Tugas Pemrograman 3 AI – K Nearest Neighbors

Badrus Shoolehk Al Ar Fanny(1301164131)
IF-40-04
badrussholehaxel@gmail.com

Diberikan file DataTrain_Tugas3_AI.csv berupa himpunan data berisi 800 data yang memiliki 5 atribut input (X1, X2, X3, X4, X5) dan 1 output yang memiliki 4 kelas / label (0, 1, 2, dan 3). Bangunlah sebuah sistem klasifikasi menggunakan metode k-Nearest Neighbors untuk menentukan kelas / label data testing dalam file DataTest_Tugas3_AI.csv. Sistem membaca masukan file DataTrain_Tugas3.csv dan DataTest_Tugas3_AI.csv dan mengeluarkan output berupa file TebakanTugas3.csv berupa satu kolom berisi 200 baris angka bernilai integer/bulat (0, 1, 2, atau 3) yang menyatakan kelas / label baris atau record yang bersesuaian pada file DataTest_Tugas3_AI.csv

1. Deskripsi Masalah

Dalam menentukan keputusan untuk memilih satu atau beberapa objek terdekat dari suatu tempat maka dibutuhkan informasi mengenai jarak dari suatu titik ke beberapa alternatif tempat pilihan tersebut. Informasi jarak yang valid akan menuntun keputusan ke arah yang benar. Pada umumnya pencarian jarak terdekat dilakukan dengan menggunakan pengukuran euclidean distance, namun pada kenyataannya sebuah objek di dunia nyata hanya bisa dicapai melalui jalan yang menghubungkannya dengan objek lainnya, sehingga jarak yang valid untuk mendapatkan objek terdekat seharusnya adalah jarak jalan atau network distance, bukan euclidean distance.

2. Rancangan Metode

Langkah-langkah:

1. Inisialisasi data train dan data test pada program
2. Tentukan jarak antara masing" data train dengan data test menggunakan euclidean atau manhattan distance
3. Sorting jarak-jarak yang sudah didapat dari terdekat (terkecil) ke terjauh
4. Pilih data yg telah disorting sebanyak "k" dari jarak terdekat ke jarak terjauh
5. Tentukan apakah data test atau classnya udah terisi dengan melihat data yang telah dipiih, jika data yang dipilih lebih banyak class yang sesua atau tidak

3. Analisa

- DataTrain_Tugas3_AI.csv berisikan 800 data dengan tiap datanya berisikan 7 kolom (Index, X1, X2, X3, X4, X5, Y(Kelas /Label))

- DataTest_Tugas3_AI.csv berisikan 800 data dengan tiap datanya berisikan 7 kolom (Index, X1, X2, X3, X4, X5, Y(Kelas / Label))
- Menebak Kelas / Label dari file DataTest_Tugas3_AI.csv dengan algoritma k-Nearest Neighbors
- Setelah melakukan validasi data (dengan hasil random index dari DataTrain_Tugas3_AI.csv) menggunakan **k fold cross validation dengan 8 fold,** akurasi dari suatu k mulai menurun / stagnan / mengeluarkan pattern perulangan setelah k > 20, dengan diketahuinya hal tersebut pengetesan sebaiknya dilakukan hanya hingga k=20 untuk meminimalisir running time validasi data.
- Terdapat **206 data dengan kelas 0**, **194 data dengan kelas 1**, **199 data dengan kelas 2**, dan **201 data dengan kelas 3** pada DataTrain_Tugas3_AI.csv
- Rumus Manhattan dan Euclidean memberikan hasil yang berbeda.

4. Strategi Penyelesaian:

Untuk menyelesaikan masalah yang diberikan dengan k-Nearest Neighbor, pertama adalah dengan menentukan nilai k yang dianggap terbaik untuk kasus yang diberikan, salah satu caranya dengan proses validation **Training K Fold,** Proses yang dilakukan adalah:

1. Ambil data dari DataTrain_Tugas3_AI.csv dan lakukan pengacakan urutan data

2. Bagi data yang teracak tersebut kedalam beberapa fold, dalam kasus ini fold yang digunakan adalah 4, setiap fold memiliki 100 data.

3. Lakukan proses k-nearest neighbors dengan 1 fold berperan sebagai dataTest dan fold lainnya sebagai dataTrain, Ulangi langkah ini hingga semua fold mendapatkan giliran menjadi dataTest.

4. Pilih nilai k yang akan digunakan dalam algoritma, dalam kasus ini algortima distance yang digunakan dalam proses k-Nearest Neighbors adalah Euclidian.

5. Hitung rata rata akurasi dari suatu k dan lakukan proses ke-3 dengan nilai k lainnya.

6. Dari hasil test validasi 8 fold dengan Euclidian, **nilai k terbaik yang didapatkan adalah 9 dengan akurasi 89%.**

7. Dengan telah didapatkannya nilai k yang dianggap paling optimum, implementasikan k-Nearest Neighbor kepada data test yang akan digunakan, menggunakan k = 9 dan rumus Euclidian.

$$\mathrm{dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$