

1. Deskripsi sistem

Dalam tugas pembelajaran mesin ini terdapat dua dataset yaitu testset dan train set, dalam test set terdapat 8 atribut yaitu age, workclass, education, marital status, relationship, occupation, house per week, dan income sedangkan dalam train set terdapat 7 atribut yaitu age, workclass, education, marital status, relationship, occupation, house per week. Maka dalam tugas ini dibuat sistem untuk melakukan klasifikasi untuk menentukan kelas pada testset dengan menggunakan naive bayes.

2. Analisis dan Implementasi

Naive Bayes merupakan salah satu metode machine learning yang menggunakan perhitungan probabilitas. Konsep dasar yang digunakan oleh Naive Bayes adalah Teorema Bayes, yaitu teorema dalam statistika untuk menghitung peluang, Bayes Optimal Classifier menghitung peluang dari satu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal.

Implementasi dari algoritma naive bayes adalah :

1. Persamaan dari teorema naive bayes

$$P(C|X) = P(X|C) P(C)$$

Keterangan :

x : Data dengan class yang belum diketahui

c : Hipotesis data merupakan suatu class spesifik

$P(c|x)$: Probabilitas hipotesis berdasar kondisi

$P(c)$: Probabilitas hipotesis

$P(x|c)$: Probabilitas berdasarkan kondisi pada hipotesis

2. Menghitung peluang jumlah kelas

Dalam tugas ini machine learning ini, pada file trainset terdapat atribut income yang mempunyai dua kelas yaitu, yaitu >50 dan $\leq 50k$. Hitung jumlah kelas yang $>50k$ dibagi dengan jumlah keseluruhan data.

3. Menghitung jumlah kasus perkelas

Dalam tahap ini, akan dilakukan perhitungan per kelas/atribut, kita dapat mencari $P(X|C_i)$ hitung seluruh atribut dengan kelasnya masing masing, contoh dalam testset terdapat atribut occupation yang memiliki kelas craft-repair, exec-managerial dan prof specialty, lalu bandingkan satu satu antara atribut occupation =craft repair dengan atribut lainnya. lalu kita dapat mengetahui hasilnya, hasil probabilitasnya.

lalu kalikan hasil probabilitas dalam satu kelas occupation misalnya atau atribut lainnya yang kelasnya sama. Misal pada atribut occupation mempunyai 3 kelas kelas craft-repair,exec -managerial dan prof specialty setelah mendapatkan probabilitas maka kalikan hasilnya

4. Melakukan Pengalian semua variable class

Setelah melakukan perhitungan di step sebelumnya. Maka pada tahap ini mengalikan hasil tahap 2 dan tahap 3 pada setiap kelas yang sama, melakukan perhitungan dengan rumus

$P(C|X) = P(X|C) P(C)$, setelah melakukan pengalian pada tahap sebelumnya, maka akan menghasilkan probabilitas di setiap kelas nya

5. Bandingkan probabilitas

Lakukan perbandingan, dan kelas yang lebih tinggi akan menjadi kelas testset

3. Hasil

Setelah melakukan implementasi pada tahap sebelumnya maka hasilnya yang didapat dalam trainset, pada atribut income yang mempunyai dua kelas >50 dan <=50, dan kelas >50 mempunyai 27 buah dari total seluruh nya dan ,kelas <= 50 mempunyai 13 buah dari total seluruhnya, berikut hasil prediksi pada testset. Dan berikut adalah hasil klasifikasi pada testset

1	<=50K	21	>50K
2	<=50K	22	>50K
3	>50K	23	>50K
4	<=50K	24	>50K
5	>50K	25	>50K
6	>50K	26	>50K
7	<=50K	27	>50K
8	<=50K	28	>50K
9	>50K	29	>50K
10	>50K	30	<=50K
11	>50K	31	<=50K
12	>50K	32	<=50K
13	<=50K	33	>50K
14	>50K	34	>50K
15	>50K	35	<=50K
16	>50K	36	>50K
17	<=50K	37	<=50K
18	>50K	38	>50K
19	<=50K	39	>50K
		40	>50K