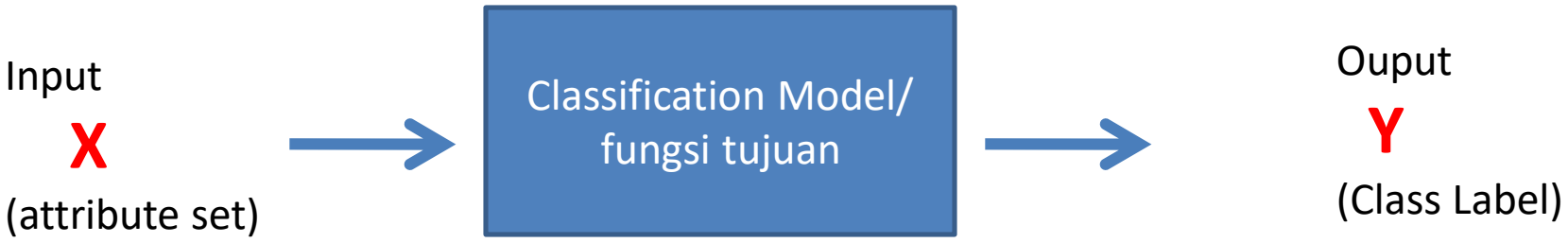
Naïve Bayes

Classification

Klasifikasi

- Merupakan proses pembelajaran suatu fungsi tujuan yang **memetakan** tiap himpunan **atribut x** ke satu dari **label kelas y** yang didefinisikan sebelumnya.



Model Klasifikasi

- Pemodelan **Deskriptif**

  Model Klasifikasi yang dpt berfungsi sbg alat **penjelasan** untuk membedakan **obyek-obyek** dalam **kelas-kelas** yang berbeda.

- Pemodelan **Prediktif**

  Model klasifikasi yang dapat digunakan untuk **memprediksi label kelas** yang tidak diketahui pada suatu ***object*/*record***.

Klasifikasi memerlukan Training Set

- Klasifikasi adalah proses pembelajaran secara terbimbing (***supervised learning***)
- Untuk melakukan klasifikasi, dibutuhkan ***training set*** sebagai data pembelajaran
- Setiap **sampel** dari training set memiliki **atribut** dan ***class label***

Dua Tahapan Klasifikasi

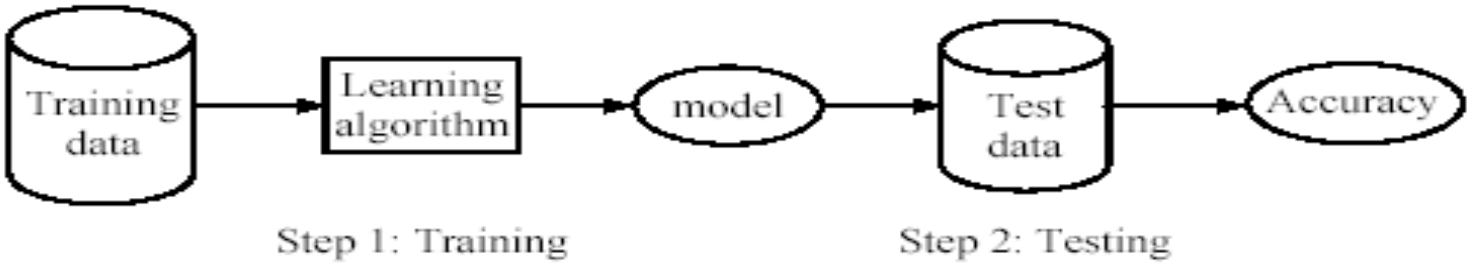
- ***Learning* (*training*):**

  Pembelajaran menggunakan **data training** (untuk ***Naïve Bayesian Classifier***, **nilai probabilitas** dihitung dalam proses pembelajaran)

- **Testing:**

  Menguji model menggunakan data testing

Sumber: Bing Liu, Web Data Mining



Akurasi

$$\text{Akurasi} = \frac{\text{Jml Prediksi Yang Benar}}{\text{Jml Prediksi keseluruhan}}$$

$$\text{Error} = \frac{\text{Jml Prediksi Yang Salah}}{\text{Jml Prediksi keseluruhan}}$$

Teori Bayesian: Sebagai Dasar

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

- **P(H|X)**, yaitu peluang hipotesa **H** berdasar kondisi **X**
- **X**:data sampel dengan klas (label) yang tidak diketahui
- **H**:merupakan hipotesa bahwa **X** adalah data dengan klas (label) C.
- **P(H)** : peluang dari hipotesa **H**
- **P(X)** adalah peluang dari **X** yang diamati
- **P(X|H)** : peluang **X**, berdasarkan kondisi pada hipotesa **H**

Naïve Bayesian Classifier

- Adalah metode classifier yang berdasarkan **probabilitas** dan Teorema Bayesian dengan asumsi bahwa setiap variabel **X** bersifat bebas(*independence)*
- Dengan kata lain, **Naïve Bayesian Classifier** mengasumsikan bahwa keberadaan sebuah atribut (variabel) tidak ada kaitannya dengan beradaan atribut (variabel) yang lain

Naïve Bayesian Classifier

- Karena asumsi **atribut tidak saling terkait**(*conditionally independent), maka:*

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

- Bila P(X|Ci) dapat diketahui melalui perhitungan di atas, maka klas (label) dari data sampel X adalah klas (label) yang memiliki *P(X|Ci)*P(Ci) maksimum*

Naïve Bayes

- Dataset

Umur	Pendapatan	Mhs	Rating Kredit	Beli Komputer
<=30	tinggi	bukan	fair	tdk
<=30	tinggi	bukan	excellent	tdk
30…40	tinggi	bukan	fair	ya
>40	sedang	bukan	fair	ya
>40	rendah	ya	fair	ya
>40	rendah	ya	excellent	tdk
31…40	rendah	ya	excellent	ya
<=30	sedang	bukan	fair	tdk
<=30	rendah	ya	fair	ya
>40	sedang	ya	fair	ya
<=30	sedang	ya	excellent	ya
31…40	sedang	bukan	excellent	ya
31…40	tinggi	ya	fair	ya
>40	sedang	bukan	excellent	tdk

class:
C1: Beli Komputer: **ya**
C2: Beli Komputer: **tdk**

bila data baru yg blm memiliki class sbb:
X =(umur<=30, pendapatan=sedang, mhs=ya,rating kredit= Fair)

Hitung P(Xk|Ci) utk setiap class i

- X =(**umur<=30**, pendapatan=sedang, mhs=ya, rating kredit= Fair)

Umur	Pendapatan	Mhs	Rating Kredit	Beli Komputer
<=30	tinggi	bukan	fair	tdk
<=30	tinggi	bukan	excellent	tdk
<=30	sedang	bukan	fair	tdk
<=30	rendah	ya	fair	ya
<=30	sedang	ya	excellent	ya
>40	rendah	ya	excellent	tdk
>40	sedang	bukan	fair	ya
>40	rendah	ya	fair	ya
>40	sedang	ya	fair	ya
>40	sedang	bukan	excellent	tdk
30…40	tinggi	bukan	fair	ya
31…40	rendah	ya	excellent	ya
31…40	sedang	bukan	excellent	ya
31…40	tinggi	ya	fair	ya

P(umur<=30| beli_komputer=ya) => 2/9 = 0.220

P(umur<=30| beli_komputer=tdk) => 3/5 = 0.600

Hitung P(Xk|Ci) utk setiap class i

- X =(umur<=30, **pendapatan=sedang**, mhs=ya,rating kredit= Fair)

ID	Umur	Pendapatan	Mhs	Rating Kredit	Beli Komputer
1	>40	rendah	ya	excellent	tdk
2	<=30	rendah	ya	fair	ya
3	>40	rendah	ya	fair	ya
4	31…40	rendah	ya	excellent	ya
5	<=30	sedang	bukan	fair	tdk
6	>40	sedang	bukan	excellent	tdk
7	<=30	sedang	ya	excellent	ya
8	>40	sedang	bukan	fair	ya
9	>40	sedang	ya	fair	ya
10	31…40	sedang	bukan	excellent	ya
11	<=30	tinggi	bukan	fair	tdk
12	<=30	tinggi	bukan	excellent	tdk
13	30…40	tinggi	bukan	fair	ya
14	31…40	tinggi	ya	fair	ya

P(pendapatan=sedang| beli_komputer=ya) => 4/9= 0.444
P(pendapatan=sedang| beli_komputer=tdk) => 2/5=0.400

Hitung P(Xk|Ci) utk setiap class i

X =(umur<=30, pendapatan=sedang, **mhs=ya**,rating kredit= Fair)

ID	Umur	Pendapatan	Mhs	Rating Kredit	Beli Komputer
1	<=30	sedang	bukan	fair	tdk
2	>40	sedang	bukan	excellent	tdk
3	<=30	tinggi	bukan	fair	tdk
4	<=30	tinggi	bukan	excellent	tdk
5	>40	sedang	bukan	fair	ya
6	31…40	sedang	bukan	excellent	ya
7	30…40	tinggi	bukan	fair	ya
8	>40	rendah	ya	excellent	tdk
9	<=30	rendah	ya	fair	ya
10	>40	rendah	ya	fair	ya
11	31…40	rendah	ya	excellent	ya
12	<=30	sedang	ya	excellent	ya
13	>40	sedang	ya	fair	ya
14	31…40	tinggi	ya	fair	ya

P(mhs=ya| beli_komputer=ya) => 6/9 = 0.670
P(mhs=ya| beli_komputer=tdk) => 1/5 = 0.200

Hitung P(Xk|Ci) utk setiap class i

X =(umur<=30, pendapatan=sedang, mhs=ya,**rating kredit= Fair**)

ID	Umur	Pendapatan	Mhs	Rating Kredit	Beli Komputer
1	<=30	tinggi	bukan	excellent	tdk
2	>40	sedang	bukan	excellent	tdk
3	>40	rendah	ya	excellent	tdk
4	31…40	sedang	bukan	excellent	ya
5	31…40	rendah	ya	excellent	ya
6	<=30	sedang	ya	excellent	ya
7	<=30	sedang	bukan	fair	tdk
8	<=30	tinggi	bukan	fair	tdk
9	>40	sedang	bukan	fair	ya
10	30…40	tinggi	bukan	fair	ya
11	<=30	rendah	ya	fair	ya
12	>40	rendah	ya	fair	ya
13	>40	sedang	ya	fair	ya
14	31…40	tinggi	ya	fair	ya

P(rating kredit=fair| beli_komputer=ya) => 6/9 = 0.670

P(rating kredit=fair| beli_komputer=tdk) => 2/5 = 0.400

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

Hitung P(Xk | Ci) utk setiap class I	
P(umur<=30| beli_komputer=ya) = 2/9	0.222
P(umur<=30| beli_komputer=tdk) = 3/5	0.600
P(pendapatan=sedang| beli_komputer=ya) = 4/9	0.444
P(pendapatan=sedang| beli_komputer=tdk) = 2/5	0.400
P(mhs=ya| beli_komputer=ya) = 6/9	0.667
P(mhs=ya| beli_komputer=tdk) = 1/5	0.200
P(rating kredit=fair| beli_komputer=ya) = 6/9	0.667
P(rating kredit=ya| beli_komputer=tdk) = 2/5	0.400

- Hitung P(X|Ci) untuk setiap Class:
  - P(X|beli_computer="ya")

    0.222 X 0.444 X 0.667 X 0.667 = 0.044
  - P(X|beli_computer="tdk")

    0.600 x 0.400 x 0.200 x 0.400 = 0.019

P(X|Ci)*P(Ci):

- P(X|beli_computer="ya") * P(beli_computer="ya")
  0.044 * (9/14) = 0.028
- P(X|beli_computer="tdk") * P(beli_computer="tdk")
  0.019 * (5/14) = 0.007

X memiliki class "beli_computer=ya"

karena

P(X|beli_computer="ya") memiliki nilai maksimum pada perhitungan di atas

Naïve Bayesian: Summary

- Kekuatan:
  - Mudah diimplementasi
  - Memberikan hasil yang baik untuk banyak kasus
- Kelemahan:
  - Harus mengasumsi bahwa antar fitur tidak terkait (*independent) Dalam realita,* keterkaitan itu ada
  - Keterkaitan tersebut tidak dapat dimodelkan oleh Naïve Bayesian Classifier

Latihan

ID	OUTLOOK	TEMPERATUR	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

Tentukan *class label* dari X:

X =(Outlook=Rainy,Temperature=Cool, Humidity=High, Windy=False)