Laporan Hasil Klasifikasi Data menggunakan metode Naïve Bayes

Deskripsi Masalah

Diberikan sebuah Trainset berupa himpunan data berisi 160 objek data yang memiliki 7 atribut input (**age**, **workclass**, **education**, **marital-status**, **occupation**, **relationship**, **hours-per-week**) dan 1 output (label kelas **income**) yang memiliki 2 kelas/label (>50K, dan <=50K). Bangunlah sebuah sistem klasifikasi menggunakan metode **Naïve Bayes** untuk menentukan kelas/label data testing dalam Testset. Sistem membaca masukan file TrainsetTugas1ML.csv dan TestsetTugas1ML.csv dan mengeluarkan *output* berupa file **TebakanTugas1ML.csv** berupa satu kolom berisi **40 baris** yang menyatakan kelas/label baris yang bersesuaian pada file TestsetTugas1ML.csv.

Metode Penyelesaian

Sejauh ini, kami mempelajari apa algoritma Naive Bayes, bagaimana teorema Bayes terkait dengannya, dan apa ekspresi teorema Bayes untuk algoritma ini. Mari kita ambil contoh sederhana untuk memahami fungsionalitas algoritma. Misalkan, kami memiliki himpunan data berisi 160 objek yang memiliki 7 atribut input dan 1 output yang memiliki 2 kelas.

Langkah 1 : Menggunakan rumus persamaan dari teorema Naïve 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 berdasarkan kondisi
P(c) : Probabilitas hipotesis
P(x|c) : Probabilitas berdasarkan kondisi pada hipotesis

Langkah 2 : Menghitung probabilitas hipotesis yang ada pada Income

Kelas **income** yang memiliki 2 kelas/label (>50K, dan <=50K). Hitung jumlah label >50K ada berapa dari 160 data, kita dapat sejumlah 120 dan <=50K kita dapat sejumlah 40.

Langkah 3 : Menghitung probabilitas berdasarkan kondiri pada hipotesis

Terdapat 7 atribut input, dari 7 atribut tersebut terdapat kelas/label lalu kita cari probabilitas kelas/label dari setiap atribut, lalu kita kalikan semua atribut input dengan output income. Hasil nilai akan didapatkan kisaran antara 0-1.

Langkah 4 : Pilih nilai yang tinggi dari hasil langkah 4

LAPORAN TUGAS 1 MACHINE LEARNING

Kita mengambil 40 data objek dari hasil langkah 4 untuk di jadikan sebagai data test berikut data test yang di ambil :

id	age	workclass	education	marital-status	occupation	relationship	hours-per-week
26027	young	Private	HS-grad	Never-married	Craft-repair	Not-in-family	normal
26314	young	Private	Bachelors	Divorced	Exec-managerial	Not-in-family	normal
31405	young	Private	Bachelors	Married-civ-spouse	Prof-specialty	Husband	normal
14736	adult	Private	Some-college	Divorced	Prof-specialty	Not-in-family	normal
27217	young	Private	HS-grad	Married-civ-spouse	Exec-managerial	Husband	many
5951	young	Private	Bachelors	Never-married	Prof-specialty	Not-in-family	normal
30067	young	Local-gov	Bachelors	Never-married	Craft-repair	Not-in-family	normal
28777	young	Self-emp-not-inc	Some-college	Never-married	Craft-repair	Not-in-family	normal
15390	adult	Private	Some-college	Married-civ-spouse	Craft-repair	Husband	normal
18042	young	Private	Some-college	Married-civ-spouse	Exec-managerial	Husband	normal
5793	adult	Local-gov	HS-grad	Married-civ-spouse	Exec-managerial	Husband	normal
31274	adult	Private	HS-grad	Married-civ-spouse	Craft-repair	Husband	normal
17068	young	Private	HS-grad	Married-civ-spouse	Craft-repair	Husband	low
21894	young	Private	Bachelors	Married-civ-spouse	Prof-specialty	Husband	normal
24128	adult	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	normal
8550	young	Private	Bachelors	Married-civ-spouse	Prof-specialty	Husband	normal
1181	young	Private	Bachelors	Divorced	Exec-managerial	Not-in-family	normal
11149	adult	Private	HS-grad	Married-civ-spouse	Craft-repair	Husband	normal
20836	young	Private	Some-college	Never-married	Prof-specialty	Not-in-family	normal
25766	adult	Local-gov	Bachelors	Married-civ-spouse	Prof-specialty	Husband	normal
139	adult	Private	Some-college	Married-civ-spouse	Craft-repair	Husband	normal
27160	young	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	normal
8814	young	Private	HS-grad	Married-civ-spouse	Exec-managerial	Husband	normal
11470	young	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	normal

Dari 160 data objek train kita mengambil 40 data untuk data test.

Langkah 5 : Menentukan output Income dari data test yang sudah di tentukan.

Berikut hasil akhir program yang di dapatkan dari data test.

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
20	>50K	40	>50K

Kita lihat pada hasil akhir program tersebut terdapat data kelas dari Income >50K lebih banyak dibandingkan <=50K. Hal ini di karenakan data train yang kurang stabil mengakibatkan data test yang tidak seimbang.