Inferensi Probabilistik

Kecerdasan Artifisial(CIF63310 / 2 sks)

Layout

- Inferensi Probabilistik
- Teori Bayesian
- Naïve Bayes
- Bayesian Belief Network

Inferensi Probalistik

Inferensi

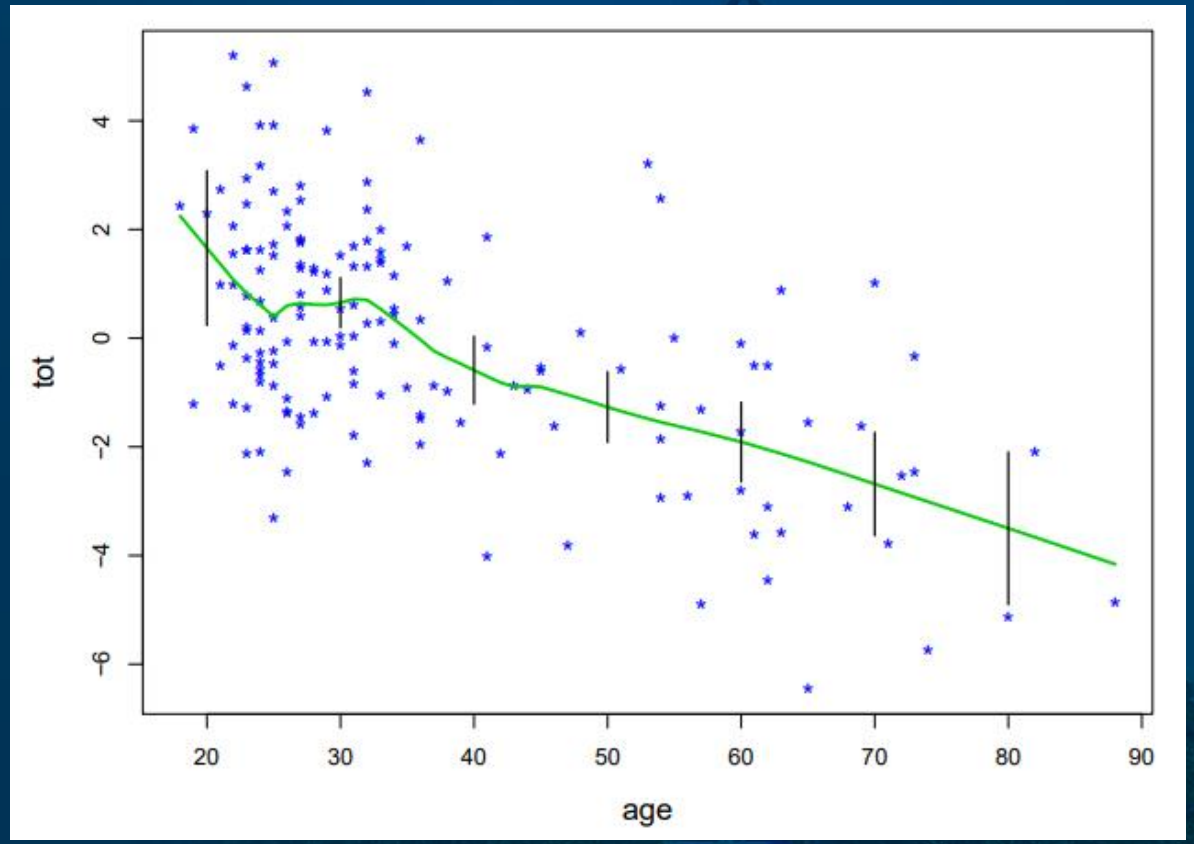
- Proses penarikan kesimpulan berdasarkan fakta yang dimiliki
- Inference system: sistem yang dapat melakukan proses reasoning/penalaran berdasarkan fakta/pengetahuan yang dimiliki
- Statistical inference system: proses penalaran yang menggunakan konsep-konsep dan teori statistika (frekuensi, peluang, rata-rata, distribusi, dll)

Statistical Inference system

- Frequentist Inference:
  - Kesimpulan didasarkan pada perhitungan frekuensi suatu kejadian yang dilakukan secara random/acak berulang-ulang dalam waktu yang lama.
  - Membutuhkan banyak random observation
- Bayesian Inference:
  - Kesimpulan didasarkan pada derajad kepercayaan terhadap sebuah kejadian

Frequentist Inference

- Contoh kasus: indeks kondisi kesehatan ginjal pasien(tot) berdasarkan umur (age).
- Prediksi tot pasien dapat dilakukan berdasarkan model yang dibangun dari dataset yang dikumpulkan dari banyak pasien sebelumnya.



Efron B. & Hastie T. Computer Age Statistical Inference, 2016, Cambridge University Press

Beberapa Metode Frequentist Inference

- Bootstrap:
  - Menambah data dengan car melakukan pemilihan sample secara acak dari dataset yang ada secara berulang-ulang, dari pada mencari data baru yang membutuhkan waktu/sumber daya yang banyak.
- Monte Carlo Simulation
  - Menggunakan probability model (Probability Distribution Function-PDF, Cumulative Distribution Function-CDF, atau model lainnya) yang ada untuk men-generate sample baru.
  - Perlu memodelkan kasus (model matematis atau algoritma) berdasaekan dataset yang dimiliki.
  - Bootsrap adalah kasus khusus dari monte carlo simulation

Beberapa Metode Frequentist Inference (2)

- Analisis Regresi
  - Membangun model berdasarkan hasil eksperimen

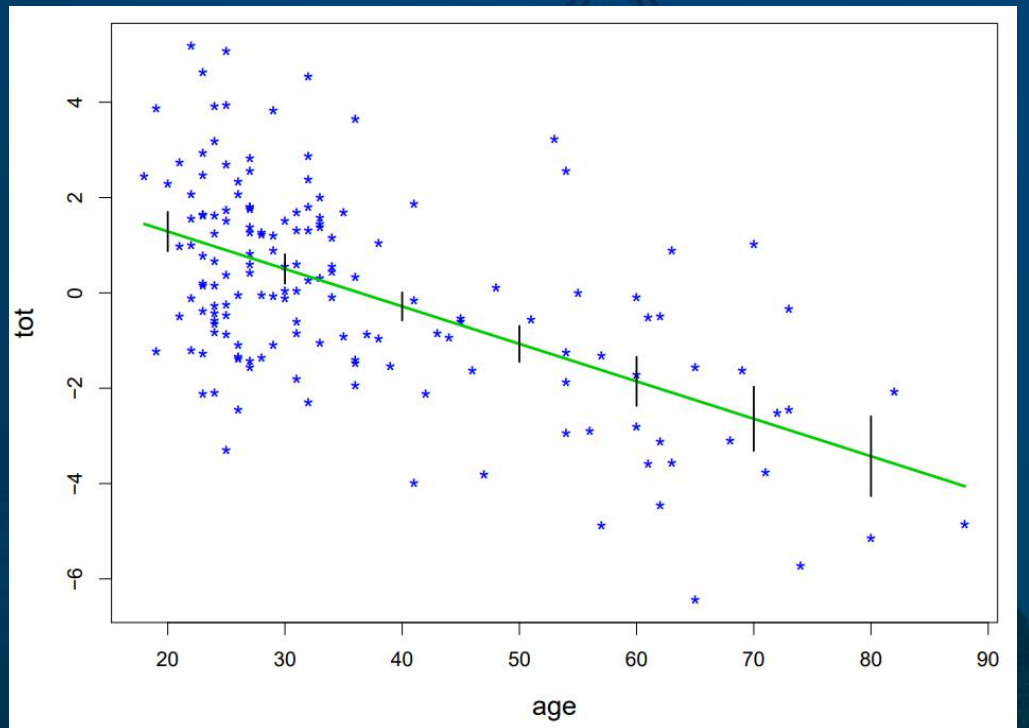


Figure 1.1 Kidney fitness `tot` vs `age` for 157 volunteers. The line is a linear regression fit, showing ±2 standard errors at selected values of `age`.

Beberapa Video Tutorial

- Bootstrap:
  - https://www.youtube.com/watch?v=Xz0x-8-cgaQ&t=38s
- Montecarlo:
  - https://www.youtube.com/watch?v=7ESK5SaP-bc
  - https://www.youtube.com/watch?v=EaR3C4e600k

Bayesian Inference

- Penalaran dengan menggunakan prior probability, likelihood, dan evidence
- Prior probability: Tingkat kepercayaan awal terhadap peluang terjadinya sebuah kejadian
- Likelihood: Kemungkinan sebuah kejadian data sample muncul di dalam sebuah populasi
- Evidence: fakta yang diketahui saat ini

Peluang kondisional (conditional probability):

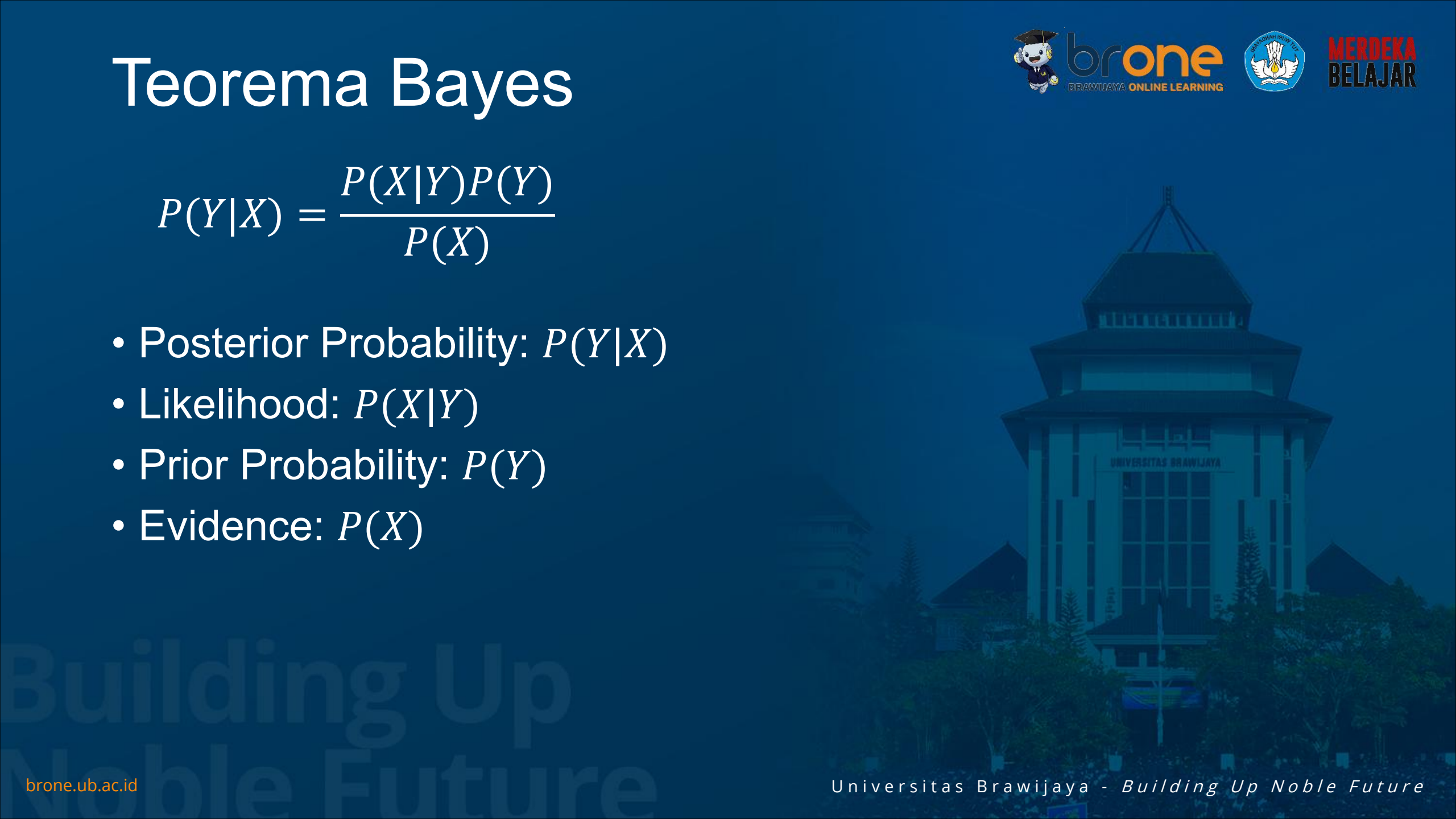
$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

Teorema Bayes

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

- Posterior Probability: $P(Y|X)$
- Likelihood: $P(X|Y)$
- Prior Probability: $P(Y)$
- Evidence: $P(X)$



Contoh:

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having a stiff neck is 1/20

- If a patient has a stiff neck, what's the probability he/she has meningitis?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Naïve Bayes

- Naïve Bayes merupakan bentuk khusus dari teorma bayes
- Dalam teorema bayes, sulit untuk menghitung likelihood pada data sample dengan banyak parameter.
- $P(X|Y) = P(x1,\ x2,\ ....,xn|Y)$

- Naïve Bayes menganggap setiap kejadian adalah kejadian yang saling lepas, sehingga likelihoodnya menjadi
- $P(X|Y) = P(x1|Y)\ P(x2|Y)..... P(xn|Y)$

Naïve Bayes Classification

- Misalkan ada sebuah data sampel (record) dengan atribut (x1, x2, …., xd)
- Goal: ingin memprediksi kelas untuk record tersebut
- Kelas Y: {y1, y2, .. yn)}
- Mencari nilai tertinggi dari peluang P(yi|x1, x2, …, xd)

Contoh Kasus

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Diketahui record yang ingin diklasifikasi adalah X={refund=No, deforced, income=120K}
  - x1: refund=No
  - x2: deforced
  - x3: income=120K
- Nilai Y={Evade=yes, Evade=No}
- Hitung posterior probability:
  - P(Evade=yes | X) = ….
  - P(Evade=No | X) = …

Likelihood X terhadap yes

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(yes|X) = \frac{P(X|yes)P(yes)}{P(X)}$
- Menghitung likelihood P(X | yes):
  - = P(x1,x2,x3 | yes)
  - = P(x1 | yes) P(x2 | yes) P(x3 | yes)
  - = P(refund=no | yes) * P(deforced | yes) * P(income = 120K | yes)
  - = 3/3 * 1/3 * 0/3
  - = 0

Likelihood X terhadap no

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(no|X) = \frac{P(X|no)P(no)}{P(X)}$
- Menghitung likelihood P(X | no):
  - = P(x1,x2,x3 | no)
  - = P(x1 | no) P(x2 | no) P(x3 | no)
  - = P(refund=no | no) *
    P(deforced | no) *
    P(income = 120K | no)
  - = 4/7 * 1/7 * 1/7
  - = 0.012

Prior yes dan no

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(yes) = \frac{3}{10}$
- $P(no) = \frac{7}{10}$

Evidence

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- P(X) = P(x1, x2, x3)
- = P(x1)P(x2)P(x3)
- = P(refund=no) * P(deforced) * P(income=120K)
- = 7/10 * 2/10 * 1/10
- = 0.014

asumsi varible saling lepas

Posteriror Probability

- $P(yes|X) = \frac{P(X|yes)P(yes)}{P(X)} = \frac{0*0.3}{0.014} = 0$
- $P(no|X) = \frac{P(X|no)P(no)}{P(X)} = \frac{0.012*0.7}{0.014} = 0.58$

- Karena P(no|x) > P(yes|X), maka kelas yang sesuai untuk data tersebut adalah evade = no
- Perhatikan nilai evidence P(X) pada kedua perhitungan di atas. Karena yang diperlukan untuk keputusan adalah posterior terbesar, dan nilain P(X) sama antara kedua nilai posterior, maka perhitungan evidence boleh diabaikan

Continue Probability

- Perhatikan nilai likelihood untuk kelas yes
  - = P(refund=no | yes) * P(deforced | yes) * P(income = 120K | yes)
  - = 3/3 * 1/3 * 0/3
  - = 0
- Nilainya = 0 kerena kolom income tidak ada yang tepat 120K
- Nilai income dalam dunia nyata selalu spesifik untuk tiap orang dan tidak bisa dihitung secara diskrit.
- Ada 2 teknik penyelesaiannya:
  - Discretization: buat interval pada nilai income
  - Menghitung peluang kontiny dengan probability density estimation

Probability Density Estimation

- Distribusi Normal (Normal Distribution)

$$P(X_i|Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

σ_{ij}^2: $sample\ variance$
μ_{ij}: $sample\ mean$

- Untuk (income, class = no):
  - Sample mean = 110
  - Sample variance = 2975

$$P(Income = 120|No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}}$$

$$= 0.0072$$

- Untuk (income, class = yes):
  - Sample mean = 90
  - Sample variance = 25

$$P(Income = 120|Yes) = \frac{1}{\sqrt{2\pi}(5)} e^{-\frac{(120-90)^2}{2(25)}}$$

$$= 1.2 \times 10^{-9}$$

Likelihood X terhadap yes (hitung ulang)

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(yes|X) = \frac{P(X|yes)P(yes)}{P(X)}$
- Menghitung likelihood P(X | yes):
  - = P(x1,x2,x3 | yes)
  - = P(x1 | yes) P(x2 | yes) P(x3 | yes)
  - = P(refund=no | yes) *
    P(deforced | yes) *
    P(income = 120K | yes)
  - = 3/3 * 1/3 * 1.2 x 10^{-9}
  - = 4 x 10^{-10}

Likelihood X terhadap no

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(no|X) = \frac{P(X|no)P(no)}{P(X)}$
- Menghitung likelihood P(X | no):
  - = P(x1,x2,x3 | no)
  - = P(x1 | no) P(x2 | no) P(x3 | no)
  - = P(refund=no | no) *
    P(deforced | no) *
    P(income = 120K | no)
  - = 4/7 * 1/7 * 0.0072
  - = 0.006

Posteriror Probability (hitung ulang)

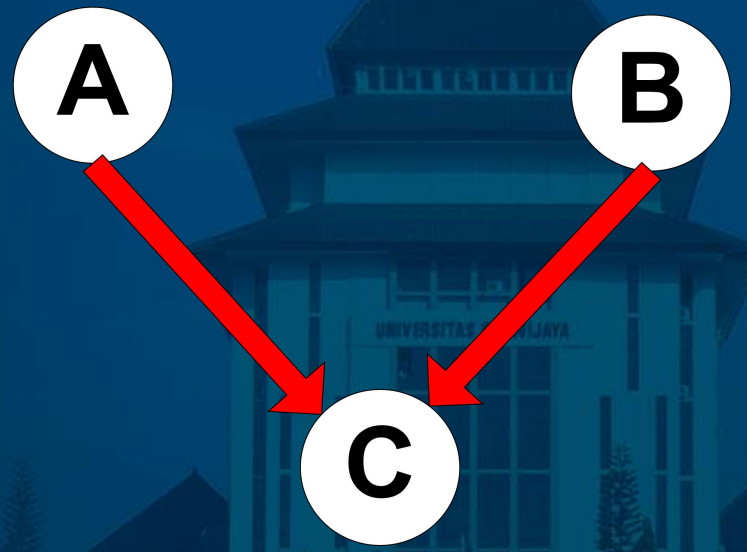
- $P(yes|X) = \frac{P(X|yes)P(yes)}{P(X)} = \frac{4\times10^{-10}*0.3}{0.014} = 8.5 \times 10^{-9}$
- $P(no|X) = \frac{P(X|no)P(no)}{P(X)} = \frac{0.006*0.7}{0.014} = 0.3$

- Karena P(no|x) > P(yes|X), maka kelas yang sesuai untuk data tersebut adalah evade = no
- Perhatikan nilai evidence P(X) pada kedua perhitungan di atas. Karena yang diperlukan untuk keputusan adalah posterior terbesar, dan nilain P(X) sama antara kedua nilai posterior, maka perhitungan evidence boleh diabaikan

Bayesian Belief Network

- Representasi grafis dari hubungan probabilitas antar radom variable

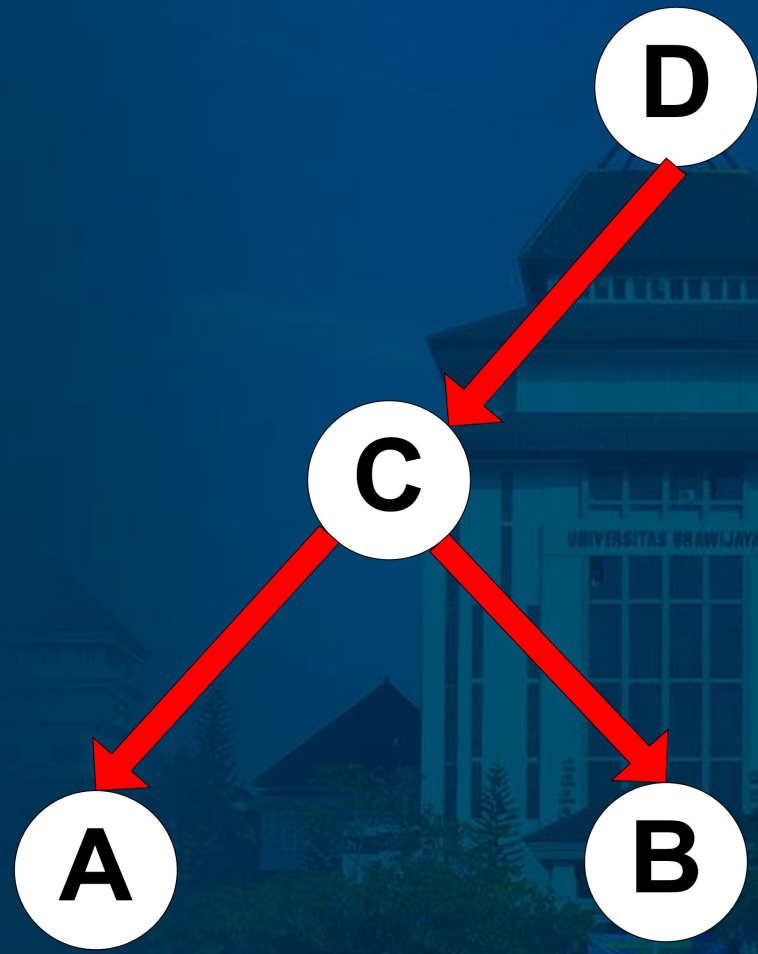
- Terdiri dari:
  - Directed acyclic graph
    - Node adalaha variable
    - Edge adalah hubungan antar variable
  - Table probabilitas (probability table) yang menghubungkan tiap node dengan parentnya



Conditional independence

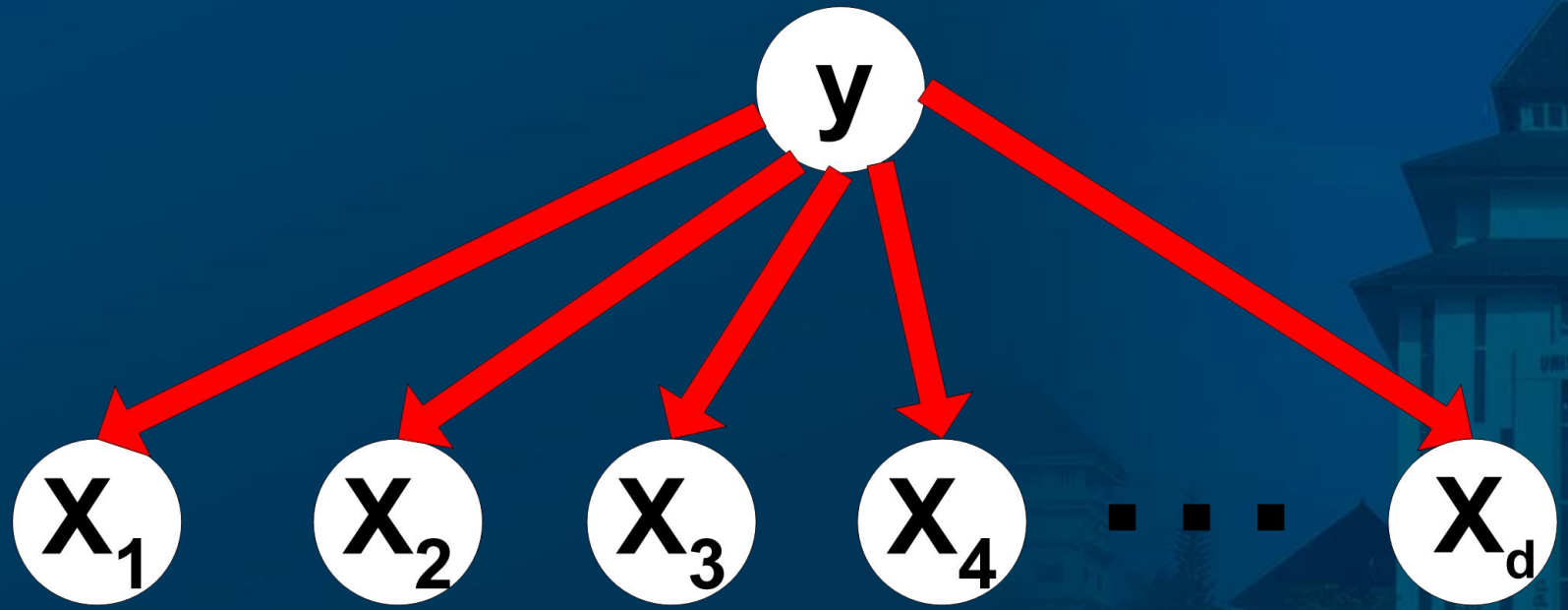
- D is parent of C
- A is child of C
- B is descendant of D
- D is ancestor of A

- A node in a Bayesian network is conditionally independent of all of its nondescendants, if its parents are known



Conditional independence

- Asumsi dalam Naïve Bayes



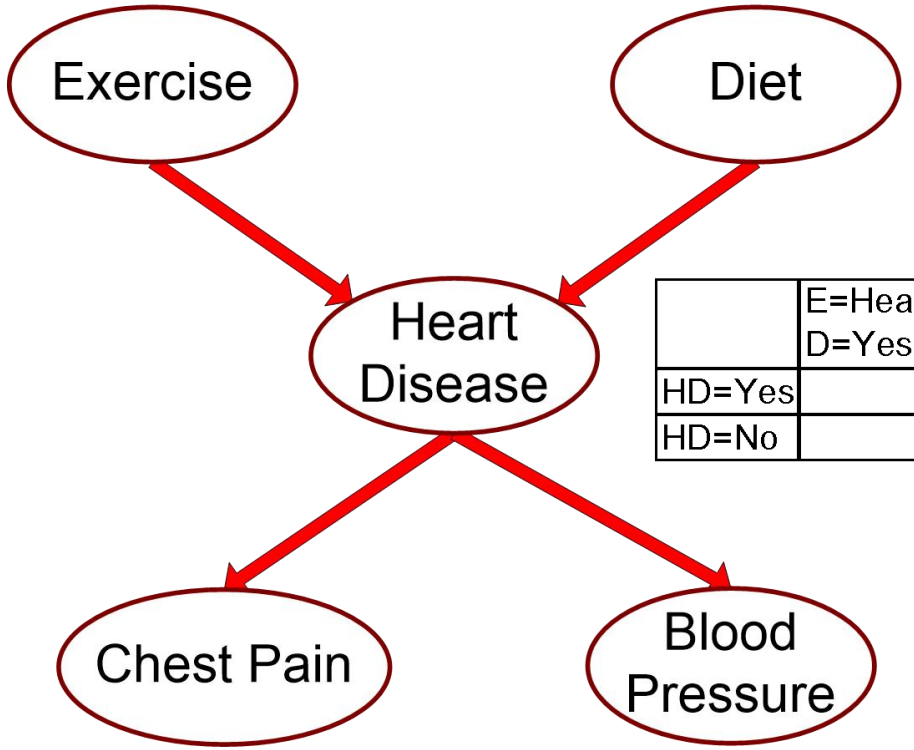
Probability Tables

- Jika sebuah node tidak memiliki parent, table probability-nya diisi dengan prior probability P(x)
- Jika variable x hanya memiliki sebuah parent (y), maka table probability-nya diisi dengan conditional probability P(x|y)
- Dan jika variable x memiliki banyak parents (y1, y2, …, yk), maka table probability-nya diisi dengan conditional probability P(x|y1, y2, …, yk)

Contoh Bayesian Belief Network

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	E=Healthy D=Yes	E=Healthy D=No	E=Unhealthy D=Yes	E=Unhealthy D=No
HD=Yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Terima Kasih

Building Up
Noble Future