

# Bayesian

Hendrik Santoso Sugiarto

IBDA2032 – *Artificial Intelligence*

# Capaian Pembelajaran

- Bayesian
- Estimasi Parametrik
- Model Probabilistik Generatif
- Model Probabilistik Diskriminatif
- Probabilistic Graphical Model

# Bayesian

# Bayesian Learning

- Paradigma machine learning yang menggunakan perspektif peluang munculnya sebuah data
- Dapat digunakan untuk regresi, klasifikasi, unsupervised
- Mudah diinterpretasi
- Probabilistic: setiap prediksi memiliki nilai peluang

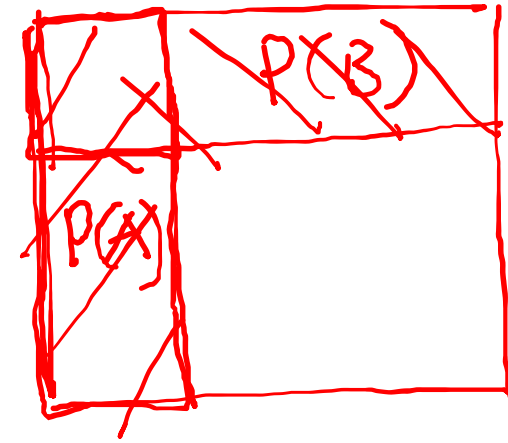
# Peluang bersyarat

- $P(A)$  → peluang A terjadi
- $P(A, B)$  → peluang A dan B terjadi
- $P(A|B)$  → peluang A terjadi jika B terjadi
- Conditional probability dan chain rule:

$$P(B|A) = \frac{P(A, B)}{P(A)} \rightarrow P(A, B) = P(A)P(B|A)$$

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$$

$$\begin{aligned} P(A, B, C, D) &= P(D|A, B, C) P(A, B, C) \\ &= P(D|A, B, C) P(C|A, B) P(A, B) \\ &= P(D|A, B, C) P(C|A, B) P(B|A) P(A) \end{aligned}$$



# Independen

- Jika A dan B tidak saling berhubungan maka

$$P(A, B) = P(A)P(B)$$

- Jika A dan B memiliki korelasi maka

$$P(A, B) \neq P(A)P(B)$$

$P(A)$

$P(B)$

$\rightarrow P(A, B) \neq P(A)P(B)$

# Uji Pemahaman

- Tidak terdapat hubungan antara hujan dan gempa bumi. Jika peluang hujan adalah 0.5 dan gempa bumi adalah 0.001:

- berapa peluang gempa bumi terjadi bersamaan dengan hujan?
- Saat ini sedang hujan, berapa peluang terjadi gempa bumi?
- Saat ini terjadi gempa bumi, berapa peluang hujan?

$$0.5 \times 0.001 = 0.0005$$

$$\rightarrow P(A, B)$$

$$\rightarrow P(B|A) = \frac{P(A, B)}{P(A)} = \frac{0.0005}{0.5} = 0.001$$

$$\rightarrow P(A|B) = \frac{P(A, B)}{P(B)} = \frac{0.0005}{0.001} = 0.5$$

# Teorema Bayes

likelihood  $\nwarrow$  Prior

- $Posterior \propto Likelihood \times Prior$

posterior  $\longleftarrow P(h|D) = \frac{P(D|h)P(h)}{P(D)}$

- $P(h)$  = prior probability hypothesis  $h$  (prior)
- $P(D)$  = prior probability data  $D$  (evidence)
- $P(h|D)$  = conditional probability of  $h$  given  $D$  (posterior)
- $P(D|h)$  = conditional probability of  $D$  given  $h$  (likelihood)



# Contoh

- Tes antigen menunjukkan hasil positif. Pada saat itu hanya 1% penduduk terkena covid. Jika akurasi hasil positif adalah 98% dan akurasi hasil negative 97%, berapa peluang terkena covid?
- $P(covid) = 0.01, P(\neg covid) = 0.99$
- $P(+|covid) = 0.98, P(-|covid) = 0.02$
- $P(-|\neg covid) = 0.97, P(+|\neg covid) = 0.03$
- $P(\underline{covid}|\underline{+}) = \frac{P(+|covid)P(covid)}{P(+)} = \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.03 \times 0.99} = \frac{0.0098}{0.0395} = \underline{0.2481}$

# Uji Pemahaman

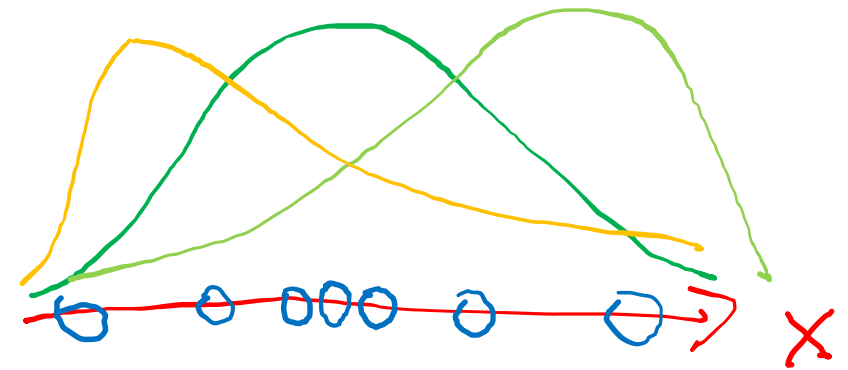
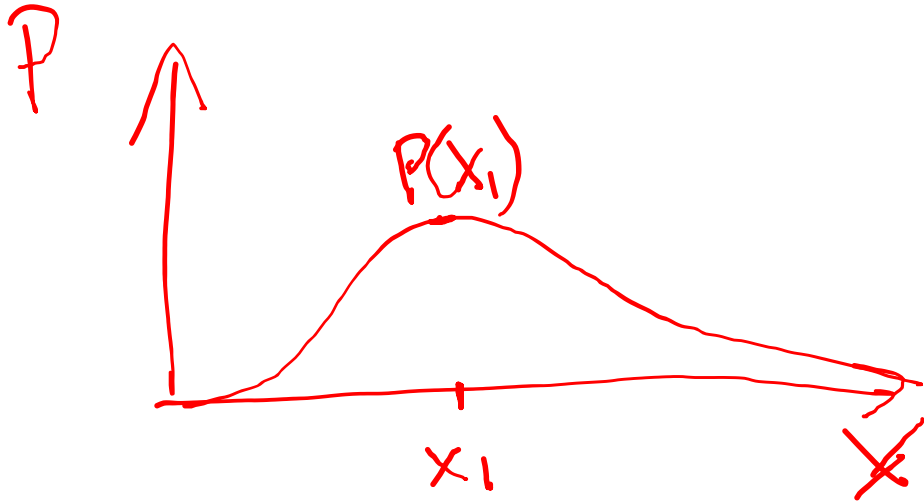
- Berapa peluang terkena covid jika hasil tes berikutnya adalah positif lagi?

$$\begin{aligned} P(\text{covid} | +) &= \frac{P(+ | \text{covid}) P(\text{covid})}{P(+)} \rightarrow 0.25 \\ &= \frac{0.98 \times 0.25}{0.98 \times 0.25 + 0.03 \times 0.75} \\ &= 0.92 \end{aligned}$$

# Estimasi Parametrik

# Density Estimation

- Density estimation bertujuan untuk estimasi fungsi peluang dibalik data observasi → menemukan sebuah distribusi  $p$  yang paling mendekati distribusi asli  $d$
- Data: data sample  $x$  diambil secara iid (independent identically distributed) dari sebuah distribusi  $d$



# Parametric Density Estimation

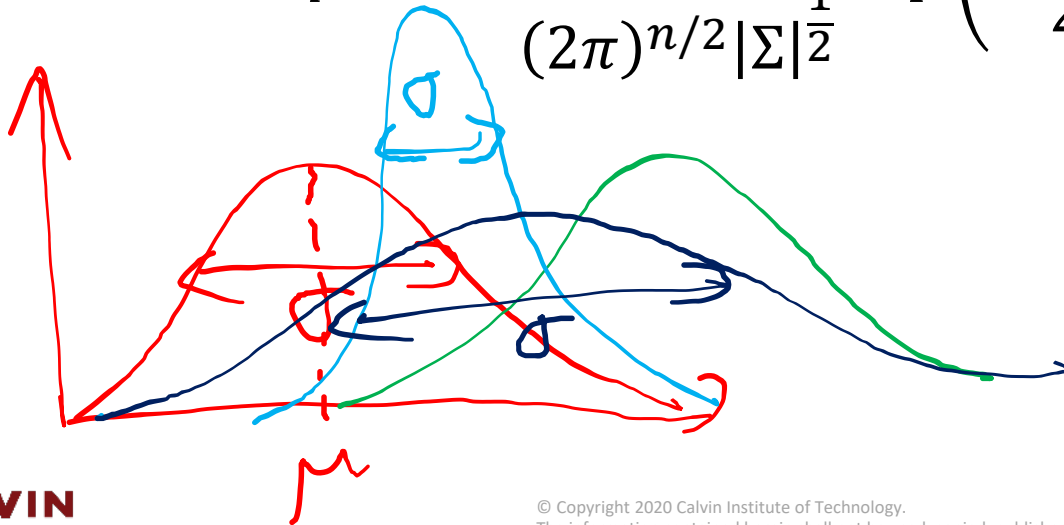
- Parametrik berarti sebuah density function memiliki bentuk tertentu, dengan parameters yang bisa diestimasi
- Gaussian density function

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

*data* (pointing to  $x$ )  
*rata-rata* (pointing to  $\mu$ )  
*standard deviation* (pointing to  $\sigma$ )

- Multivariate gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$



# Maximum-Likelihood Estimation

$$P(x^{(1)}) = 0.0001 = 10^{-4}$$

$$\log P(x^{(1)}) = -4$$

- Likelihood: peluang untuk menemukan data pada distribusi  $d$ , dengan asumsi independent:

*data pertama data kedua*

$$P(x^{(1)}, x^{(2)}, \dots, x^{(m)}) = \prod_{i=1}^m p(x^{(i)}) = p(x^{(1)}) \cdot p(x^{(2)}) \cdot p(x^{(3)}) \dots$$

- MLE principle: pilih sebuah distribusi yang dapat memaksimalkan likelihood

$$0.1 \times 0.1 \times 0.1$$

$$\log(a \times b \times c) = \log(0.001)$$

$$= -3$$

$$\log a + \log b + \log c$$

$$-1 -1 -1$$

$$p_* = \arg \max_p \prod_{i=1}^m p(x^{(i)})$$

$$p_* = \arg \max_p \sum_{i=1}^m \log p(x^{(i)})$$

# MLE vs MAP

- MLE: mencari hipotesis  $h$  yang memaksimalkan likelihood data  $D$   
$$h_{MLE} = \arg \max_h P(D|h)$$
- MAP: mencari hipotesis  $h$  yang memaksimalkan posterior data  
$$h_{MAP} = \arg \max_h P(h|D)$$
- Jika priornya adalah uniform distribution maka MLE sama dengan MAP

# Optimisasi dari Perspektif Bayesian

## Bayesian

- Probability = a degree of belief
- Aturan Bayes  $P(x|y) = P(x)P(y|x)/P(y)$
- Maximum Likelihood Estimation (MLE)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} (P(y|\theta))$$

- Maximum A Posteriori (MAP)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left( \frac{P(y|\theta)P(\theta)}{\int_{\Theta} P(y|\phi)P(\phi)d\phi} \right) = \underset{\theta}{\operatorname{argmax}} (P(y|\theta)P(\theta))$$
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} (\log P(y|\theta) + \log P(\theta))$$



# Kasus Gaussian

- Gaussian conditional likelihood

$$P(\overset{\text{label}}{\mathbf{y}}|\overset{\text{data}}{\mathbf{X}}, \boldsymbol{\Theta}) = \prod_i^N \exp\left(-\frac{1}{2}(y_i - \mathbf{x}_i\boldsymbol{\theta})^2\right)$$

- Gaussian Prior

$$P(\boldsymbol{\Theta}) = \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\theta})^2\right)$$

- Log posterior

$$\ln P(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta})P(\boldsymbol{\Theta}) = \sum_i^N -\frac{1}{2}(y_i - \mathbf{x}_i\boldsymbol{\theta})^2 - \frac{1}{2\sigma^2}(\boldsymbol{\theta})^2$$

- MAP

$$\frac{d}{d\theta_j} \ln P(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta})P(\boldsymbol{\Theta}) = \sum_i^N -\frac{1}{2} \underline{x_{ij}(y_i - \mathbf{x}_i\boldsymbol{\theta})} - \underline{\lambda(\theta_j)} = 0$$

- Log Posterior = Regresi Linear + Regularisasi

# Kasus Binomial

- Binomial conditional likelihood

$$P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_i^N [\sigma(x_i \boldsymbol{\theta})]^{y_i} [1 - \sigma(x_i \boldsymbol{\theta})]^{1-y_i}$$

Product

$$\prod_{i=1}^N x_i = x_1 \cdot x_2 \cdot x_3$$
$$\sum_{i=1}^N x_i = x_1 + x_2 + x_3$$

- Gaussian Prior

$$P(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{\theta})^2\right)$$

- Log posterior

$$\ln P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})P(\boldsymbol{\theta}) = \sum_i^N [-\ln(1 + \exp(x_i \boldsymbol{\theta})) + y_i x_i \boldsymbol{\theta}] - \frac{1}{2\sigma^2} (\boldsymbol{\theta})^2$$

- MAP

$$\frac{d}{d\theta_j} \log P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})P(\boldsymbol{\theta}) = \sum_i^N [y_i - \sigma(x_i \theta_j)] x_i - \lambda \theta_j$$

- Log Posterior = Regresi Logistik + Regularisasi

# Model Probabilistik Generatif

# Probabilistic Generative Models

- Terdapat data latih dari K kelas:

$$(x^{(i)}, y^{(i)})$$

*data* (pointing to  $x^{(i)}$ )  
*label* (pointing to  $y^{(i)}$ )

- Klasifikasi data  $x$  menuju salah satu dari K kelas

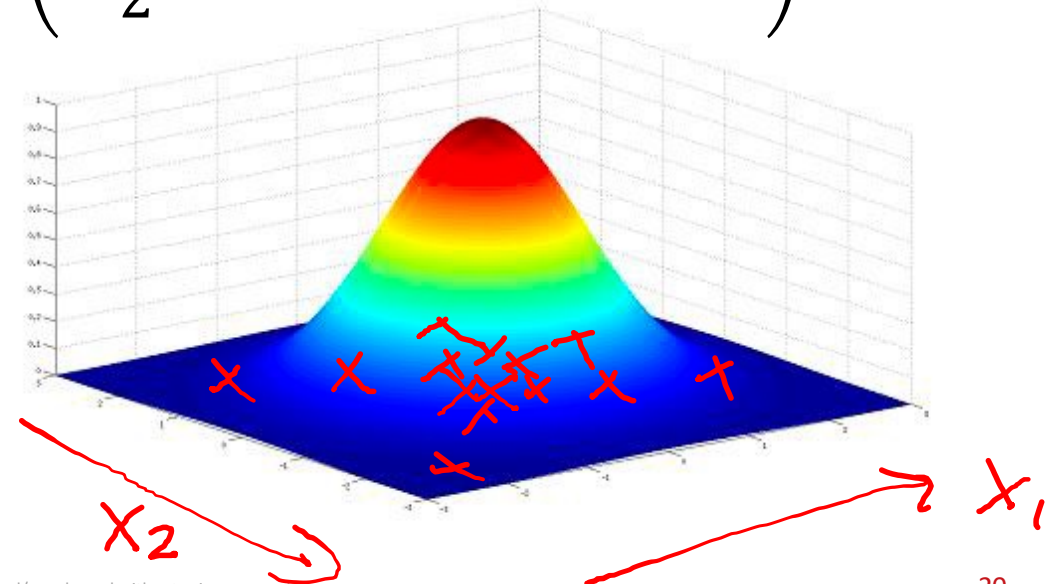
$$p(C_k|x) \propto p(x|C_k)p(C_k)$$

- Density function untuk kelas  $C_k$

$$p(x|C_k) = N(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right)$$

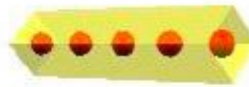
*parameter* (pointing to  $\mu_k$  and  $\Sigma_k$ )

*p*

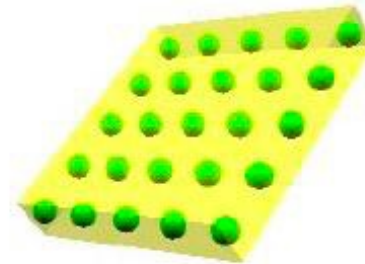


# Curse of Dimensionality

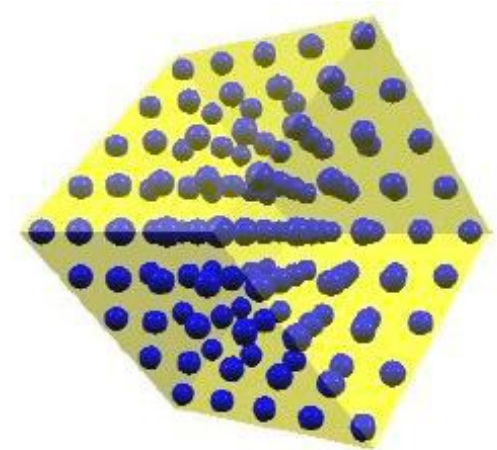
- Problem dari metode ini biasanya tidak cukupnya data untuk menebak parameter
- Jika 5 data cukup untuk menebak pola 1D maka dibutuhkan:
  - 1D: 5 data
  - 2D: 25 data
  - 3D: 125 data
  - 10D: 9765625 data



5 points



25 points



125 points



# Klasifikasi Naïve Bayes

- Algoritma efisien namun memiliki performa / generalisasi yang lebih buruk jika dibandingkan linear classifier lain seperti regresi logistic atau SVM
- Asumsi naïve bayes: setiap fitur independent terhadap fitur yang lain sehingga memerlukan data lebih sedikit dibanding algoritma lain
- Contoh: benda apakah dengan warna merah, bentuk bulat, diameter sekitar 10 cm?
- Prediksi: apel
- Mudah dilakukan karena fitur warna, bentuk, dan ukuran independent terhadap peluang benda tersebut adalah apel. Korelasi antar fitur dapat diabaikan sekalipun ada

# Naïve Bayes

- Sulit untuk menebak  $p(\mathbf{x}|C_k)$  pada data dimensi tinggi / fitur banyak
- Asumsi naïve bayes: semua fitur independent
- Aproksimasi naïve bayes:

$$p(\mathbf{x}|C_k) \approx \prod_{j=1}^d p(x_j|C_k)$$
$$p(\mathbf{x}|C_k) = \underline{N(\mathbf{x}|\mu, \Sigma)} \approx \prod_{j=1}^d p(x_j|C_k) = \prod_{j=1}^d N(x_j|\mu_j, \sigma_j^2)$$


# Naïve Bayes Classifier

- Untuk klasifikasi:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})} \propto P(\mathbf{x}|C_k)P(C_k)$$

- Naïve Bayes Classifier:

$$C_{NB} = \arg \max_{\underline{C_k}} P(C_k) \prod_{j=1}^d p(x_j|C_k)$$





# Model Probabilistik Diskriminatif

# Batas Keputusan pada Naïve Bayes

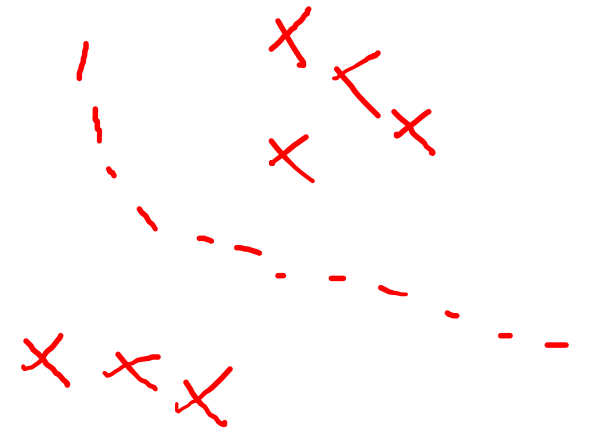
- Misalkan kita mengklasifikasikan 2 kelas
- Gaussian density function:  $p(\mathbf{x}|C_k) = N(\mathbf{x}|\mu_k, \Sigma_k)$
- Misalkan kovariannya sama:  $\Sigma_1 = \Sigma_2 = \Sigma$
- Rasio pengambilan keputusan:

$$\frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} = \frac{P(C_1)}{P(C_2)} \times \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)}$$
$$\ln \frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} \propto \ln \frac{P(C_1)}{P(C_2)} - \mathbf{x}^T \Sigma^{-1} (\mu_1 - \mu_2)$$

- Persamaan ini merupakan bentuk dari batas keputusan:

$$\ln \frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} = b + \mathbf{x}^T \boldsymbol{\theta}$$

$$\exp\left(\frac{(\mathbf{x} - \mu_k)^T \Sigma^{-1} (\mathbf{x} - \mu_k)}{2}\right)$$



# Regresi Logistik

- Linear discriminatory model: memodelkan langsung batas keputusan linear

$$\ln \frac{P(y = 1|\mathbf{x})}{P(y = -1|\mathbf{x})} = b + \mathbf{x}^T \boldsymbol{\theta}$$

$$\frac{P(y|\mathbf{x})}{1 - P(y|\mathbf{x})} = \exp(b + \mathbf{x}^T \boldsymbol{\theta})$$

$$P(y|\mathbf{x}) = (1 - P(y|\mathbf{x})) \exp(b + \mathbf{x}^T \boldsymbol{\theta})$$

$$P(y|\mathbf{x})(1 + \exp(b + \mathbf{x}^T \boldsymbol{\theta})) = 1$$

$$\hat{y} \Rightarrow P(y|\mathbf{x}) = \frac{1}{1 + \exp(b + \mathbf{x}^T \boldsymbol{\theta})} = \sigma(b + \mathbf{x}^T \boldsymbol{\theta})$$

- Jadi dalam paradigma ini, peluang dimodelkan dengan sigmoid

# Discriminative vs Generative

- Regresi Logistik: memodelkan  $P(y|x)$
- Kelebihan :
  - Performa lebih baik
  - Robust terhadap noise
- Kekurangan:
  - Susah konvergen
  - Expensive computation
- Naïve Bayes: memodelkan  $P(x|y)$
- Kelebihan:
  - Mudah konvergen
  - Cheap computation ✓
- Kekurangan:
  - Performa lebih buruk
  - Sensitif terhadap noise



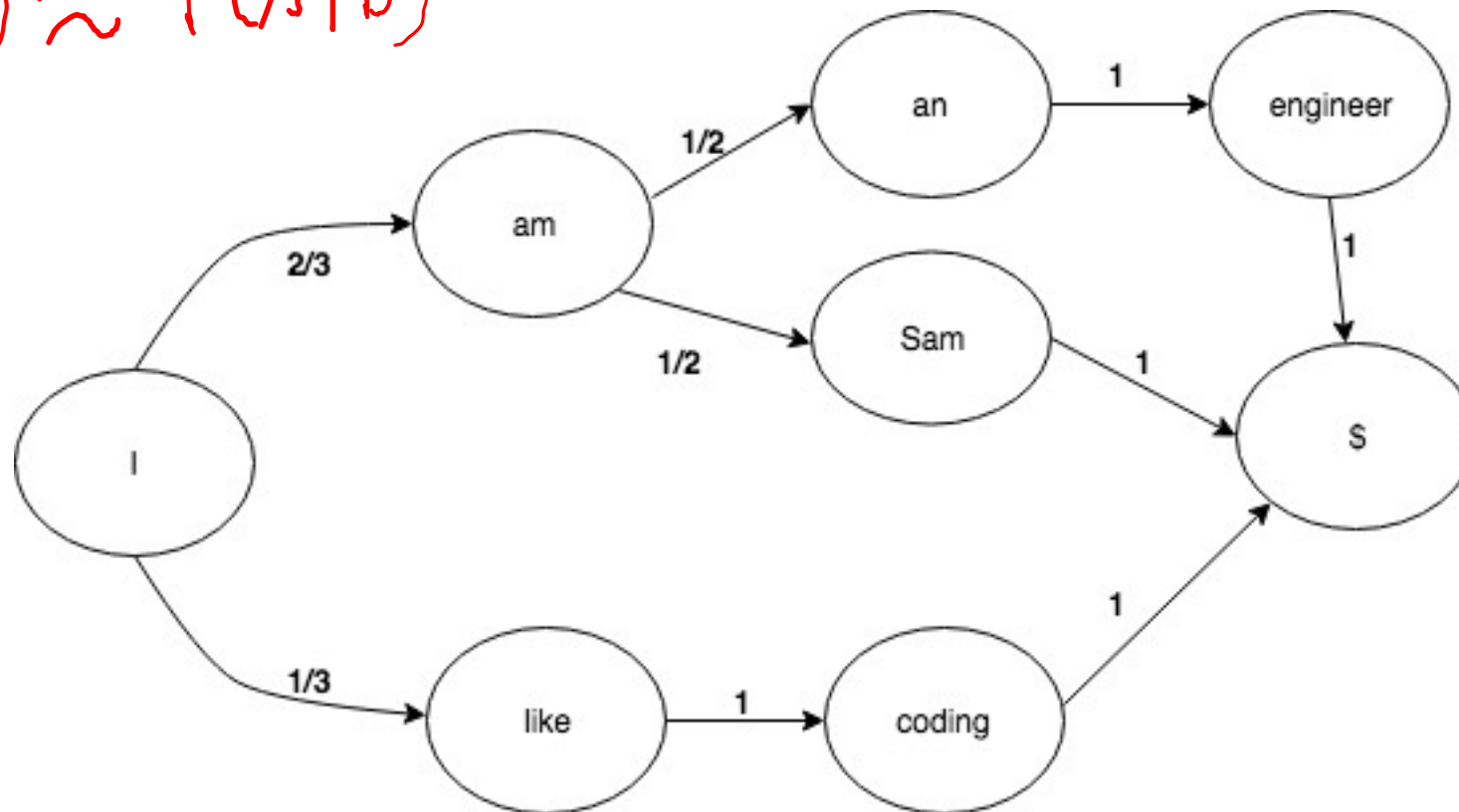
# Probabilistic Graphical Model

# Markov Process

- Peluang setiap kata hanya bergantung pada beberapa kata sebelumnya

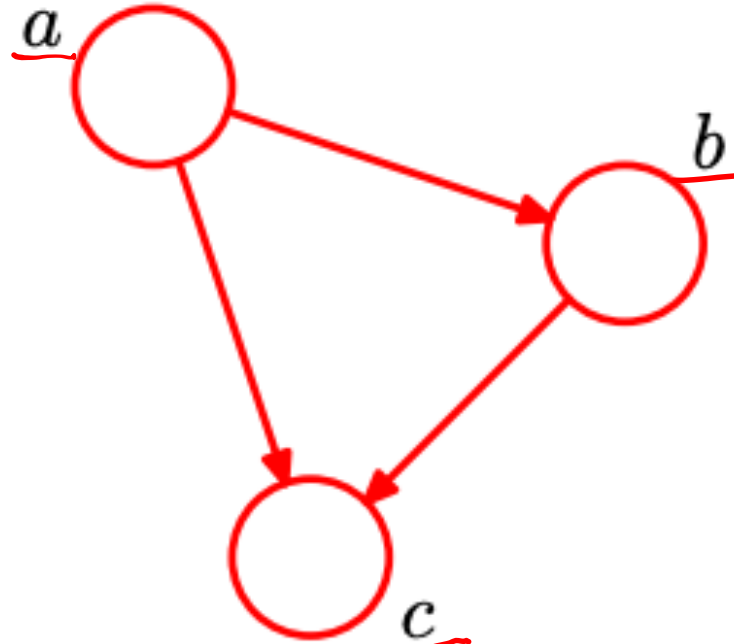
$$P(w_N | w_1^{N-1}) \approx P(w_N | w_{N-1})$$

$$P(A|BCD) \approx P(A|B)$$



# Bayesian Networks

- Berlaku untuk relasi Bayesian apapun

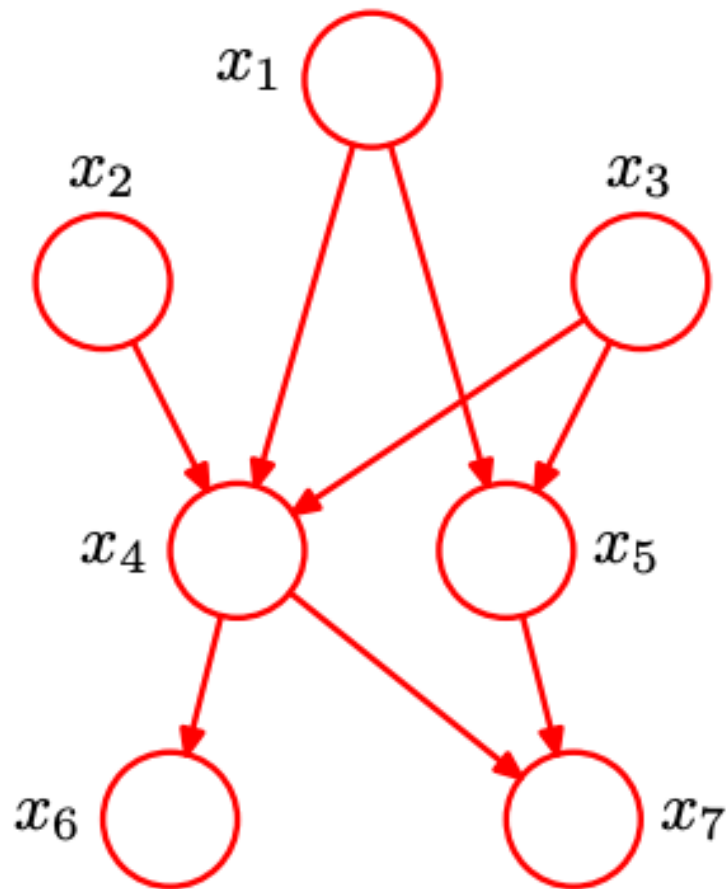


$$p(\underline{a}, \underline{b}, \underline{c}) = p(\underline{c} | \underline{a}, \underline{b}) p(\underline{b} | \underline{a}) p(\underline{a})$$

*Handwritten red text above the equation:*  $p(c|a,b)p(a,b)$

# Uji Pemahaman

- Apakah relasi Bayesian untuk graphical model berikut?



$$P(x_1) P(x_2) P(x_3) P(x_4 | x_1, x_2, x_3) P(x_5 | x_1, x_3) \\ P(x_6 | x_4, x_1, x_2, x_3) P(x_7 | x_1, x_2, x_3, x_4, x_5, x_3)$$

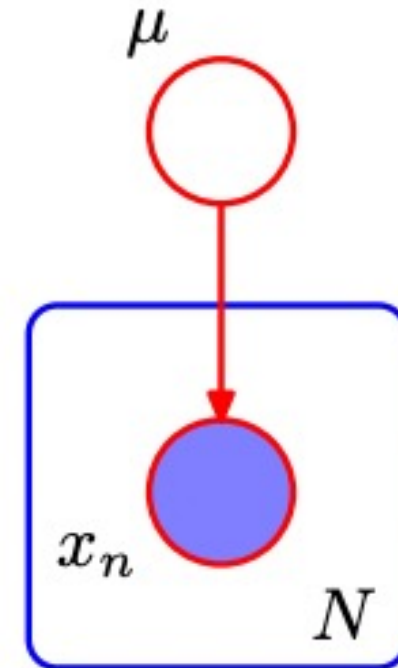
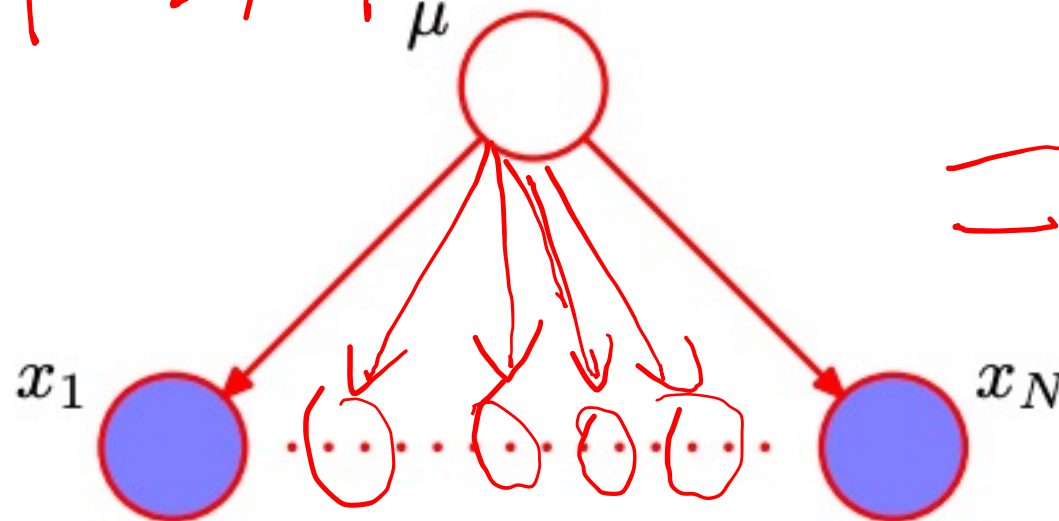


# Multiple output (Plate Diagram)

- Distribusi parametrik dapat menghasilkan banyak output

$$P(D, \mu) = p(w) \prod_{n=1}^N p(x_n | \mu)$$

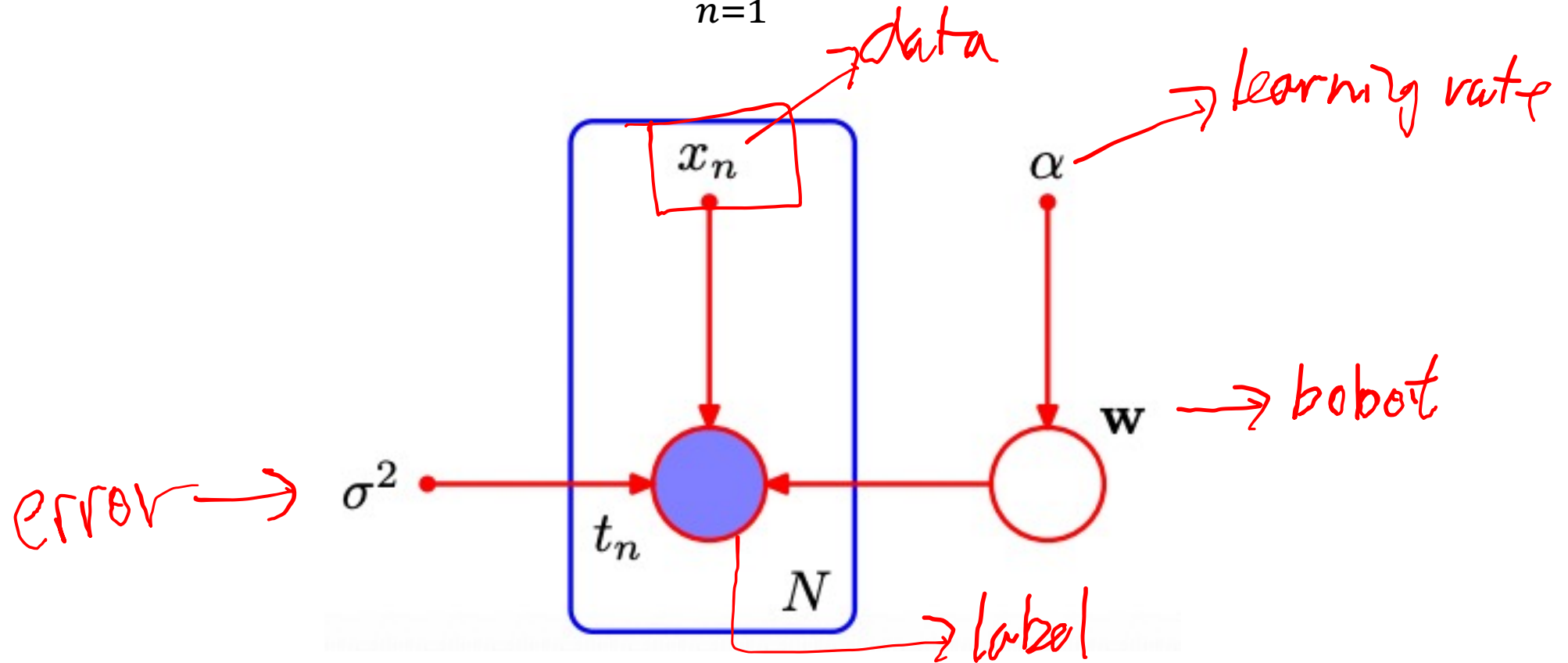
$p(x_1 | \mu) p(x_2 | \mu) p(x_3 | \mu) \dots$



# Polynomial regression model

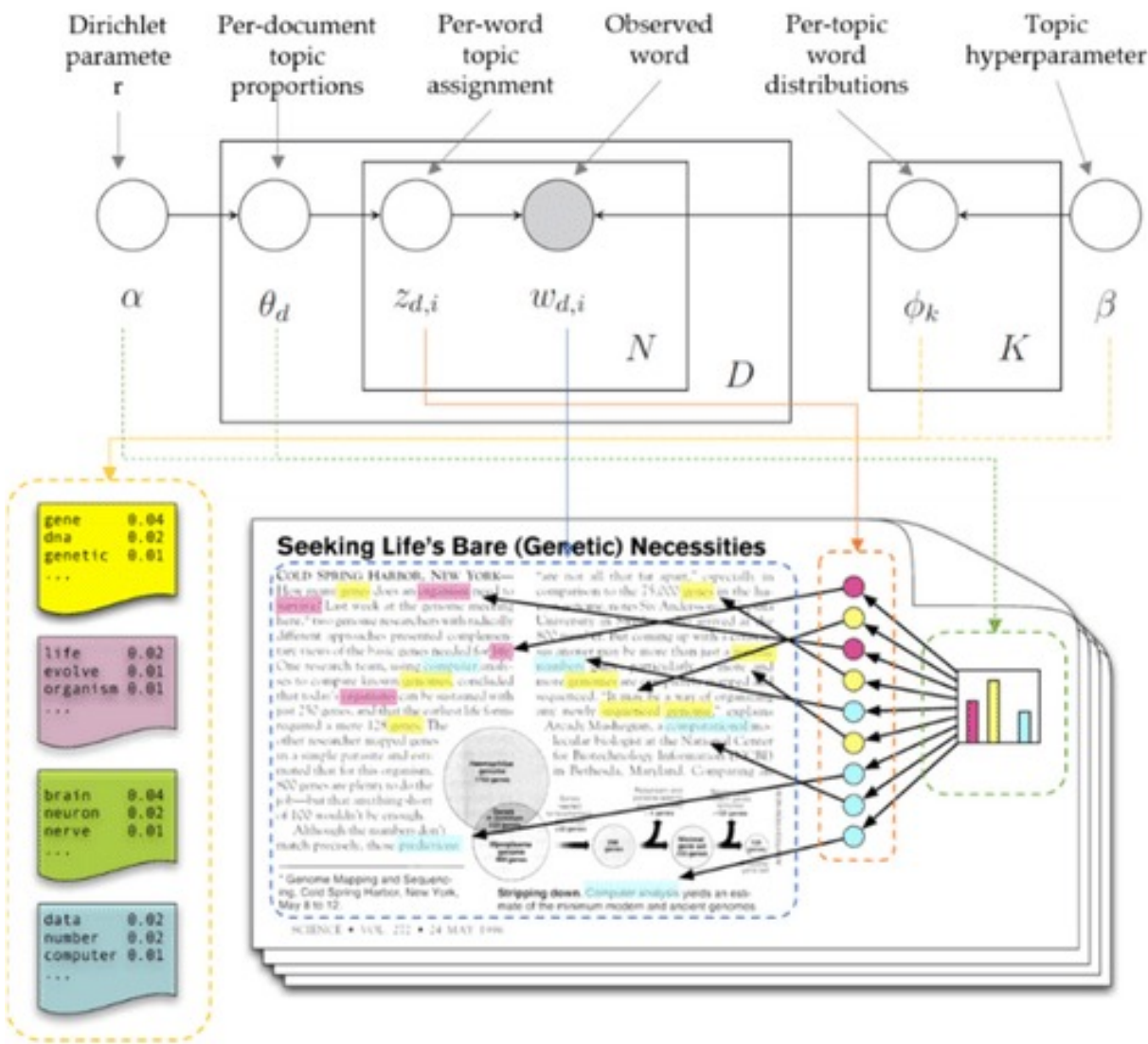
- Regresi: input data  $\mathbf{x} = (x_1, \dots, x_N)^T \rightarrow$  observed output  $\mathbf{t} = (t_1, \dots, t_N)^T$

$$P(\mathbf{t}, \mathbf{w} | \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w} | \alpha) \prod_{n=1}^N p(t_n | \mathbf{w}, x_n, \sigma^2)$$



# Latent Dirichlet Allocation

(a) LDA document generation process



(b) An illustrative example of LDA document generation process

(c) Two outputs of LDA

(c-1) Per-document topic proportions ( $\theta_d$ )

	Topic 1	Topic 2	Topic 3	...	Topic K
Doc 1	0.20	0.50	0.10	...	0.10
Doc 2	0.50	0.02	0.01	...	0.40
Doc 3	0.05	0.12	0.48	...	0.15
...	...	...	...	...	...
Doc N	0.14	0.25	0.33	...	0.14

(c-2) Per-topic word distributions ( $\phi_k$ )

	Topic 1	Topic 2	Topic 3	...	Topic K
word 1	0.01	0.05	0.05	...	0.10
word 2	0.02	0.02	0.01	...	0.03
word 3	0.05	0.12	0.08	...	0.02
...	...	...	...	...	...
word N	0.04	0.01	0.03	...	0.07



# Tuhan Memberkati



*God's People for God's Glory*

**CALVIN**  
INSTITUTE OF TECHNOLOGY