TUGAS 3

Kecerdasan Buatan

K-Nearest Neighbor

DESKRIPSI MASALAH

Klasifikasi data menggunakan metode K-Nearest Neighbor. Diberikan file DataTrain\_Tugas3\_AI.csv berupa himpunan data berisi 800 data yang memiliki 5 atribut input (X1, X2, X3, X4, dan 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\_AI.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.

ANALISIS DAN STRATEGI

Pada deskripsi masalah kali ini mengklasifikasikan data menggunakan program dengan metode K-Nearest Neightbor. Terdapat file DataTrain\_Tugas3\_AI.csv dan DataTest\_Tugas3\_AI.csv. Proses pertama kali dibuat adalah system untuk membaca file tersebut dan mengeluarkan output berupa file .csv.

1. The Algorithm

Sistem K-Nearest Neightbor memiliki parameter K dengan nilai integer untuk menentukan berapa banyak yang akan di banding kan pada setiap data.

Contoh pelatihan {x𝑖, 𝑦𝑖}

x𝑖: representasi nilai atribut contoh

𝑦𝑖: label kelas

Titik pengujian x yang ingin kita klasifikasikan

Algoritma:

Hitung jarak x untuk setiap contoh pelatihan x𝑖, 𝐷 (x, x𝑖)

Pilih 𝑘 contoh terdekat x𝑖1… x𝑖𝑘 dan label mereka 𝑦𝑖1… 𝑦𝑖𝑘

Output kelas 𝑦∗ yang lebih sering di 𝑦𝑖1… 𝑦𝑖𝑘

Pada problem ini saya menggunakan Manhattan.

Manhattan Distance :

1. Choosing the Value of k

Nilai k memiliki efek yang kuat pada kinerja K-NN

- nilai besar -> semuanya diklasifikasikan sebagai kelas yang paling mungkin: P (y)

- nilai kecil -> sangat bervariasi, batas keputusan tidak stabil

* perubahan kecil pada set pelatihan -> perubahan besar dalam klasifikasi
* mempengaruhi "kelancaran" dari batas

Memilih nilai k

- menyisihkan sebagian dari data pelatihan (set validasi)

- Variasikan k, amati pelatihan -> kesalahan validasi

Untuk menentukan nilai K yang paling optimal diperlukan data validasi yang diambil dari data train. Data train dibagi menjadi 2, bagian pertama 300 sebagai data validasi dan bagian kedua 500 sebagai data train. Data pertama berisi 300\*800=240.000 data yang dijadikan sebagai data validasi terhadap data train yang 500. Data validasi berisi 150 baris pertama diambil dari data train no urut 1-150, 150 baris kedua diambil dari data train no urut 651-800. Dan sisanya dijadikan sebagai data train untuk validasi, jadi terdapat 500 baris untuk data train validasi. Berikut data hasil output terhadap nilai K dari 1-5.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K = 1 | K = 2 | K = 3 | K = 4 | K = 5 |
|  |  |  |  |  |
| Tingkat Accuracy = 82.67% | Tingkat Accuracy = 80% | Tingkat Accuracy = 85.33% | Tingkat Accuracy = 84.67% | Tingkat Accuracy = 84.33% |

Untuk hasil saat ini Tingkat Accuracy yang paling tinggi di antara K 1-5 adalah K3 dengan Tingkat Accuracy 85.33%. Berdasarakn data validasi yang didapatkan sehingga kami menarik kesimpulan untuk menggunakan nilai K=3