Summary Machine Learning

Klasifikasi II (Naïve Bayes & Bayesian Network)

Muh.Ikhsan (H071191049)

A. Naïve Bayes

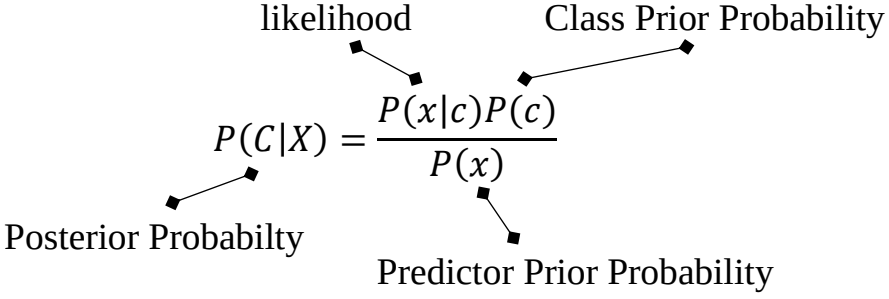
Algoritma Naive Bayes merupakan sebuah metoda klasifikasi menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes. Algoritma Naive Bayes memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dr Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian.

Keuntungan penggunan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yg diasumsikan sebagai variabel independent, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

1. Tahapan algoritma Naïve Bayes
   a. Menghitung Jumlah kelas / label.
   b. Menghitung Jumlah Kasus Per Kelas
   c. Kalikan Semua Variable Kelas
   d. Bandingkan Hasil Per Kelas

2. Kelebihan dan Kekurangan Naïve Bayes
   a. Kelebihan
      1) Mudah untuk dibuat
      2) Hasil bagus
   b. Kekurangan
      1) Asumsi independence antar atribut membuat akurasi berkurang (karena biasanya ada keterkaitan)

3. Persamaan Teorema Bayes



Keterangan :
$\boldsymbol{x}$: Data dengan class yang belum diketahui
$\boldsymbol{c}$: Hipotesis data merupakan suatu class spesifik
$\boldsymbol{P(c|x)}$: Probabilitas hipotesis berdasar kondisi (posteriori probability)
$\boldsymbol{P(c)}$: Probabilitas hipotesis (prior probability)
$\boldsymbol{P(x|c)}$: Probabilitas berdasarkan kondisi pada hipotesis
$\boldsymbol{P(x)}$: Probabilitas c

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 (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik sampel secara global (disebut juga evidence). Karena itu, rumus tersebut dapat pula ditulis sebagai berikut :

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

Nilai Evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan $(c|x_1, \dots, x_n)$ menggunakan aturan perkalian sebagai berikut :

$$P(C|X_1, \dots, X_n) = P(C)P(X_1, \dots, X_n|C)$$
$$= P(C)P(X_1|c)(X_2, \dots, X_n|C, X_1)$$
$$= P(C)P(X_1|c)P(X_2|C, X_1)(X_3, \dots, X_n|C, X_1, X_2)$$
$$= P(C)P(X_1|c)P(X_2|C, X_1)P(X_3|C, X_1, X_2) \dots P(X_n|C, X_1, X_2, \dots, X_{n-1)}$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (naif), bahwa masing masing petunjuk saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(c|X_1, \dots, X_n) = P(C) \prod \prod_{i=1}^{n} P(X_i|C)$$

$$P(c|X) = P(x_1|c)P(x_2|c) \dots P(x_n|c)P(c)$$

Persamaan diatas merupakan model dari Teorema Naive Bayes yang selanjutnya akan digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinyu digunakan rumus Densitas Gauss :

$$P = (X_i = x_i|Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2 ij}}$$

Keterangan :
P : Peluang
X_i : Atribut ke i
x_i : Nilai atribut ke i
Y : Kelas yang dicari
y_j : Sub kelas Y yang dicari
μ : Mean, menyatakan rata rata dari seluruh atribut
σ : Deviasi standar, menyatakan varian dari seluruh atribut

Mean

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Deviasi Standar

$$\sigma \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2 \right]^{0.5}$$

4. Laplace Correction
   Laplace Correction (Laplacian Estimator) atau additive smoothing adalah suatu cara untuk menangani nilai probabilitas 0 (nol). Dari sekian banyak data di training set, pada setiap perhitungan datanya ditambah 1 (satu) dan tidak akan membuat perbedaan yang berarti pada estimasi probabilitas sehingga bisa menghindari kasus nilai probabilitas 0 (nol).

$$\rho_i = \frac{m_i + 1}{n + k}$$

Dimana nilai k adalah jumlah kelas atau bin dari atribut m_i

5. Contoh Perhitungan Naïve Bayes
   Berikut ini Tenis Dataset
   Berisikan informasi apakah permainan Tenis akan dilaksanakan berdasarkan kondisi cuaca.
   Attribut Meliput :
   - *Outlook* merupakan informasi keaadaan cuaca, apakah (Sunny/Hujan, Cloudy/Mendung, Rainy/Cerah)
   - *Temp* merupakan informasi tempratur, apakah (Hot/Panas, Mild/Sejuk, Cool/Dingin)
   - *Humidity* merupakan informasi kelembaban, apakah (High/Tinggi, Normal)
   - *Windy* merupakan informasi apakah berangin atau tidak

No	Outlook	Temp	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	No
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	Hot	Normal	False	Yes
13	Cloudy	Mild	High	True	Yes

   Misalnya terdapat data uji dengan kondisi sebagai berikut :
   Outlook : Sunny, **Temp** : Mild, **Humidity** : Normal, **Windy**: False

a. **Langkah 1 :** Menghitung jumlah kelas atau label
Terdapat 2 class dari data training tersebut yaitu :
$C_1\ (Class\ 1) \rightarrow Play = Yes \rightarrow 9\ record$
$C_2\ (Class\ 2) \rightarrow Play = No \rightarrow 4\ record$
$Total\ =\ 13\ record$
Maka :

$$P(C_1) = \frac{9}{13} = 0.6923$$

$$P(C_2) = \frac{4}{13} = 0.3077$$

Data X = {*Outlook* : Sunny, *Temp* : Mild, *Humidity* : Normal, *Windy*: False}
Main Tenis atau tidak ?

b. **Langkah 2 :** Menghitung jumlah kasus per kelas
Misal attribute Humidity
$P(Humidity = \text{“}High\text{”} \mid Play\ =\ \text{“}Yes\text{”})\ =\ 3/9 =\ 0.3333$
$P(Humidity = \text{“}High\text{”} \mid Play\ =\ \text{“}No\text{”})\ =\ 3/4 =\ 0.75$
$P(Humidity = \text{“}Normal\text{”} \mid Play\ =\ \text{“}Yes\text{”})\ =\ 6/9 =\ 0.67$
$P(Humidity = \text{“}Normal\text{”} \mid Play\ =\ \text{“}No\text{”})\ =\ 1/4 =\ 0.25$

Jika semua atribut dihitung maka didapat hasil akhirnya seperti berikut

Atribut	Parameter	No	Yes
Outlook	value=Rainy	0.25	0.33
Outlook	value=Cloudy	0	0.45
Outlook	value=Sunny	0.75	0.22
Temprature	value=Hot	0.5	0.22
Temprature	value=Mild	0.25	0.45
Temprature	value=Cool	0.25	0.33
Humidity	value=High	0.75	0.33
Humidity	value=Normal	0.25	0.67
Windy	value=True	0.5	0.33
Windy	value=False	0.5	0.67

c. **Langkah 3 :** Menghitung *posterior probability* untuk masing-masing kelas Yes dan No
Kalikan semua nilai hasil sesuai dengan data X yang dicari class-nya

$$P(Yes|X) = P(Sunny|Yes) \times P(Mild|Yes) \times P(Normal|Yes) \times P(False|Yes) \times P(Yes)$$

$$= 0.22 \times 0.45 \times 0.67 \times 0.67$$

$$= 0.044$$

$$P(No|X) = P(Sunny|No) \times P(Mild|No) \times P(Normal|No) \times P(False|No) \times P(No)$$
$$= 0.75 \times 0.25 \times 0.25 \times 0.5$$
$$= 0.023$$

d. **Langkah 4 :** Bandingkan Hasil
Berdasarkan hasil perhitungan diperoleh nilai Yes lebih besar dari nilai No maka class dari data X tersebut adalah **Yes,** Jadi kemungkinan A akan bermain Tenis pada hari itu.

B. Bayesian Network

Bayesian Network dikenal juga dengan sebutan *belief network* atau *probalibistic network*. Dalam penulisan ini untuk selajutnya akan digunakan penamaan Bayesian network yang disingkat dengan BN. Bayesian network merupakan probabilistic graphical model (PGM) dengan busur berarah yang digunakan untuk merepresentasikan pengetahuan tentang hubungan kebergantungan (dependency) atau kebebasan (independency) di antara variabel-variabel dari domain persoalan yang dimodelkan. BN terdiri dari dua bagian utama, yaitu bagian struktur graf dan himpunan parameter. Kedua bagian BN tersebut dijelaskan sebagai berikut:

- Struktur Graf
  Struktur graf BN disebut dengan directed acyclic graph (DAG) yaitu graf berarah tanpa siklus. DAG terdiri dari simpul dan busur. Simpul merepresentasikan variabel acak, dan busur merepresentasikan adanya hubungan kebergantungan langsung (dapat pula diinterpretasikan sebagai pengaruh (sebab akibat) langsung, di antara variabel yang dihubungkannya.

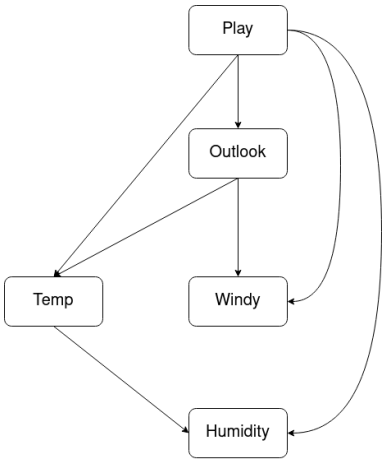
- Himpunan Parameter
  Himpunan parameter mendefinisikan distribusi probabilitas kondisional untuk setiap variabel. Setiap variabel acak direpresentasikan oleh sebuah simpul. Pada setiap simpul terdapat tabel yang berisikan distribusi probabilitas kondisional yang disebut dengan conditional probability table (CPT). Pada setiap sel dari tabel tersebut berisikan probabilitas kondisional dari nilai-nilai simpul yang diwakilinya jika diketahui setiap kombinasi nilai semua simpul parent kecuali pada akar (root).

1. Contoh Perhitungan Bayesian Network

No	Outlook	Temp	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	No
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	Hot	Normal	False	Yes
13	Cloudy	Mild	High	True	Yes

Misalnya terdapat data uji dengan kondisi sebagai berikut :
Outlook : Cloudy, **Temp** : Mild, **Humidity** : Normal, **Windy**: True

Berdasarkan dataset kita dapat membuat Graf Berarah Asiklik sebagai berikut :



a. **Langkah 1 :** Menghitung jumlah kelas atau label
Terdapat 2 class dari data training tersebut yaitu :
$C_1\ (Class\ 1) \rightarrow Play = Yes \rightarrow 9\ record$
$C_2\ (Class\ 2) \rightarrow Play = No \rightarrow 4\ record$
Total = 13 record
Maka :

$$P(C_1) = \frac{9}{13} = 0.6923$$

$$P(C_2) = \frac{4}{13} = 0.3077$$

Menerapkan koreksi Laplace

$$P(C_1) = \frac{9+1}{13+2} = 0.67$$

$$P(C_2) = \frac{4+1}{13+2} = 0.33$$

Data X = {Outlook : Cloudy, Temp : Mild, Humidity : Normal, Windy: True}
Main Tenis atau tidak ?

b. **Langkah 2 :** Menghitung jumlah kasus perkelas
Untuk setiap kategori dalam atribut, akan dilakukan koreksi Laplace agar nilai peluangnya $\neq 0$

1) Outlook

$$P(Outlook = Sunny|Play = Yes) = (2+1)(9+3) = 0.25$$
$$P(Outlook = Cloudy|Play = Yes) = (4+1)(9+3) = 0.42$$
$$P(Outlook = Rainy|Play = Yes) = (3+1)(9+3) = 0.33$$

$$P(Outlook = Sunny|Play = No) = (3+1)(4+3) = 0.57$$
$$P(Outlook = Cloudy|Play = No) = (0+1)(4+3) = 0.14$$
$$P(Outlook = Rainy|Play = No) = (1+1)(4+3) = 0.29$$

	Rainy	Cloudy	Sunny
$P(x_i\|Play = Yes)$	0.33	0.42	0.25
$P(x_i\|Play = No)$	0.29	0.14	0.57

2) Temp

	Hot	Mild	Cool
$P(x_i\|Play = Yes, Outlook = Rainy)$	0.167	0.5	0.333
$P(x_i\|Play = Yes, Outlook = Cloudy)$	0.429	0.286	0.286
$P(x_i\|Play = Yes, Outlook = Sunny)$	0.2	0.4	0.4
$P(x_i\|Play = No, Outlook = Rainy)$	0.25	0.25	0.5

$P(x_i \mid Play = No, Outlook = Cloudy)$	0.333	0.333	0.333
$P(x_i \mid Play = No, Outlook = Sunny)$	0.5	0.333	0.167

3) Humidity

	High	Normal
$P(x_i \mid Play = Yes, Temp = Hot)$	0.5	0.5
$P(x_i \mid Play = Yes, Temp = Mild)$	0.5	0.5
$P(x_i \mid Play = Yes, Temp = Cool)$	0.2	0.8
$P(x_i \mid Play = No, Temp = Hot)$	0.75	0.25
$P(x_i \mid Play = No, Temp = Mild)$	0.67	0.33
$P(x_i \mid Play = No, Temp = Cool)$	0.33	0.67

4) Windy

	True	False
$P(x_i \mid Play = Yes, Outlook = Rainy)$	0.2	0.8
$P(x_i \mid Play = Yes, Outlook = Cloudy)$	0.5	0.5
$P(x_i \mid Play = Yes, Outlook = Sunny)$	0.5	0.5
$P(x_i \mid Play = No, Outlook = Rainy)$	0.67	0.33
$P(x_i \mid Play = No, Outlook = Cloudy)$	0.5	0.5
$P(x_i \mid Play = No, Outlook = Sunny)$	0.4	0.6

c. **Langkah 3 :** Menghitung posterior probability untuk masing-masing kelas Yes dan No

$$P(play = Yes \mid Outlook = Cloudy, Temp = Mild, Humidity = Normal, Windy = True)$$
$$= \alpha \times P(play = Yes) \times P(Outlook = Cloudy \mid Play = Yes) \times P(Temp = Mild \mid Play = Yes, Outlook = Cloudy) \times P(Humidity = Normal \mid Play = Yes, Temp = Mild) \times P(Windy = True \mid Play = Yes, Outlook = Cloudy)$$
$$= \alpha \times 0.67 \times 0.42 \times 0.286 \times 0.5 \times 0.5$$
$$= \boldsymbol{\alpha \times 0.201}$$

$$P(play = No \mid Outlook = Cloudy, Temp = Mild, Humidity = Normal, Windy = True)$$

$$= \alpha \times P(play = No) \times P(Outlook = Cloudy|Play = No) \times P(Temp = Mild|Play = No, Outlook = Cloudy) \times P(Humidity = Normal|Play = No, Temp = Mild) \times P(Windy = True|Play = No, Outlook = Cloudy)$$

$$= \alpha \times 0.33 \times 0.14 \times 0.333 \times 0.33 \times 0.5$$

$$= \boldsymbol{\alpha \times 0.003}$$

e. **Langkah 4 :** Bandingkan Hasil
   Berdasarkan hasil perhitungan diperoleh nilai Yes lebih besar dari nilai No maka class dari data X tersebut adalah **Yes,** Jadi kemungkinan A akan bermain Tenis pada hari itu.