

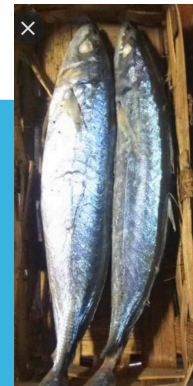
# KLASIFIKASI MENGUNAKAN NAÏVE BAYES

Margaretha Sulistyoningsih, Ph.D


Diadaptasi dari : Jiawei Han, UIUC CS412, Fall 2017 dan Huan Sun, CSE 5243 Intro to Data Mining, Classification (Basic Concepts & Advanced Methods), Ohio State University, undated.

# BAYESIAN CLASSIFICATION: WHY?

- Masih ingat klasifikasi jenis ikan?
  - Ikan yang warnanya hitam, tidak bersisik, berkumis adalah lele. Pasti.
  - Yang memiliki ciri : bersisik, warna merah, lebar, tidak berkumis adalah ikan nila. Pasti, bukan kadang nila kadang lele.



# BAYESIAN CLASSIFICATION: WHY?

- Sekarang bagaimana dengan klasifikasi yang perlu menggunakan probabilitas karena jawabannya tidak pasti? Misalnya:
    1. Seorang mahasiswa bisa saja mau membeli komputer yang anda jual. Sementara mahasiswa lainnya belum tentu mau membeli komputer anda.
    2. Seorang wanita berusia 30an bisa saja menyukai busana casual , tetapi seorang wanita yang lain lagi menyukai busana resmi atau ada yang menyukai kebaya.
- 

# BAYES' THEOREM

Diberikan data testing  $X$ , maka probabilitas (*posteriori probability*) dari sebuah hipotesa  $H$ ,  $P(H|X)$  adalah:

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

Secara informal dapat ditulis:

posteriori = likelihood x prior/evidence

# BAYES' THEOREM: BASICS

X adalah data (“evidence”) yang klasifikasinya tidak diketahui.

H adalah hipotesa bahwa X masuk klasifikasi C.

Klasifikasi adalah menentukan  $P(H|X)$  atau yang disebut dengan *posteriori probability* yaitu probabilitas dari sebuah hipotesa diberikan data X.

$P(H)$  (*prior probability*), adalah probabilitas awal

- Contoh: X akan membeli komputer, terlepas dari umur, income, dst...

$P(X)$ : Probabilitas dari data yang akan dicari klasifikasinya.

$P(X|H)$  (likelihood), the probabilitas data X jika diberikan given that the hipotesa H

Contoh., Diberikan Hipotesa “Yes”, X akan membeli komputer, Probabilitas bahwa X berusia 31..40, medium income, dst...



**NAÏVE BAYES**

# NAÏVE BAYES

Rumus Bayes adalah:

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

Karena  $P(\mathbf{X})$  konstan untuk semua kasus, maka yang perlu dimaksimalkan adalah:

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i) \quad \text{Inilah yang menjadi rumus Naïve Bayes.}$$

Mari belajar sambil mengerjakan contoh soal di slide selanjutnya ! 😊

## CONTOH SOAL :

### NAÏVE BAYES CLASSIFIER: TRAINING DATASET [1]

Ada 2 Klasifikasi:

C1:buys\_computer = 'yes'

C2:buys\_computer = 'no'

#### Soal:

Jika ada seseorang bernama Mr.X:

Umur = age  $\leq 30$ ,

Income = medium,

Student = yes

Credit\_rating = Fair,

age	income	student	credit_rating	buys_computer
$\leq 30$	high	no	fair	no
$\leq 30$	high	no	excellent	no
31...40	high	no	fair	yes
$> 40$	medium	no	fair	yes
$> 40$	low	yes	fair	yes
$> 40$	low	yes	excellent	no
31...40	low	yes	excellent	yes
$\leq 30$	medium	no	fair	no
$\leq 30$	low	yes	fair	yes
$> 40$	medium	yes	fair	yes
$\leq 30$	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
$> 40$	medium	no	excellent	no

Tentukan apakah yang bersangkutan akan membeli komputer atau tidak!



# CARA PENGGERJAAN DAN RUMUS

Rumus Naïve Bayes Classifier:

$$P(C_i | \mathbf{X}) = P(\mathbf{X} | C_i) P(C_i)$$

- Yang kita akan hitung adalah  $P(C_i | \mathbf{X})$ , yaitu Probabilitas klasifikasi  $C_i$ , (**buys\_computer**="yes" dan **buys\_computer**="no") jika diketahui ciri-ciri  $\mathbf{X}$  seperti pada soal, yaitu ciri-ciri  $\mathbf{X}$  : age ≤ 30, income = medium, dst.
- $P(C_i)$  adalah Probabilitas Klasifikasi dalam hal ini: Probabilitas dari **buys\_computer**="yes" dan probabilitas **buys\_computer** = "no".
- $P(\mathbf{X} | C_i)$ , yaitu Probabilitas Ciri  $\mathbf{X}$  jika diketahui Klasifikasi  $C_i$ .

$P(\mathbf{X} | C_i)$  dapat dihitung sebagai perkalian dari probabilitas dari setiap ciri  $\mathbf{X}$  yang klasifikasinya  $C_i$ , seperti yang tertera pada rumus di bawah:

$$P(\mathbf{X} | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \dots \times P(x_n | C_i)$$

Inilah yang mengakibatkan metode ini disebut Naïve. Karena dalam perkalian, kita menganggap bahwa tiap ciri terlepas (independen/bebas/tidak terikat) satu sama lain. Maka, ketika kita menghitung probabilitas suatu ciri terhadap klasifikasi  $C_i$ , kita tidak memperhatikan ciri lainnya.

# NAÏVE BAYES CLASSIFIER: JAWABAN

Rumus Naïve Bayes Classifier:

$$P(C_i | \mathbf{X}) = P(\mathbf{X} | C_i) P(C_i)$$

$P(C_i)$ : Probabilitas dari membeli/tidak membeli komputer =

$$P(\text{buys\_computer} = \text{"yes"}) = 9/14 = 0.643$$

$$P(\text{buys\_computer} = \text{"no"}) = 5/14 = 0.357$$

Hitung  $P(\mathbf{X} | C_i)$  untuk setiap kelas:

$P(\text{age} = \text{"<=30"} | \text{buys\_computer} = \text{"yes"}) = 2/9 = 0.222$ ,  
didapat dari ada 2 data orang yang umurnya  $\leq 30$   
**dan**  $\text{buys\_computer} = \text{"yes"}$ ,  
dibagi 9, yaitu semua data yang  $\text{buys\_computer} = \text{"yes"}$ .

$P(\text{age} = \text{"<= 30"} | \text{buys\_computer} = \text{"no"}) = 3/5 = 0.6$ , mengikuti cara diatas.  
Penyebutnya 5 dari jumlah "no" pada klasifikasi ( $\text{buys\_computer} = \text{"no"}$  ada 5).

$$P(\text{income} = \text{"medium"} | \text{buys\_computer} = \text{"yes"}) = 4/9 = 0.444$$

$$P(\text{income} = \text{"medium"} | \text{buys\_computer} = \text{"no"}) = 2/5 = 0.4$$

$$P(\text{student} = \text{"yes"} | \text{buys\_computer} = \text{"yes"}) = 6/9 = 0.667$$

$$P(\text{student} = \text{"yes"} | \text{buys\_computer} = \text{"no"}) = 1/5 = 0.2$$

$$P(\text{credit\_rating} = \text{"fair"} | \text{buys\_computer} = \text{"yes"}) = 6/9 = 0.667$$

$$P(\text{credit\_rating} = \text{"fair"} | \text{buys\_computer} = \text{"no"}) = 2/5 = 0.4$$

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

## LANJUTAN JAWABAN SOAL NAÏVE BAYES CLASSIFIER

**X = (age ≤ 30 , income = medium, student = yes, credit\_rating = fair)**

**$P(X|C_i)$  :**  $P(X|\text{buys\_computer} = \text{"yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$

$P(X|\text{buys\_computer} = \text{"no"}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$

**$P(X|C_i) * P(C_i)$  :**  $P(X|\text{buys\_computer} = \text{"yes"}) * P(\text{buys\_computer} = \text{"yes"})$   
 $= 0.044 * 0.643 = \underline{0.028292}$

$P(X|\text{buys\_computer} = \text{"no"}) * P(\text{buys\_computer} = \text{"no"})$   
 $= 0.019 * 0.357 = \underline{0.006783} = 0.007 \text{ (Pembulatan)}$

Karena probabilitas **"buys\_computer = yes"** lebih besar yaitu 0.028, dibanding probabilitas,  $P(X|\text{buys\_computer} = \text{"no"})$  yaitu 0.007, maka, Mr.X masuk klasifikasi akan membeli komputer.

# NAÏVE BAYES CLASSIFIER:

- Keuntungan
  - Mudah diimplementasikan.
  - Akurasi bagus dalam banyak kasus.
- Kerugian:
  - Asumsi bahwa : Masing-masing ciri tidak saling tergantung (independent) satu sama lain, sehingga tidak cocok untuk kasus kasus tertentu yang membutuhkan kebergantungan antar ciri dan kelas.
    - Contoh: hospitals: patients: Profile: age, family history, etc.
    - Gejala demam, batuk, dll, klasifikasi penyakit: kanker paru-paru, diabetes, dll. Kebergantungan seperti ini tidak dapat dimodelkan dengan Naïve Bayes. Metode yang cocok adalah: Bayesian Belief Network.

# LATIHAN MANDIRI:

Soal akan diberikan di kelas, dikumpulkan di SCE. Untuk nilai tugas semua mahasiswa.

## SOAL:

Data testing: Mr.X dengan ciri-ciri:

Age: >40

Income: High

Student: No

Credit\_rating: Excellent

Tentukan apakah Mr.X akan membeli komputer atau tidak!

# REFERENCES

1. Jiawei Han, UIUC CS412, Fall 2017
2. Huan Sun, CSE 5243 Intro to Data Mining, Classification (Basic Concepts & Advanced Methods), Ohio State University, undated.