UNBIN
Universitas Binaniaga Indonesia

DATA MINING

(3 SKS)

Algoritma pada Supervised dan *Unsupervised learning*:

UNBIN
Universitas Binaniaga Indonesia

Supervised Learning	*Unsupervised Learning*
Beberapa algoritma yang termasuk dalam supervised learning: • Linear regression • Decision Tree dan Random Forest • Naive Bayes Classifier • Nearest Neighbor Classifier • Artificial Neural Network • Support Vector Machine	**Beberapa contoh algoritma unsupervised learning:** • K-means clustering • Hierarchical clustering • DBSCAN • Association Rule • Apriori algorithm

UNBIN
Universitas Binaniaga Indonesia

Metoda klasifikasi

Naive Bayes

??

Naive Bayes

- Metoda klasifikasi yang berakar pada **teorema Bayes** (Memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya)
- Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes
- Ciri utama adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian.

Bapak Thomas Bayes

Naive Bayes (lanjutan...)

- Olson Delen (2008) menjelaskan bahwa Naïve Bayes untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek
- Algoritma ini mengasumsikan bahwa atribut obyek adalah independen
- Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari ” master ” tabel keputusan.
- Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya

Kegunaan Naive Bayes

- Mengklasifikasikan dokumen teks seperti teks berita ataupun teks akademis
- Sebagai metode machine learning yang menggunakan probabilitas
- Untuk membuat diagnosis medis secara otomatis
- Mendeteksi atau menyaring spam

Kelebihan Naive Bayes

- Bisa dipakai untuk data kuantitatif maupun kualitatif
- Tidak memerlukan jumlah data yang banyak
- Tidak perlu melakukan data training yang banyak
- Jika ada nilai yang hilang, maka bisa diabaikan dalam perhitungan.
- Perhitungannya cepat dan efisien
- Mudah dipahami
- Mudah dibuat
- Pengklasifikasian dokumen bisa dipersonalisasi, disesuaikan dengan kebutuhan setiap orang
- Jika digunakan dalaam bahasa pemrograman, *code*-nya sederhana

Kekurangan Naive Bayes

- Apabila probabilitas kondisionalnya bernilai nol, maka probabilitas prediksi juga akan bernilai nol
- Asumsi bahwa masing-masing variabel independen membuat berkurangnya akurasi, karena biasanya ada korelasi antara variabel yang satu dengan variabel yang lain
- Keakuratannya tidak bisa diukur menggunakan satu probabilitas saja. Butuh bukti-bukti lain untuk membuktikannya.
- Untuk membuat keputusan, diperlukan pengetahuan awal atau pengetahuan mengenai masa sebelumnya. Keberhasilannya sangat bergantung pada pengetahuan awal tersebut Banyak celah yang bisa mengurangi efektivitasnya
- Dirancang untuk mendeteksi kata-kata saja, tidak bisa berupa gambar

Keuntungan Naive Bayes

- Algoritma ini bekerja sangat cepat dan dapat dengan mudah memprediksi kelas dari kumpulan data pengujian.
- Algoritma ini bisa digunakan untuk memecahkan masalah yang berhubungan dengan prediksi multi-kelas karena cukup berguna untuk menyelesaikannya
- Pengklasifikasi algoritma *naive bayes* berkinerja lebih baik daripada model lain dengan lebih sedikit training data jika asumsi independensi fitur berlaku
- Algoritma ini bekerja sangat baik dengan variabel input kategoris, dibandingkan dengan variabel numerik.

Tipe Algoritma Naive Bayes

UNBIN
Universitas Binaniaga Indonesia

Bernoulli Naive Bayes

- Prediktor adalah variabel Boolean
- Satu-satunya nilai yang ada adalah benar atau salah
- Algoritma ini digunakan ketika data sesuai dengan distribusi *bernoulli multivariat*

Naive Bayes Multinomial

- Untuk memecahkan masalah klasifikasi dokumen
- Contohnya, apabila ingin menentukan apakah suatu dokumen termasuk dalam suatu kategori, algoritma ini bisa digunakan untuk memilahnya
- *Naive bayes* menggunakan frekuensi kata-kata sekarang sebagai fitur

Gaussian Naive Bayes

- Digunakan apabila prediktor tidak diskrit namun memiliki nilai kontinu
- Prediktor diasumsikan sebagai sampel dari distribusi gaussian

Metode *Naive Bayes:*

- Merupakan metode pengklasifikasian yang sangat sederhana
- Dengan metode *Naive Bayes* terlebih dahulu mencari nilai probabilitas dan *likelihood* maksimum dari setiap atribut untuk masing-masing kelas
- **Persamaan dari probabilitas prior**

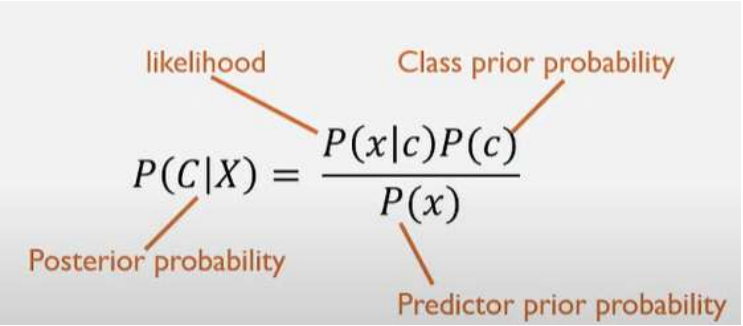
$$P(C) = \frac{Nj}{N}$$

Keterangan:

Nj : Jumlah data pada suatu class

N : Jumlah total data

Persamaan dari Teorema Bayes



Keterangan:

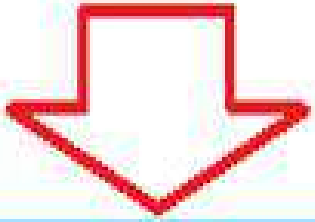
* X : Vektor input
* *c* : Sebuah *class* spesifik
* P(C|X) : Probabilitas class berasarkan vektor input yang diketahui *(Posteriori probability)*
* P(c) : Probabilitas *class* yang dicari
* P(x|c) : Probabilitas tiap input berdasarkan kondisi pada *class*
* P(x) : Probabilitas suatu input dari keseluruhan data

UNBIN
Universitas Binaniaga Indonesia

Metode *Naive Bayes*…

- Penentuan class dilakukan dengan cara membandingkan nilai probabilitas suatu sampel yang berada di class yang satu dengan nilai probabilitas suatu sampel yang berada di class yang lain
- Untuk menentukan class yang cocok dari suatu sampel dilakukan dengan cara membandingkan nilai posterior untuk masing-masing class, dan mengambil class dengan nilai posterior yang ***tertinggi***
- Tahapan Algoritma penyelesaian dengan Metode *Naive Bayes*

- Tahapan Algoritma penyelesaian dengan Metode *Naive Bayes*:



1. Menghitung Nilai Peluang kasus baru dari setiap Hipotesa dengan Klas (Label) yang ada "$P(X_k|C_i)$"
2. Menghitung Nilai Akumulasi Peluang dari setiap Klas "$P(X|C_i)$"
3. Menghitung Nilai $P(X|C_i) \times P(C_i)$
4. Menentukan Klas dari kasus baru tersebut

Contoh Kasus Naive Bayes: Keputusan Bermain Sepak Bola

Budi Santosa. (2007)

"Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis"

Contoh Kasus Naive Bayes: Keputusan Bermain Sepak Bola

- Kasus ini yaitu keputusan untuk bermain sepak bola atau tidak
- Data training yang digunakan ada 14 data
- Data telah di klasifikasikan berdasarkan: *cuaca*, *temperatur*, *kelembaban dan angin* (setelah dilakukan pengambilan keputusan atau prediksi dalam pengklasifikasian akan menghasilkan **output main atau tidak**)
- Data training dapat dilihat pada tabel berikut :

Data Training Kasus Bermain Sepak Bola

Cuaca (X1)	Temperatur (X2)	Kelembaban (X3)	Angin (X4)	Main atau Tidak (Y)
Cerah	Panas	Tinggi	Kecil	Tidak
Cerah	Panas	Tinggi	Besar	Tidak
Mendung	Panas	Tinggi	Kecil	Ya
Hujan	Sedang	Tinggi	Kecil	Ya
Hujan	Dingin	Normal	Kecil	Ya
Hujan	Dingin	Normal	Besar	Tidak
Mendung	Dingin	Normal	Besar	Ya
Cerah	Sedang	Tinggi	Kecil	Tidak
Cerah	Dingin	Normal	Kecil	Ya
Hujan	Sedang	Normal	Kecil	Ya
Cerah	Sedang	Normal	Besar	Ya
Mendung	Sedang	Tinggi	Besar	Ya
Mendung	Panas	Normal	Kecil	Ya
Hujan	Sedang	Tinggi	Besar	Tidak

Data Sampel Kasus Bermain Sepak Bola

Cuaca (X1)	Temperatur (X2)	Kelembaban (X3)	Angin (X4)	Main atau Tidak (Y)
Cerah	Dingin	Tinggi	Besar	???

Cari Penyelesaiannya !!

Tahap-tahap Penyelesaian:

1. **Menghitung nilai P(Xk | Ci) untuk setiap class i**

P(Cuaca = "Cerah" | Keterangan = "Ya")

P(Cuaca = 2 / 9 = 0,22

P(Cuaca = "Cerah" | Keterangan = "Tidak")

P(Cuaca = 3 / 5 = 0,6

P(Temperatur = "Dingin" | Keterangan = "Ya")

P(Temperatur = 3 / 9 = 0,33

P(Temperatur = "Dingin" | Keterangan = "Tidak")

P(Temperatur = 1 / 5 = 0,2

P(Kelembaban = "Tinggi" | Keterangan = "Ya)

P(Kelembaban = 3 / 9 = 0,33

P(Kelembaban = "Tinggi" | Keterangan = "Tidak)

P(Kelembaban = 4 / 5 = 0,8

P(Angin = "Besar" | Keterangan = "Ya")

P(Angin = 3 / 9 = 0,33

P(Angin = "Besar" | Keterangan = "Tidak")

P(Angin = 3 / 5 = 0,6

2. **Menghitung nilai P(X|Ci) untuk setiap class (Label)**

P(X | Keterangan = "Ya")

= 0,22 x 0,33 x 0,33 x 0,33 = 0,007906

P(X | Keterangan = "Tidak")

= 0,6 x 0,2 x 0,8 x 0,6 = 0,0576

3. **Menghitung nilai P(X|Ci * P(Ci)**

(P (X|Keterangan = "Ya") x P(Keterangan = "Ya)

= 0,007906 x 9 / 14 = 0,005083

(P (X|Keterangan = "Tidak") x P(Keterangan = "Tidak")

= 0,0576 x 5 / 14 = 0,020571

4. Menentukan class dari kasus tersebut

Cuaca (X1)	Temperatur (X2)	Kelembaban (X3)	Angin (X4)	Main atau Tidak (Y)
Cerah	Dingin	Tinggi	Besar	Tidak

Untuk data Input diatas, ***Naive Bayes*** Menyimpulkan **"Tidak"**