KNN

algoritma klasifikasi yang bekerja dengan mengambil sejumlah k data terdekat (tetangganya) sebagai acuan untuk menentukan kelas dari data baru. algoritma ini mengklasifikasikan data berdasarkan kemiripan atau kedekatannya terhadap data lainnya

cara kerja knn:

- tentukan jumlah tetangga (k) yang akan digunakan untuk pertimbangan penentuan kelas.
- hitung jarak euclidean dari jumlah k terdekat.
- ambil k terdekat ini sesuai dengan jarak euclidean yang dihitung.
- di antara k terdekat ini, hitunglah jumlah titik data dalam setiap kategori.
- tetapkan titik data baru ke dalam kategori yang jumlah tetangganya maksimum.

Kelebihan

algoritmanya yang sederhana dan mudah diimplementasikan. Selanjutnya, Anda tidak perlu membangun model, membuat beberapa parameter, atau membuat asumsi tambahan. Terakhir, algoritma K-nearest neighbor ini sangat serbaguna. Anda bisa menggunakannya untuk membuat klasifikasi, regresi, dan pencarian data.

Kelemahan

Algoritmanya bisa menjadi lebih lambat secara signifikan karena jumlah contoh atau prediksi variabel independennya meningkat. Selain itu, K-nearest neighbor juga selalu memerlukan penentuan nilai K yang mungkin kompleks untuk beberapa kasus. Biaya komputasinya juga tinggi karena harus menghitung jarak antara titik data dengan semua sampel yang tersebar di sekitarnya (neighbor).

Rumus

$$dis = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + \cdots}$$

contoh

Age	Income	Class
29	350	A
51	430	B
33	290	A
24	255	A
40	410	B
45	380	B
34	390	?

Penyelesaian:

Data 1

$dis = \sqrt{(34-29)^2 + (390-350)^2} = 40.31$

Data 2

$dis = \sqrt{(34-51)^2 + (390-430)^2} = 43.46$

Data 3

$dis = \sqrt{(34-33)^2 + (390-290)^2} = 100.01$

Data 4

$dis = \sqrt{(34-24)^2 + (390-255)^2} = 135.37$

Data 5

$dis = \sqrt{(34-40)^2 + (390-410)^2} = 20.88$

Data 6

$dis = \sqrt{(34-45)^2 + (390-380)^2} = 14.87$

Data	Age	Income	Jarak dengan data baru
6	45	380	14.87
5	40	410	20.88
1	29	350	40.31
2	51	430	43.46
3	33	290	100.01
4	24	255	135.37

Naives bayes

metode pemelajaran mesin yang memanfaatkan perhitungan probabilitas dan statistik yang menjelaskan tentang prediksi probabilitas pada masa depan berdasarkan pengalaman pada masa sebelumnya.

Kelebihan

Bisa dipakai untuk data kuantitatif maupun kualitatif

Tidak memerlukan jumlah data yang banyak

Tidak perlu melakukan data training yang banyak

Kelemahan

Tidak berlaku jika probabilitas kondisionalnya adalah nol, apabila nol maka probabilitas prediksi akan bernilai nol juga

Mengasumsikan variabel bebas

Rumus

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)}$$

Keterangan:
X = Sampel data yang memiliki class (label) yang tidak diketahui.
C = Hipotesis bahwa X adalah data class (label).
P(C) = Probabilitas hipotesis C.
P(X) = Peluang dari data sampel yang diamati (probabilitas C).
P(X|C) = Probabilitas berdasarkan kondisi pada hipotesis.

Contoh

Warna	Tipe	Asal	Tercuri?
a1	a2	a4	vj
merah	Sport	domestik	ya
merah	Sport	domestik	tidak
merah	Sport	domestik	ya
kuning	Sport	domestik	tidak
kuning	Sport	import	ya
kuning	SUV	import	tidak
kuning	SUV	import	ya
kuning	SUV	domestik	tidak
merah	SUV	import	tidak
merah	Sport	import	ya

Terdapat studi case dengan mengelompokkan mobil warna merah, tipe SUV, dan asal domestik. Tentukan probabilitas tercuri dan probabilitas tidak tercuri, dan kemudian tentukan berapa persen mobil yang tercuri dan berapa persen mobil yang tidak tercuri, serta tentukan mobil dengan warna merah, tipe SUV, dan asal domestik tersebut tercuri atau tidak?

1. Menghitung Mobil tercuri YA & TIDAK

$$P(v_j) = \frac{N}{\text{jumlah}}$$

a. P (YA) = 5/10 = 0.5

b. P (tidak) = 5/10 = 0.5

2. Menghitung probabilitas ya tercuri pada mobil warna merah, tipe SUV, dan asal domestik

a. P (merah|ya) = 3/5 = 0.6

b. P (SUV|ya) = 1/5 = 0.2

c. P (domestik|ya) = 2/5 = 0.4

3. Menghitung probabilitas tidak tercuri pada mobil warna merah, tipe SUV, dan asal domestik

a. P (merah|tidak) = 2/5 = 0.4

b. P (SUV|tidak) = 3/5 = 0.6

c. P(domestik|tidak) = 3/5 = 0.6

4. menentukan berapa persen mobil tercuri dan tidak tercuri

a. tercuri (ya) = P (ya) * P (merah|ya) * P (SUV|ya) * P(domestik|ya)

= 0.5 * 0.6 * 0.2 * 0.4

= 0.024 atau 2,4 %

4. menentukan berapa persen mobil tercuri dan tidak tercuri

b. tercuri (tidak) = P (tidak) * P (merah|tidak) * P (SUV|tidak) *P (domestik|tidak)

= 0.5 * 0.4 * 0.6 * 0.6

= 0.072 atau 7,2%

Jadi, berdasarkan hasil perhitungan tercuri di atas dengan hasil tercuri (tidak) > tercuri (ya) yaitu 7.2% > 2.4% maka dapat disimpulkan mobil dengan warna merah, tipe SUV, dan asal domestik TIDAK TERCURI

Contoh lain:

suatu daerah dengan harga tanah mahal, jarak dari pusat kota sedang dan ada angkutan umum. Maka tentukan apakah daerah tersebut dipilih untuk mendirikan perumahan?

Aturan ke-	Harga tanah (C1)	Jarak dari pusat kota (C2)	Ada angkutan umum (C3)	Dipilih untuk Perumahan (C4)
1	Murah	Dekat	Tidak	Iya
2	Sedang	Dekat	Tidak	Iya
3	Mahal	Dekat	Tidak	Iya
4	Mahal	Jauh	Tidak	Tidak
5	Mahal	Sedang	Tidak	Tidak
6	Sedang	Jauh	Ada	Tidak
7	Murah	Jauh	Ada	Tidak
8	Murah	Sedang	Tidak	Iya
9	Mahal	Jauh	Ada	Tidak
10	Sedang	Sedang	Ada	Iya

Penyelesaian:

mencari probabilitas kemunculan setiap nilai untuk atribut (class).

Harga tanah	Jumlah kejadian 'dipilih' Iya	Jumlah kejadian 'dipilih' Tidak	Probabilitas Iya	Probabilitas Tidak
Murah	2	1	2/5	1/5
Sedang	2	1	2/5	1/5
Mahal	1	3	1/5	3/5
Jumlah	5	5	1	1

Probabilitas kemunculan setiap nilai untuk atribut Harga tanah (C1).

Jarak dari pusat kota	Jumlah kejadian 'dipilih' Iya	Jumlah kejadian 'dipilih' Tidak	Probabilitas Iya	Probabilitas Tidak
Dekat	3	0	3/5	0
Sedang	2	1	2/5	1/5
Jauh	0	4	0/5	4/5
Jumlah	5	5	1	1

Probabilitas kemunculan setiap nilai untuk atribut Jarak dari pusat kota (C2).

mencari probabilitas kemunculan setiap nilai untuk atribut (class).

Ada angkutan umum	Jumlah kejadian 'dipilih' Iya	Jumlah kejadian 'dipilih' Tidak	Probabilitas Iya	Probabilitas Tidak
Ada	1	3	1/5	3/5
Tidak	4	2	4/5	2/5
Jumlah	5	5	1	1

Probabilitas kemunculan setiap nilai untuk atribut Ada angkutan umum (C3).

Dipilih untuk perumahan	Jumlah kejadian 'dipilih' Iya	Jumlah kejadian 'dipilih' Tidak	Probabilitas Iya	Probabilitas Tidak
Jumlah	5	5	1/2	1/2

Probabilitas kemunculan setiap nilai untuk atribut Dipilih untuk perumahan (C4).

***Likelihood* Iya** = 1/5 * 2/5 * 1/5 * 5/10 = 1/125 = **0,008**

***Likelihood* Tidak** = 3/5 * 1/5 * 3/5 * 5/10 = 9/250 = **0,036**

Probabilitas Iya = 0,008 / (0,008+0,036) = 0,182

Probabilitas Tidak = 0,036 / (0,008+0,036) = 0,818

Jadi dapat disimpulkan dari hasil yang kita dapat diatas bahwa di lokasi tersebut tidak dibangun perumahan.

Decision tree(C4.5)

C4.5 merupakan perbaikan dari ID3 (Iterative Dichotomiser 3) • C4.5 dapat menangani data kontinu (angka, desimal) maupun diskrit (kategori / nominal) • C4.5 membuat pohon keputusan berdasarkan nilai entropy dan information gain. • Entropy adalah ukuran seberapa heterogen suatu data. Entropy is the measure of disorder. • Semakin bercampur-baur, entropi semakin tinggi. • Semakin "pure" ("murni") / homogen, entropi semakin kecil.

Contoh menghitung entropy

- Jika 50:50, maka p_1 = ½ dan p_2 = ½ , maka
- E = -1/2 * $\log_2$ (1/2) + (-1/2 * $\log_2$ (1/2)) = -1/2 * -1 + (-1/2 * -1)
- = 0.5 + 0.5 = 1

- Jika 100:0, maka p_1 = 1 dan p_2 = 0, maka
- E = -1 * $\log_2$ 1 + 0 * $\log_2$ 0 = (-1 * 0) + 0 = 0