

“NAÏVE BAYES CLASSIFIER”

Latar Belakang

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek. Klasifikasi merupakan proses awal dari pengelompokan data. Proses klasifikasi dokumen merupakan proses yang sangat penting dalam bidang sistem informasi, khususnya dalam proses penambangan data (data mining) untuk memperoleh pengetahuan bisnis (business knowledge).

Metode Bayes merupakan metode yang baik di dalam mesin pembelajaran berdasarkan data training, dengan menggunakan probabilitas bersyarat sebagai dasarnya. Karena metode lain tidak dapat digunakan untuk data yang tidak konsisten dan data yang bias, sehingga untuk bentuk data dengan jenis tertentu salah satu metode sederhana yang dapat digunakan adalah metode bayes.

Naïve Bayes dapat digunakan untuk berbagai macam keperluan antara lain untuk klasifikasi dokumen, deteksi spam atau *filtering spam*, dan masalah klasifikasi lainnya.

Ada beragam Teknik klasifikasi diantaranya adalah Naïve Bayes classifier, Decision Trees, dan Support Vector Machines.

Teorema Bayes

(Patil & Sherekar, 2013) Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas.

(Liantoni & Nugroho, 2015) Naïve Bayes Classifier adalah metode yang paling cepat dan sederhana. Hal ini berasal dari teorema Bayes dan hipotesis kebebasan, menghasilkan klasifier statistik berdasarkan peluang. Metode Bayes merupakan

pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi.

Persamaan dari teorema Bayes adalah:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)}$$

Dimana:

X: Data dengan class yang belum diketahui

H: Hipotesis data merupakan suatu class spesifik

$P(H|X)$: Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)

$P(H)$: Probabilitas hipotesis H (prior probabilitas)

$P(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$: Probabilitas X

Untuk menjelaskan metode Naïve Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode Naïve Bayes disesuaikan sebagai berikut:

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)}$$

Dimana variabel C merepresentasikan kelas, sementara variabel $F1 \dots Fn$ merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (*Posterior*) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (*likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global(*evidence*)

$$Posterior = \frac{prior \times likelihood}{evidence}$$

Nilai *Evidence* selalu tetap untuk setiap kelas pada suatu sampel. Nilai *posterior* tersebut nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan.

Teori lain mengatakan Teori Bayes adalah kesimpulan statistik yang membuktikan atau pengamatan yang digunakan untuk memperbarui atau menarik kesimpulan yang baru suatu probabilitas yang mungkin benar. Teori Bayes berasal dari kebiasaan menggunakan rumus bayes untuk memproses suatu kesimpulan/dugaan.

Keadaan Posteriror (Probabilitas X_k di dalam Y) dapat dihitung dari keadaan prior (Probabilitas Y di dalam X_k dibagi dengan jumlah dari semua probabilitas Y di dalam semua X_i).

Naïve Bayes untuk klasifikasi

Kaitan antara Naïve Bayes dengan Klasifikasi, korelasi hipotesis dan bukti klasifikasi adalah bahwa hipotesis dalam teorema Bayes merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadikan masukkan dalam model klasifikasi. Jika X adalah vector masukkan yang berisi fitur dan Y adalah label kelas, Naïve Bayes dituliskan dengan $P(X|Y)$. Notasi tersebut berarti probabilitas label kelas Y didapatkan setelah fitur-fitur X diamati. Notasi ini disebut juga probabilitas akhir (*posterior probability*) untuk Y , sedangkan $P(Y)$ disebut probabilitas awal (*prior probability*) Y .

Selama proses pelatihan dilakukan pembelajaran probabilitas akhir $P(Y|X)$ pada model untuk setiap kombinasi X dan Y berdasarkan informasi yang didapat dari data latih. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan nilai $P(X'|Y')$ yang didapat.

Formula Naïve Bayes untuk klasifikasi adalah :

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^q P(X_i|Y)}{P(X)}$$

$P(Y|X)$ adalah probabilitas data dengan vector X pada kelas Y . $P(Y)$ adalah probabilitas awal kelas Y . $\prod_{i=1}^q P(X_i|Y)$ adalah probabilitas independent kelas Y

dari semua fitur dalam vector X . Nilai $P(X)$ selalu tetap sehingga dalam perhitungan prediksi nantinya kita tinggal menghitung bagian $P(Y) \prod_{i=1}^q P(X_i|Y)$ dengan memilih yang terbesar sebagai kelas yang dipilih sebagai hasil prediksi.

Umumnya Naïve Bayes mudah dihitung untuk fitur bertipe kategoris seperti pada kasus klasifikasi hewan dengan fitur “penutup kulit dengan nilai {bulu,rambut,cangkang} atau kasus fitur “jenis kelamin” dengan nilai {pria,wanita}. Namun untuk fitur dengan tipe numerik (kontinu) ada perlakuan khusus sebelum dimasukkan dalam Naïve Bayes. Caranya adalah :

1. Melakukan diskretisasi pada setiap fitur kontinu dan mengganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasikan fitur kontinu ke dalam fitur ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data pelatihan.

HMAP (Hypothesis Maximum Appropri Probability)

HMAP (Hypothesis Maximum Appropri Probability) menyatakan hipotesa yang diambil berdasarkan nilai probabilitas berdasarkan kondisi prior yang diketahui. HMAP adalah model penyederhanaan dari metode bayes yang disebut dengan Naive Bayes. HMAP inilah yang digunakan di dalam machine learning sebagai metode untuk mendapatkan hipotesis untuk suatu keputusan.

$$\begin{aligned} P(S|X) &= \frac{\operatorname{argmax}_{x \in X} P(Y|X)P(X)}{P(X)} \\ &= \frac{\operatorname{argmax}_{x \in X} P(Y|X)P(X)}{P(X)} \end{aligned}$$

Karakteristik Naïve Bayes

Klasifikasi dengan Naïve Bayes bekerja berdasarkan teori probabilitas yang memandang semua fitur dari data sebagai bukti dalam probabilitas. Hal ini memberikan karakteristik Naïve Bayes sebagai berikut :

1. Metode Naïve Bayes bekerja teguh (robust) terhadap data-data yang terisolasi yang biasanya merupakan data dengan karakteristik berbeda. Naïve bayes juga

bisa menangani nilai atribut yang salah dengan mengabaikan data latih selama proses pembangunan model dan prediksi.

2. Tangguh menghadapi atribut yang tidak relevan.
3. Atribut yang mempunyai korelasi bisa mendegradasi kinerja klasifikasi Naïve Bayes karena asumsi independensi atribut tersebut sudah tidak ada.

Teorema Naïve Bayes memiliki beberapa kelebihan dan kekurangan yaitu sebagai berikut:

a. Keuntungan Naïve Bayes

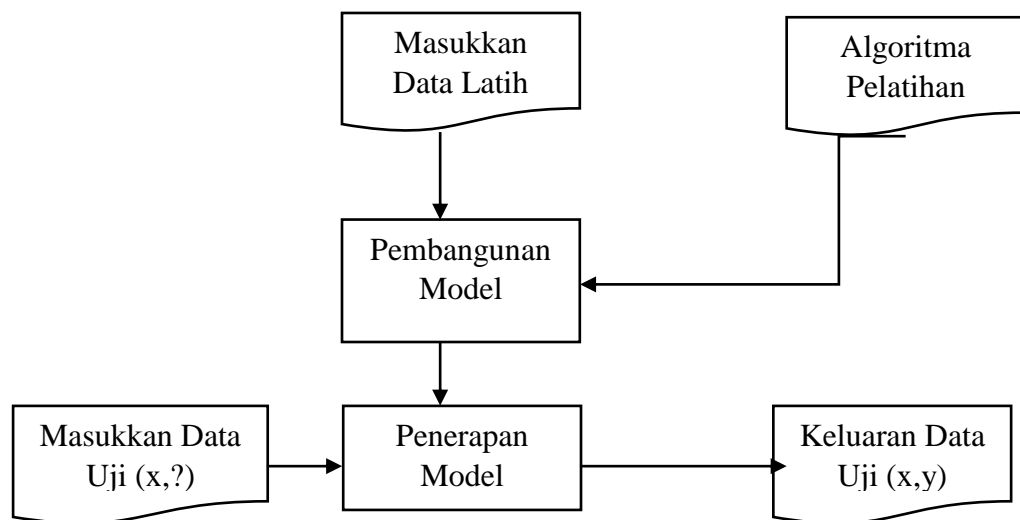
- 1) Menangani kuantitatif dan data diskrit.
- 2) Kokoh untuk titik noise yang diisolasi, misalkan titik yang di rata-rata kan ketika mengestimasi peluang bersyarat data.
- 3) Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata-rata dan variasi dari variable) yang dibutuhkan untuk klasifikasi.
- 4) Menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang.
- 5) Cepat dan efisiensi ruang.
- 6) Kokoh terhadap atribut yang tidak relevan.

b. Kekurangan Naïve Bayes

- 1) Tidak berlaku jika probabilitas kondisionalnya adalah nol, apabila nol maka probabilitas prediksi akan bernilai nol juga.
- 2) Mengasumsikan variable bebas.

Model Klasifikasi

Model dalam klasifikasi mempunyai arti yang sama dengan kotak hitam, dimana ada suatu model yang menerima masukan, kemudian mampu melakukan pemikiran terhadap masukan tersebut dan memberikan jawaban sebagai keluaran dari hasil pemikirannya. Kerangka kerja klasifikasi ditunjukkan pada gambar 2.1. pada gambar tersebut disediakan sejumlah data latih (x,y) untuk digunakan sebagai data pembangunan model. Model tersebut kemudian dipakai untuk memprediksi kelas dari data uji (x,y) sehingga diketahui kelas y yang sesungguhnya.



Gambar 2.1 Proses Klasifikasi

Model yang sudah dibangun pada saat pelatihan kemudian dapat digunakan untuk memprediksi label kelas baru yang belum diketahui. Dalam pembangunan model selama proses pelatihan tersebut diperlukan suatu algoritma untuk membangunnya, yang disebut algoritma pelatihan (*learning algorithm*). Ada banyak algoritma pelatihan yang sudah dikembangkan oleh para peneliti, seperti *K-Nearest Neighbor*, *Artificial Neural Network*, *Support Vector Machine* dan sebagainya. Setiap algoritma mempunyai kelebihan dan kekurangan, tetapi semua algoritma berprinsip sama, yaitu melakukan suatu pelatihan sehingga di akhir pelatihan, model dapat memetakan (memprediksi) setiap vector masukan ke label kelas keluaran dengan benar.

Pengukuran Kinerja Klasifikasi

Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja suatu sistem tidak bisa 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (*confuse matrix*).

Matrix konfusi merupakan tabel pencatat hasil kerja klasifikasi. Kuantitas matriks konfusi dapat diringkus jadi dua nilai, yaitu akurasi dan laju error. Dengan mengetahui jumlah data yang diklasifikasikan secara benar, kita dapat

mengetahui akurasi hasil prediksi data dengan mengetahui jumlah data yang diklasifikasikan secara salah, kita dapat mengetahui laju error dari prediksi yang dilakukan. Dua kuantitas ini digunakan sebagai matrik kinerja klasifikasi. Untuk menghitung akurasi digunakan formula:

$$\text{Akurasi} = \frac{\text{jumlah data yang diprediksi secara benar}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}}$$

Untuk menghitung laju *error* (kesalahan prediksi) digunakan formula:

$$\text{laju error} = \frac{\text{jumlah data yang diprediksi secara salah}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}}$$

semua algoritma klasifikasi berusaha membentuk model yang mempunyai akurasi tinggi atau (laju *error* yang rendah). Umumnya, model yang dibangun memprediksi dengan benar pada semua data yang menjadi data latihnya. Tetapi ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah algoritma klasifikasi ditentukan.

Contoh kasus Probabilitas Bersyarat

Misalkan dari 100 orang mahasiswa menunjukkan 20 orang mahasiswa menyukai keduanya, 30 orang mahasiswa menyukai bulu tangkis tapi tidak menyukai bola volley, 40 orang mahasiswa menyukai bola volley tapi tidak menyukai bulu tangkis, dan 10 orang mahasiswa tidak menyukai keduanya.

Dari data ini dapat disusun bentuk distribusi bersama sebagai berikut:

1. Tabel 3.1. Tabel probabilitas

Suka bulu tangkis (A)	Suka bola volley (B)		P(A)
	Ya	Tidak	
Ya	0.2	0.3	0.5
Tidak	0.4	0.1	0.5
P(B)	0.6	0.4	1

$$\text{Kejadian}(A,B) = \{(ya,ya), (ya,tidak), (tidak,ya), (tidak,tidak)\}$$

$$p(A=ya,B=ya) = 0.2$$

$$p(A=ya,B=tidak) = 0.3$$

$$p(A=tidak,B=ya) = 0.4$$

$$p(A=\text{tidak}, B=\text{tidak}) = 0.1$$

Probabilitas seseorang menyukai bulutangkis bila diketahui dia menyukai bola volley adalah:

$$P(A = ya | B = ya) = \frac{P(A = ya, B = ya)}{P(B = ya)} = \frac{0.2}{0.6} = 0.333$$

Probabilitas seseorang tidak menyukai bolavolley bila diketahui dia menyukai bola bulutangkis adalah:

$$P(B = \text{tidak} | A = ya) = \frac{P(A = ya, B = \text{tidak})}{P(A = ya)} = \frac{0.3}{0.5} = 0.6$$

Dari tabel tersebut dapat di simpulkan bahwa 0.2 suka keduanya, 0.3 suka bulu tangkis saja, 0.4 suka bola voley saja dan 0.1 tidak suka keduanya. Data tersebut dinamakan *Distribusi Bersama*. Sedangkan nilai 0.5 suka bulu tangkis dan 0.5 tidak suka bulu tangkis serta 0.6 suka bola voley dan 0.4 tidak suka bola voley adalah *Distribusi Marginal*.

Contoh Kasus HMAP

Diketahui hasil survey yang dilakukan sebuah lembaga kesehatan menyatakan bahwa 30% penduduk di dunia menderita sakit paru-paru. Dari 90% penduduk yang sakit paru-paru ini 60% adalah perokok, dan dari penduduk yang tidak menderita sakit paru-paru 20% perokok.

Fakta ini bisa didefinisikan dengan: X=sakit paru-paru dan Y=perokok.

Maka : $P(X) = 0.9$

$$P(\sim X) = 0.1$$

$$P(Y/X) = 0.6 \rightarrow P(\sim Y/X) = 0.4$$

$$P(Y/\sim X) = 0.2 \rightarrow P(\sim Y/\sim X) = 0.8$$

Dengan metode bayes dapat dihitung:

$$P(\{Y\}/X) = P(Y/X).P(X) = (0.6) . (0.9) = 0.54$$

$$P(\{Y\}/\sim X) = P(Y/\sim X) P(\sim X) = (0.2).(0.1) = 0.02$$

Bila diketahui seseorang merokok, maka dia menderita sakit paru-paru karena $P(\{Y\}|X)$ lebih besar dari $P(\{Y\}|\sim X)$. HMAP diartikan mencari probabilitas terbesar dari semua instance pada atribut target atau semua kemungkinan keputusan. Pada persoalan keputusan adalah sakit paru-paru atau tidak.

Contoh Kasus Probabilitas Pembelian Komputer

No	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
15	<=30	high	Yes	fair	No
16	>40	high	No	excellent	No

buys computer

No	age	income	student	credit_rating	buys_computer
1	31...40	high	no	fair	yes
2	>40	medium	no	fair	yes
3	>40	low	yes	fair	yes
4	31...40	low	yes	excellent	yes
5	<=30	low	yes	fair	yes
6	>40	medium	yes	fair	yes
7	<=30	medium	yes	excellent	yes
8	31...40	medium	no	excellent	yes
9	31...40	high	yes	fair	yes

no buys computer

No	age	income	student	credit_rating	buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	>40	low	yes	excellent	no
4	<=30	medium	no	fair	no
5	>40	medium	no	excellent	no

Probability

P(buy)	64%	
P(age=yes <=30)	22%	100%
P(age=yes 31-40)	44%	
P(age=yes >40)	33%	
P(income=yes low)	33%	100%
P(income=yes medium)	44%	
P(income=yes high)	22%	
P(student=yes no)	33%	100%
P(student=yes yes)	67%	
P(credit-rating=yes fair)	67%	100%
P(credit_rating=yes excellent)	33%	

probability

P(No)	36%	
P(age=no <=30)	60%	100%
P(age=no 31-40)	0%	
P(age=no >40)	40%	
P(income=no low)	20%	100%
P(income=no medium)	40%	
P(income=no high)	40%	
P(student=no no)	80%	100%
P(student=no yes)	20%	
P(credit-rating=no fair)	40%	100%
P(credit_rating=no excellent)	60%	

Prediksi

No	age	income	student	credit_rating	buys_computer
15	<=30	high	yes	fair	No
16	>40	high	no	excellent	No

maka nomor 15

peluang yes	0.02	22%
peluang no	0.08	78%

maka nomor 16

peluang yes	0.01	10%
peluang no	0.08	90%

DAFTAR PUSTAKA

- Liantoni, F., & Nugroho, H. (2015). KLASIFIKASI DAUN HERBAL MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER DAN K-NEAREST NEIGHBOR. *Jurnal Ilmiah SimanteC*, 9-16.
- Nataius, S. (2010/2011). Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen. *Makalah II 2092 Probabilitas dan Statistik-Sem.1*.
- Patil, T. R., & Sherekar, M. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal of Computer Science and Applications*, 256-261.