2. Ketersediaan sumber daya seperti sumber daya komputasi dan ukuran dataset.
3. Batas waktu yang tersedia untuk melakukan optimasi hyperparameter.

Ada dua algoritma paling populer yang dapat digunakan untuk melakukan Penyetelan Hyperparameter:
1. Grid Search:
   - ➔ Metode ini akan melatih setiap kombinasi nilai hyperparameter yang telah diatur.
   - ➔ Karena metode ini menggunakan setiap kombinasi untuk membangun dan mengevaluasi kinerja model, metode ini sangat mahal secara komputasi.
2. Random Search:
   - ➔ Metode ini akan melatih kombinasi acak dari hyperparameter.
   - ➔ Tidak semua kombinasi mungkin akan dilatih.
   - ➔ Biasanya diterapkan jika dataset besar.

Tuning Method

Algorithms (2)

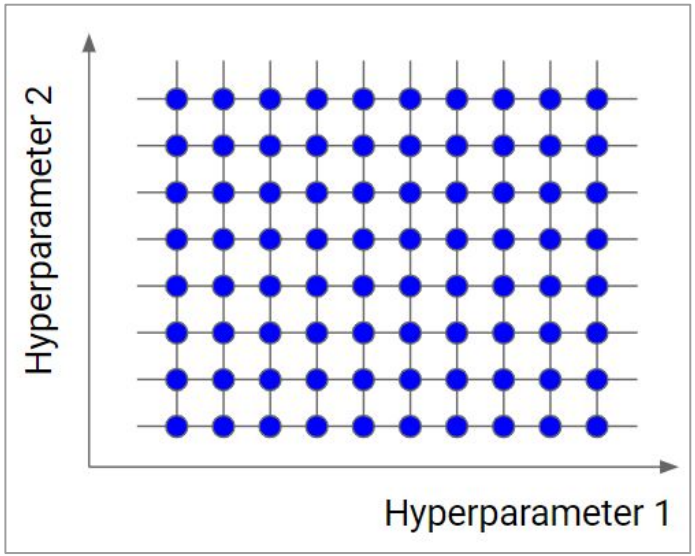


Illustration of Grid Search

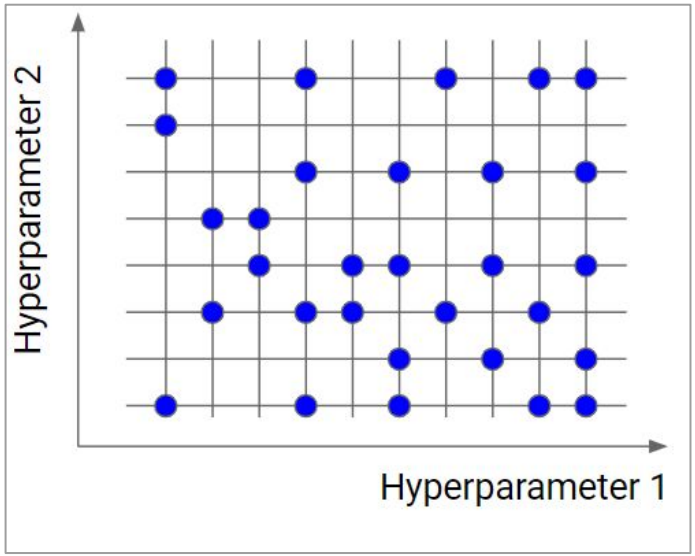


Illustration of Random Search

Source

Hands-On

Colab Link _____ *Visit Here*