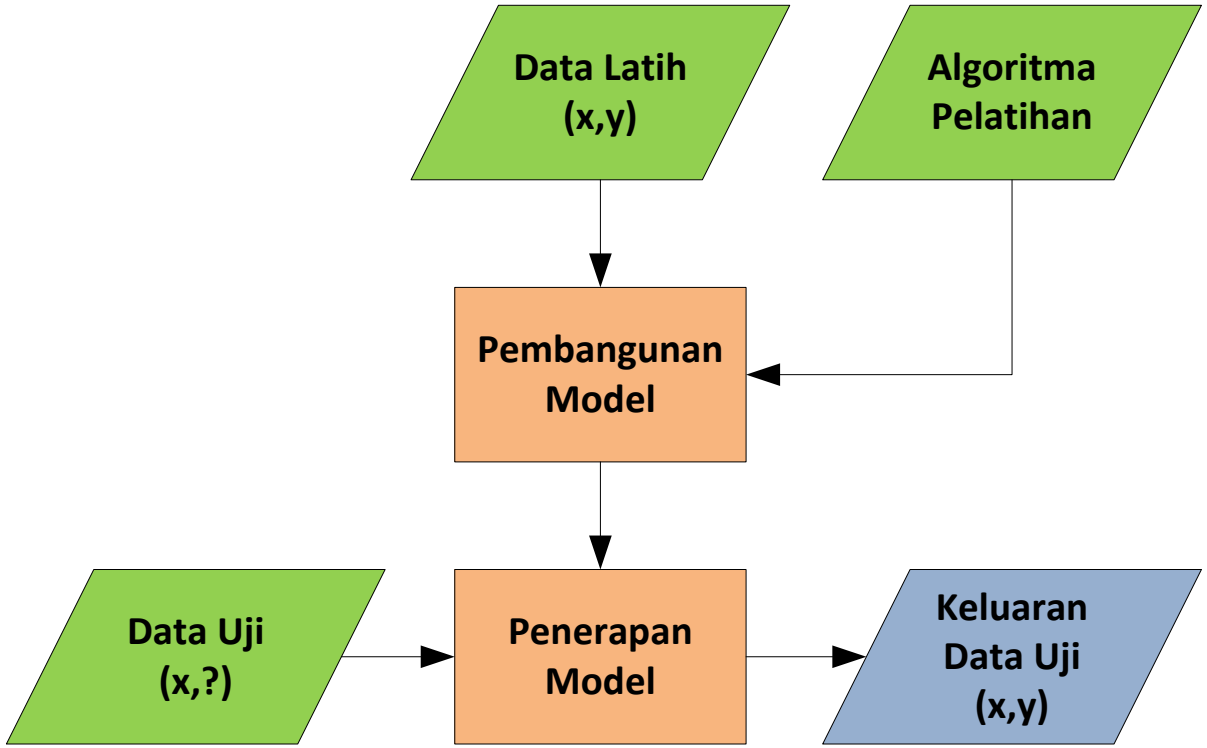
DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi

TEKNIK KLASIFIKASI –NAÏVE BAYES CLASSIFIER

Nizar Rabbi Radliya
nizar@email.unikom.ac.id

TEKNIK KLASIFIKASI



BAYESIAN CLASSIFICATION

Merupakan teknik prediksi berbasis probabilistik yang berdasar pada teorema Bayes.

$$\mathbf{P(H|E)} = \frac{P(E|H)xP(H)}{P(E)}$$ ← ***Formula teorema Bayes***

P(H|E) → Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis H terjadi, jika bukti (*evidence*) E terjadi.

P(E|H) → Probabilitas sebuah bukti E terjadi, akan mempengaruhi hipotesis H.

P(H) → Probabilitas awal (*priori*) hipotesis H terjadi, tanpa memandang bukti apapun.

P(E) → Probabilitas awal (*priori*) bukti E terjadi, tanpa memandang hipotesis/bukti yang lain.

NAÏVE BAYES CLASSIFIER

*Merupakan teknik klasifikasi berbasis probabilistic sederhana yang berdasar pada teorema Bayes dengan a*sumsi *independency* (ketidaktergantungan).

$$\mathbf{P(H|E_1, E_2, E_n)} = \frac{P(E_1,E_2,E_n|H)xP(H)}{P(E_1,E_2,E_n)}$$

Dalam prakteknya, probabilitas dari E_n tidak perlu diikutsertakan karena tidak tergantung pada hipotesis H yang akan diprediksi. Sehingga, formula di atas menjadi:

$$\mathbf{P(H|E_1, E_2, E_n)} = P(E_1, E_2, E_n|H)xP(H)$$

NAÏVE BAYES CLASSIFIER

1. Baca *data training*
2. Hitung jumlah *class*
3. Hitung jumlah kasus yang sama dengan *class* yang sama
4. Kalikan semua nilai hasil sesuai dengan data X yang dicari *class*-nya

CONTOH KASUS

1. Baca data training (data latih): Keputusan Membeli Komputer

Id	Age	Income	Student	Credit Rating	Buys Computer
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31…40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31…40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<=30	Medium	Yes	Excellent	Yes
12	31…40	Medium	No	Excellent	Yes
13	31…40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

CONTOH KASUS

2. Hitung jumlah class

Terdapat dua class yaitu:

C1 → buys computer = yes → 9 record

C2 → buys computer = no → 5 record

Total → 14 record

Menghitung probabilitas awal (*priori probability*) untuk setiap class:

P(C1) = P (buys computer = yes) = 9/14 = 0.643

P(C2) = P (buys computer = no) = 5/14 = 0.357

Buys Computer
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
No

CONTOH KASUS

3. Hitung jumlah kasus yang sama dengan class yang sama

Menghitung $P(E_1, E_2, E_n|C_i)$; untuk $i = 1$ dan $i = 2$

P(age = <=30 | buys computer = yes) = 2/9 = 0.222

P(age = <=30 | buys computer = no) = 3/5 = 0.600

CONTOH KASUS

3. Hitung jumlah kasus yang sama dengan class yang sama

Fitur	Value	Kasus C1	Kasus C2	P(E|C1)	P(E|C2)
Age	<=30	2	3	0.222	0.600
Age	31..40	4	0	0.444	0.000
Age	>40	3	2	0.333	0.400
Income	Low	3	1	0.333	0.200
Income	Medium	4	2	0.444	0.400
Income	High	2	2	0.222	0.400
Student	Yes	6	1	0.667	0.200
Student	No	3	4	0.333	0.800
Credit Rating	Fair	6	2	0.667	0.400
Credit Rating	Excellent	3	3	0.333	0.600

CONTOH KASUS

4. Kalikan semua nilai hasil sesuai dengan data X yang dicari class-nya

Fitur	Value	Kasus C1	Kasus C2	P(E|C1)	P(E|C2)
Age	<=30	2	3	0.222	0.600
Age	31..40	4	0	0.444	0.000
Age	>40	3	2	0.333	0.400
Income	Low	3	1	0.333	0.200
Income	Medium	4	2	0.444	0.400
Income	High	2	2	0.222	0.400
Student	Yes	6	1	0.667	0.200
Student	No	3	4	0.333	0.800
Credit Rating	Fair	6	2	0.667	0.400
Credit Rating	Excellent	3	3	0.333	0.600

Misal terdapat data X (belum diketahui *class*-nya)

X (age = <=30 ; income = medium ; student = yes ; credit rating = fair)

P(X | buys computer = yes) = (0.222 x 0.444 x 0.667 x 0.667) x 0.643 = **0.028**

P(X | buys computer = no) = (0.600 x 0.400 x 0.200 x 0.400) x 0.357 = 0.007

Kesimpulan: **buys computer = yes**

CONTOH KASUS

Data Latih: Keputusan Bermain Tenis

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Misal terdapat data X (belum diketahui *class*-nya)

X (outlook=rainy, temperature=hot, humidity=high, windy=true)

Main tenis atau tidak?

NEXT

TEKNIK CLUSTERING