NAMA : Edy Santoso Kelas : IF 40-08 Nim : 1301162111

Laporan Klasifikasi income dengan menggunakan metode Naïve bayes

Deskripsi masalah

Terdapat 2 buah data yang diberikan yaitu data tarin dan data set. Data train sendiri terdiri dari 7 atribut input, yaitu age, workclass education, martial-status, occupation, relationship, hours-per-week) dan memiliki 1 atribut output yaitu kelas income, yang mana kelas income sendiri terdiri dari 2 kelas yaitu >50k dan =<50k. sedangkan untuk data test memiliki atribut yang sama, namun tidak ada atribut label/kelas. Untuk itu, diperlukan sebuah model/sistem klasifikai untuk menentukan kelas/label pada data test dengan menggunakan metode Naïve bayes.

Metode Penyelesaian

Algoritma Naive Bayes memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dr Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian.

Adapnu langkah langkah dari pengimplementasian algoritma Naïve Bayes adalah sebagai berikut :

1. Diketahui persamaan dari teorema naïve bayes adalah sebagi berikut:

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

   Keterangan :
   x : Data dengan class yang belum diketahui
   c : Hipotesis data merupakan suatu class spesifik
   P(c|x) : Probabilitas hipotesis berdasar kondisi (posteriori probability)
   P(c) : Probabilitas hipotesis (prior probability)
   P(x|c) : Probabilitas berdasarkan kondisi pada hipotesis

2. Hitung peluang jumlah kelas/label.
   Untuk kasus ini, terdapat 2 kelas, yaitu >50dan <=50k. hitung P(Ci) atau dapat dikatakan hitung jumlah kelas yang >50k dibagi dengan jumlah keseluruhan data.

3. Menghitung jumlah kasus perkelas
   Dilakukan penghitungan jumlah kasus perkelas atau apabila didalam teorema, kita mencari P(X|Ci).
   Hitung seluruh atribut dengan masing masing kelas, contohnya, kita mempunyai ddata test dengan atribut workclass=”private” dan atribut lainnya. bandingkan atribut workclass= “private” dan atribut lainnya dengan masing masing class. Steleah melakukan perbandingan, maka akan mendapat probabilistiknya,. Setelah mendapat probabilistiknya, lakukan pengalian dari setiap probabilistic yang kelasnya sama.

4. Kalikan semua variable class

Lakukan pengalian dari hasil perhitungan pada step 2 dan step 3 untuk setiap kelasnya. Seperti yang sudah ditulis sebelumnya, teorema bayes adalah sebagi berikut :

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

P(X|C) sudah didapatkan pada step 3, dan P(C) pada step satu, maka akan dikalikan hasil dari setiap probabilistiknya pada setiap kelas.

5. Bandingkan probabilistik pada step 4
   Setelah mendapatkan probabilistic dari setiap kelasnya pada step 4, makan alakukan perbandingan. Probabilitas kelas yang lebih tinggi, akan menjadi kelas data test yang dittest.

Hasil Pemrosesan

Berdasarkan hasil pemrosesan, dapat dikatakan, terjadi ketidakseimbangan pada data train, yang mana kelas dari income >50k lebih banyak dari income <=50k atau dalam perbandingan 75% dan 25%. Hal ini berakibat pada hasil prediksi data test. Dan berikut adalah hasil dari klasifikasi 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