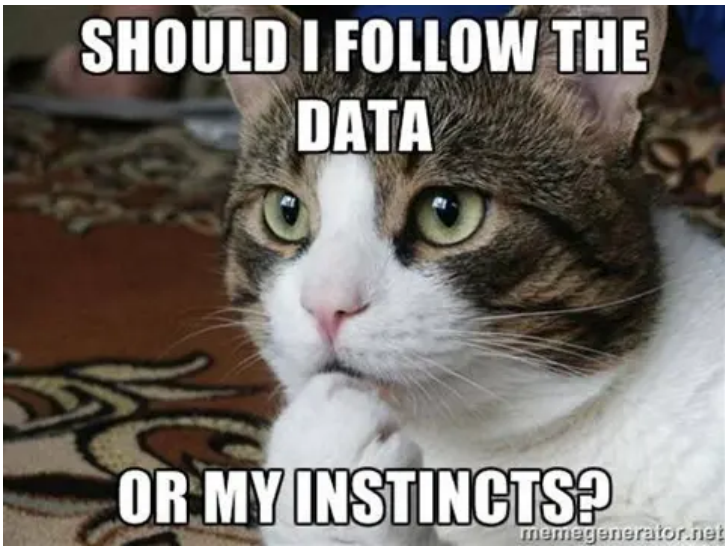
Tugas Besar 2 IF3170 Inteligensi Artifisial
Implementasi Algoritma Pembelajaran Mesin

Dipersiapkan Oleh Tim Asisten Lab AI ‘21

Versi: 1.0 17/11/2024
Versi: 1.1 19/11/2024
Versi: 1.2 25/11/2024
Versi: 1.3 28/11/2024
Versi: 1.4 29/11/2024
Versi 1.5: 30/11/2024
Versi 1.6: 6/12/2024



~~Deadline: Senin, 9 Desember 2024 23.59 WIB~~
Deadline: Minggu, 15 Desember 2024 23.59 WIB

Tujuan

Tugas Besar 2 pada kuliah IF3170 Inteligensi Buatan bertujuan untuk memberikan pengalaman langsung kepada peserta kuliah dalam menerapkan algoritma pembelajaran mesin pada permasalahan nyata.

Spesifikasi

Pembelajaran mesin merupakan salah satu cabang dari kecerdasan buatan yang memungkinkan sistem untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit.

Dataset **UNSW-NB15** adalah kumpulan data lalu lintas jaringan yang mencakup berbagai jenis serangan siber dan aktivitas normal. Pada tugas ini, Anda diminta untuk mengimplementasikan algoritma pembelajaran mesin yang telah kalian pelajari di kuliah, yaitu **KNN, Gaussian Naive-Bayes, dan ID3** pada dataset **UNSW-NB15**. Rincian spesifikasi untuk tugas besar 2 dapat dilihat sebagai berikut:

1. Implementasi KNN *from scratch*.
    a. Minimal bisa menerima 2 input parameter
        i. Jumlah tetangga
        ii. Metrik jarak antar data point. Minimal dapat menerima 3 pilihan, yaitu Euclidean, Manhattan, dan Minkowski
2. Implementasi Gaussian Naive-Bayes *from scratch*.
3. Implementasi ID3 *from scratch*, termasuk pemrosesan data numerik sesuai materi yang dijelaskan dalam PPT kuliah.
4. Implementasi algoritma poin 1-3 menggunakan *scikit-learn*. Bandingkan hasil dari algoritma *from scratch* dan algoritma *scikit-learn*. Untuk ID3 di *scikit-learn*, gunakan DecisionTreeClassifier dengan parameter criterion='entropy' (memang tidak sama persis dengan ID3, tetapi cukup mendekati)
5. Model harus bisa di-save dan di-load. Implementasinya dibebaskan (misal menggunakan .txt, .pkl, dll).
6. [Bonus] Kaggle Submission pada link berikut.

Implementasi KNN, Gaussian Naive-Bayes, dan ID3 yang *from scratch* bisa dalam bentuk kelas-kelas (class KNN, dst.) yang nantinya akan di-import ke notebook pengerjaan. Untuk implementasi *from scratch*, *library* yang boleh digunakan adalah untuk perhitungan matematika saja seperti numpy dan sejenisnya.

Asisten telah menyediakan notebook berikut untuk Anda lengkapi, dan deskripsi lengkap mengenai dataset dapat dilihat sebagai berikut:

Deskripsi Dataset

Dataset UNSW-NB15 merupakan dataset berisi raw network packets yang dibuat menggunakan IXIA PerfectStorm oleh Cyber Range Lab UNSW Canberra. Dataset ini terdiri dari

10 jenis aktivitas (9 jenis attack dan 1 aktivitas normal). Sembilan jenis attack yang termasuk ke dalam dataset ini adalah Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode dan Worms. Dataset yang akan digunakan (beserta metadata-nya) pada tugas ini dapat anda akses di tautan berikut (**Catatan: Terdapat perubahan dataset dari yang telah digunakan di Tugas Kecil 2, sehingga untuk Tugas Besar 2, gunakan data yang diberikan pada tautan tersebut**). Untuk tugas ini, variabel target yang akan digunakan hanya *attack_cat* saja.

Untuk menghasilkan prediksi yang berkualitas, Anda diharuskan untuk melakukan beberapa tahap berikut ini (tahapan lebih lengkap dapat dilihat di template notebook):

Data Cleaning

Tahap ini bertujuan untuk membersihkan dataset dari nilai yang hilang (missing values), data duplikat, atau data yang tidak valid sehingga dataset siap digunakan untuk analisis.

Data Transformation

Transformasi data melibatkan langkah-langkah seperti encoding variabel kategori, normalisasi atau standarisasi fitur numerik, serta penanganan ketidakseimbangan data (imbalanced data) untuk memastikan data berada dalam format yang sesuai dengan algoritma pembelajaran mesin.

Feature Selection

Pemilihan fitur yang relevan bertujuan untuk mengurangi kompleksitas model, menghindari overfitting, serta meningkatkan kinerja model. Langkah ini melibatkan identifikasi fitur yang memiliki pengaruh signifikan terhadap variabel target.

Dimensionality Reduction

Jika dataset memiliki jumlah fitur yang besar, reduksi dimensi dapat digunakan untuk mengurangi dimensi tanpa kehilangan informasi penting. Teknik seperti Principal Component Analysis (PCA) sering digunakan pada tahap ini.

Modeling dan Validation

Pada tahap ini, algoritma pembelajaran mesin seperti K-Nearest Neighbors (KNN), Naive Bayes, dan ID3 (Iterative Dichotomiser 3) diterapkan pada dataset. Anda akan melatih model pembelajaran mesin yang akan **mengklasifikasi kategori attack (attack_cat)** berdasarkan

fitur-fitur lain yang telah diberikan. Model yang telah dibuat divalidasi menggunakan metode seperti **train-test split** atau **k-fold cross-validation** untuk memastikan kinerja yang optimal.

Bonus

Untuk bonus, nilai diberikan berdasarkan ranking *leaderboard* **Kaggle** yang dirincikan sebagai berikut:

- Rank 1-3 = 10 poin
- Rank 4-5 = 5 poin
- Rank 6-10 = 3 poin

Dalam leaderboard, gunakan nama kelompok. Identifikasi dilakukan berdasarkan nama kelompok, jadi cukup 1 orang saja yang berada dalam tim Kaggle. Hasil prediksi di kaggle harus **reproducible**, sehingga notebook yang dikumpulkan harus bisa menghasilkan nilai akhir yang sama dengan submisi kaggle. Jika tidak sama, maka akan **didiskualifikasi** dari leaderboard. **Model yang boleh digunakan hanya KNN, Naive Bayes, dan ID3 yang diimplementasi from scratch.**

Kelompok

Pembagian kelompok ditentukan sendiri oleh mahasiswa dengan mengisi sheets kelompok berikut ini dengan 1 kelompok terdiri dari **1-3 kelompok Tugas Kecil 2** dengan **maksimal anggota sebanyak 5 orang**. Batas waktu pengisian kelompok adalah **Jumat, 22 November 2024 pukul 23:59 WIB**. Setelah waktu yang ditentukan, mahasiswa yang belum mengisi sheets kelompok akan diacak.

QnA

Pertanyaan dapat ditanyakan pada **link QnA** berikut. Pastikan pertanyaan yang ditanyakan tidak berulang.

Aturan

Terdapat beberapa hal yang harus diperhatikan dalam pengerjaan tugas ini, yakni:

1. Jika terdapat hal yang tidak dimengerti, silahkan ajukan pertanyaan kepada asisten melalui **link QnA** yang telah diberikan di atas. Pertanyaan yang diajukan secara

personal ke asisten **tidak akan dijawab** untuk menghindari perbedaan informasi yang didapatkan oleh peserta kuliah.

2. Dilarang melakukan **plagiarisme, menggunakan AI dalam bentuk apapun untuk men-*generate* jawaban Anda, dan melakukan kerjasama antar kelompok**. Pelanggaran pada poin ini akan menyebabkan pemberian **nilai E** pada setiap anggota kelompok.

Deliverables

- Tugas dikumpulkan dalam bentuk link ke *repository* GitHub yang **minimal** berisi beberapa hal berikut (boleh ditambahkan jika dirasa perlu):
    - Folder **src**, digunakan untuk menyimpan source code
    - Folder **doc,** digunakan untuk menyimpan laporan dalam bentuk **.pdf** yang terdiri atas komponen berikut:
        - Cover
        - Penjelasan singkat implementasi KNN.
        - Penjelasan singkat implementasi Naive-Bayes.
        - Penjelasan singkat implementasi ID3.
        - Penjelasan tahap cleaning dan preprocessing yang dilakukan beserta dengan alasannya.
        - Perbandingan hasil prediksi dari algoritma yang diimplementasikan dengan hasil yang didapatkan dengan menggunakan pustaka. Jelaskan insight yang kalian dapatkan dari perbandingan tersebut.
            - Perbandingan hasil dapat menggunakan metrics yang sesuai dengan permasalahan yang ada.
        - Kontribusi setiap anggota dalam kelompok.
        - Referensi
    - **README.md**, yang berisi deskripsi singkat repository, cara setup dan run program, dan pembagian tugas tiap anggota kelompok.
- Pengumpulan dilakukan melalui form dengan tautan sebagai berikut.
- Batas akhir pengumpulan adalah hari **Minggu, 15 Desember 2024 23.59 WIB**. Tugas yang terlambat dikumpulkan tidak akan diterima.
- Pengumpulan dilakukan oleh NIM terkecil.

Referensi

- The UNSW-NB15 Dataset | UNSW Research
- UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set) | IEEE Conference Publication | IEEE Xplore

- K-Nearest Neighbor(KNN) Algorithm - GeeksforGeeks
- What Are Naïve Bayes Classifiers? | IBM
- Decision Trees: ID3 Algorithm Explained | Towards Data Science