Bayesian Decision

Jin Young Choi

Seoul National University

Outline

- Bayes Formula
  - Priori probability
  - Likelihood
  - Posterior Probability
  - Bayes Decision
- Risk Formulation
  - Conditional Risk
  - Likelihood Ratio Test
  - Zero-one Loss Function (Bayes Decision)
- Maximum Likelihood Estimation (MLE)
- Error Probability

Bayesian Decision

- Question:
  - There live two kinds of fishes in a lake: tuna or salmon.
  - If you catch a fish by fishing, is the fish likely to be tuna or salmon?

Bayesian Decision

- We have experienced that salmon has been caught in 70% and tuna in 30%.
- What is the next fish likely to be?

Bayesian Decision

- If other types of fish are irrelevant:

$$p(\omega = \omega_1) + p(\omega = \omega_2) = 1,$$

  ω is random variable, ω_1 and ω_2 denote salmon and tuna.

- Probabilities reflect our prior knowledge obtained from past experience.
- Simple Decision Rule:
  - Make a decision without seeing the fish.
  - Decide ω_1 if $p(\omega = \omega_1) > p(\omega = \omega_2)$

    ω_2 otherwise.

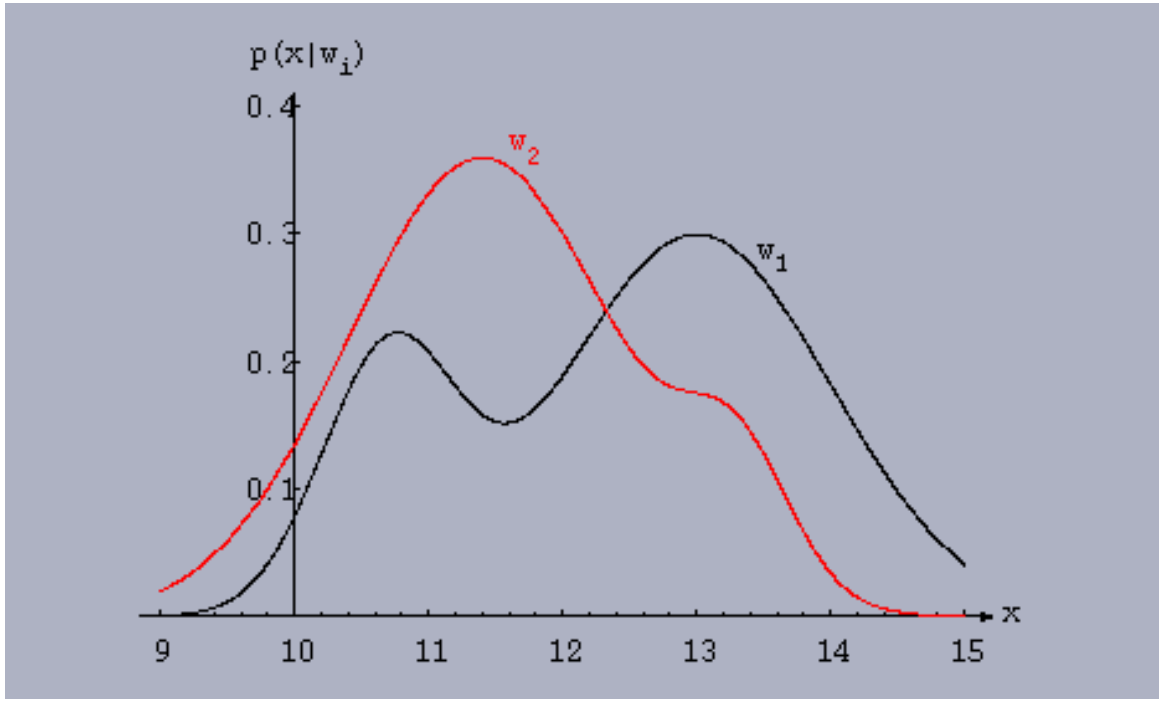
Bayesian Decision

- In general, we will have some features and more information.
- Feature: lightness measurement = x
  - Different fish yields different lightness readings (x is a random variable)

Bayesian Decision

- Define
  - $p(x|\omega_i)$= Class Conditional Probability Density
  - The difference between $p(x|\omega_1)$ and $p(x|\omega_2)$ describes the difference in lightness between tuna and salmon.

Bayesian Decision



- Hypothetical class-conditional probability
- Density functions are normalized (area under each curve is 1.0)

Bayesian Decision

- Suppose that we know
  - The prior probabilities $p(\omega_1)$ and $p(\omega_2)$
  - The conditional densities $p(x|\omega_1)$ and $p(x|\omega_2)$
  - Measure lightness of a fish $= x$
- What is the category of the fish with lightness of x ?
- The probability that the fish has category of ω_i is $p(\omega_i|x)$.

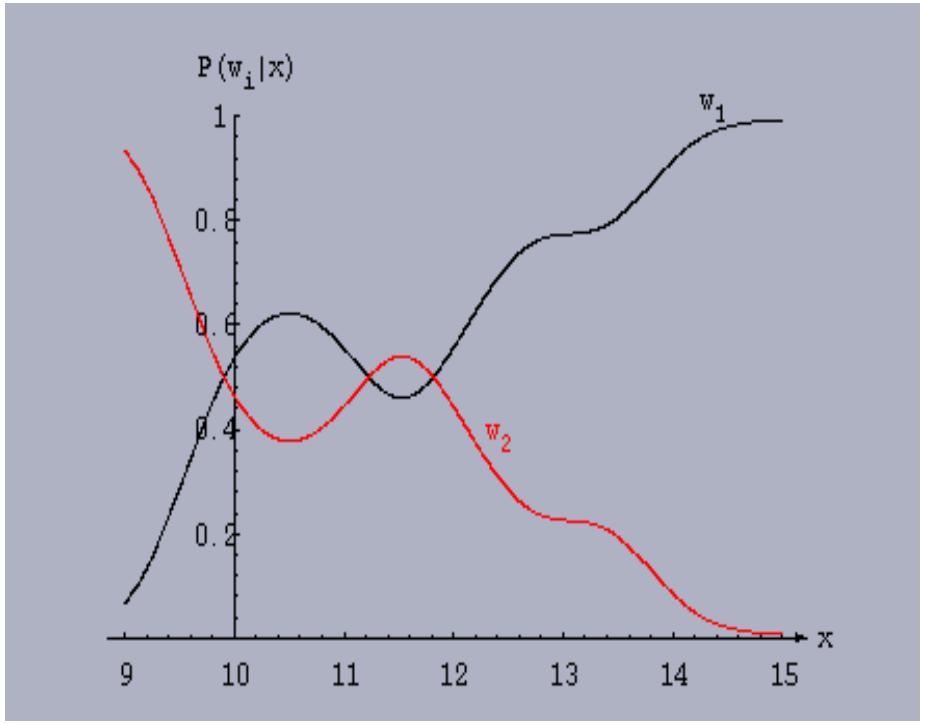
Bayes formula

- $p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{p(x)}$,

  where $p(x) = \sum_j p(x|\omega_j)p(\omega_j) = \sum_j p(x, \omega_j)$.

- $Posterior = \frac{Likelihood * Prior}{Evidence}$
- $p(x|\omega_i)$ is called the ***likelihood*** of ω_i with respect to x.
- The ω_i category for which $p(x|\omega_i)$ is large is more "likely" to be the true category
- $p(x)$ is the ***evidence***
  - How frequently is a pattern with feature value x observed.
  - Scale factor that the posterior probabilities sum to 1.

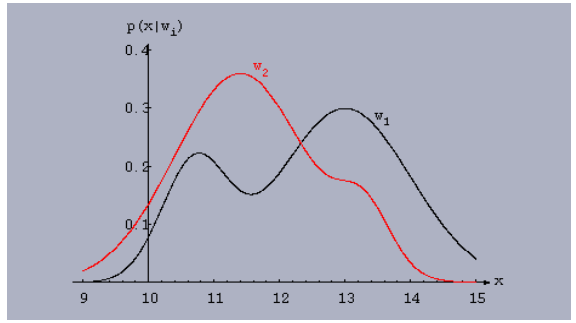
Bayes formula

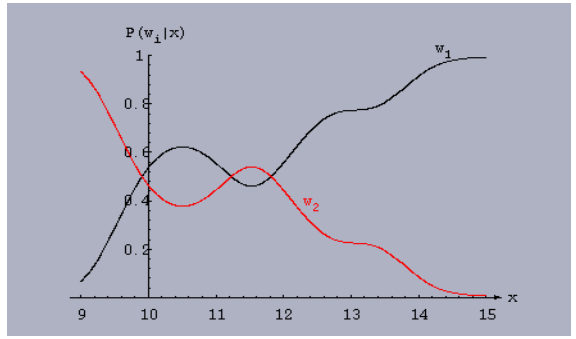


- Posterior probabilities for the particular priors $p(\omega_1) = 2/3$ and $p(\omega_2) = 1/3$. At every x the posteriors sum to 1.

Bayes Decision Rule (Minimal probability error)

- Likelihood Decision:
  - $\omega_1 : if\ p(x|\omega_1) > p(x|\omega_2)$
  - $\omega_2 :$ *otherwise*
- Posteriori Decision:
  - $\omega_1 : if\ p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$
  - $\omega_2 :$ *otherwise*
- Decision Error Probability
  - $p(error|x) = \min(p(\omega_1|x), p(\omega_2|x))$





where the decision error is given by

$$p(error|x) = \begin{cases} p(\omega_2|x) & if\ we\ decide\ \omega_1\ for\ \omega_2 \\ p(\omega_1|x) & if\ we\ decide\ \omega_2\ for\ \omega_1 \end{cases}$$

Exercise

- 한 해안가에서 연어가 잡힐 확률은 0.6 이고 농어가 잡힐 확률은 0.4 이다. 잡힌 연어 중 40 cm 이하의 크기일 확률은 20%이고, 농어 중 40cm 이하일 확률은 3%이다. 잡은 고기가 40cm 이하 일 때 연어라고 판단할 것인지? 아니면 농어로 판단할 것인지 결정하시오.

Exercise

- 한 해안가에서 연어가 잡힐 확률은 0.6 이고 농어가 잡힐 확률은 0.4 이다. 잡힌 연어 중 40 cm 이하의 크기일 확률은 20%이고, 농어 중 40cm 이하일 확률은 3%이다. 잡은 고기가 40cm 이하 일 때 연어라고 판단할 것인지? 아니면 농어로 판단할 것인지 결정하시오.

- Sol.
  - ✓ (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.

Exercise

- 한 해안가에서 연어가 잡힐 확률은 0.6 이고 농어가 잡힐 확률은 0.4 이다. 잡힌 연어 중 40 cm 이하의 크기일 확률은 20%이고, 농어 중 40cm 이하일 확률은 3%이다. 잡은 고기가 40cm 이하 일 때 연어라고 판단할 것인지? 아니면 농어로 판단할 것인지 결정하시오.

- Sol.
  - ✓ (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 연어: $X = 0,$ 농어: $X = 1,$ 크기: $Y.$
  - ✓ $P(X = 0) = 0.6, P(X = 1) = 0.4, P(Y \leq 40cm | X = 0) = 0.2, P(Y \leq 40cm | X = 1) = 0.03$

Exercise

- 한 해안가에서 연어가 잡힐 확률은 0.6 이고 농어가 잡힐 확률은 0.4 이다. 잡힌 연어 중 40 cm 이하의 크기일 확률은 20%이고, 농어 중 40cm 이하일 확률은 3%이다. 잡은 고기가 40cm 이하 일 때 연어라고 판단할 것인지? 아니면 농어로 판단할 것인지 결정하시오.

- Sol.
  - ✓ (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 연어: $X = 0$, 농어: $X = 1$, 크기: Y.
  - ✓ $P(X = 0) = 0.6, P(X = 1) = 0.4, P(Y \leq 40cm | X = 0) = 0.2, P(Y \leq 40cm | X = 1) = 0.03$
  - ✓ 질문: posteriori: $P(X = 0 \mid Y \leq 40cm) =?$, $P(X = 1 \mid Y \leq 40cm) =?$

Exercise

- 한 해안가에서 연어가 잡힐 확률은 0.6 이고 농어가 잡힐 확률은 0.4 이다. 잡힌 연어 중 40 cm 이하의 크기일 확률은 20%이고, 농어 중 40cm 이하일 확률은 3%이다. 잡은 고기가 40cm 이하 일 때 연어라고 판단할 것인지? 아니면 농어로 판단할 것인지 결정하시오.

- Sol.
  - ✓ (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 연어: $X = 0$, 농어: $X = 1$, 크기: Y.
  - ✓ $P(X = 0) = 0.6, P(X = 1) = 0.4, P(Y \leq 40cm | X = 0) = 0.2, P(Y \leq 40cm | X = 1) = 0.03$
  - ✓ 질문: posteriori: $P(X = 0 \mid Y \leq 40cm) =?, P(X = 1 \mid Y \leq 40cm) =?$
  - ✓ $P(X = 0 \mid Y \leq 40cm) = \frac{P(Y \leq 40cm | X=0)P(X=0)}{P(Y \leq 40cm)} = \frac{0.2 \times 0.6}{0.2 \times 0.6 + 0.4 \times 0.03} = 90.9\%$

Exercise

- 한 해안가에서 연어가 잡힐 확률은 0.6 이고 농어가 잡힐 확률은 0.4 이다. 잡힌 연어 중 40 cm 이하의 크기일 확률은 20%이고, 농어 중 40cm 이하일 확률은 3%이다. 잡은 고기가 40cm 이하 일 때 연어라고 판단할 것인지? 아니면 농어로 판단할 것인지 결정하시오.

- Sol.
  - ✓ (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 연어: $X = 0$, 농어: $X = 1$, 크기: Y.
  - ✓ $P(X = 0) = 0.6, P(X = 1) = 0.4, P(Y \leq 40cm | X = 0) = 0.2, P(Y \leq 40cm | X = 1) = 0.03$
  - ✓ 질문: posteriori: $P(X = 0 \mid Y \leq 40cm) = ?, P(X = 1 \mid Y \leq 40cm) = ?$
  - ✓ $P(X = 0 \mid Y \leq 40cm) = \frac{P(Y \leq 40cm | X = 0)P(X = 0)}{P(Y \leq 40cm)} = \frac{0.2 \times 0.6}{0.2 \times 0.6 + 0.4 \times 0.03} = 90.9\%$
  - ✓ $P(X = 1 \mid Y \leq 40cm) = \frac{P(Y \leq 40cm | X = 1)P(X = 1)}{P(Y \leq 40cm)} = \frac{0.4 \times 0.03}{0.2 \times 0.6 + 0.4 \times 0.03} = 9.09\%$
  - ✓ Bayes decision 에 의해 연어라고 판단한다.

General Formulation

- Let $\{\omega_1, \ldots, \omega_c\}$ be the finite set of c categories.
- Let $\{\alpha_1, \ldots, \alpha_a\}$ be the finite set of a possible actions.

  Ex. Action α_i = deciding that the true state is ω_i or others.
- The risk function $\lambda(\alpha_i|\omega_j)$ = risk incurred for taking action when the state of nature is ω_j.
- $x = d$ –dimensional feature vector (random variable)
- $p(x|\omega_i)$ = likelihood probability density function for x for given ω_i
- $p(\omega_i)$ = prior probability that nature is in state ω_i.

Conditional Risk

- After the observation, the expected risk (conditional risk) is given by

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) p(\omega_j|x)$$

- The decision action $\alpha(x)$ for given x is given

$$\alpha(x) = arg \min_{\alpha_i} R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) p(\omega_j|x)$$

Two-Category Classification

- Action α_1 = deciding that the true state is ω_1
- Action α_2 = deciding that the true state is ω_2
- Let $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ be the risk incurred for deciding ω_i when true state is ω_j.
- *The conditional risks:*

$$R(\alpha_1|x) = \lambda_{11}p(\omega_1|x) + \lambda_{12}p(\omega_2|x)$$
$$R(\alpha_2|x) = \lambda_{21}p(\omega_1|x) + \lambda_{22}p(\omega_2|x)$$

- Decide ω_1 if $R(\alpha_1|x) < R(\alpha_2|x)$

  or if $(\lambda_{21} - \lambda_{11})p(\omega_1|x) > (\lambda_{12} - \lambda_{22})p(\omega_2|x)$

  or if $(\lambda_{21} - \lambda_{11})p(x|\omega_1)p(\omega_1) > (\lambda_{12} - \lambda_{22})p(x|\omega_2)p(\omega_2)$

  and ω_2, *otherwise*

Two-Category Likelihood Ratio Test

- Under reasonable assumption that $\lambda_{12} > \lambda_{22}$ and $\lambda_{21} > \lambda_{11}$, (why?)

  decide ω_1 if $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})p(\omega_2)}{(\lambda_{21}-\lambda_{11})p(\omega_1)} = T$

  and ω_2 , *otherwise.*

- The ratio $\frac{p(x|\omega_1)}{p(x|\omega_2)}$ is called the *likelihood ratio*.

- We can decide ω_1 if the likelihood ratio exceeds a threshold T value that is independent of the observation x.

Minimum-Error-Rate Classification

- To give an equal cost to all errors, we define zero-one risk function as

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}, \quad \text{for } i, j = 1, \ldots, C$$

- The conditional risk representing error rate is

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) p(\omega_j|x)$$
$$= \sum_{j \neq i}^{c} p(\omega_j|x) = 1 - p(\omega_i|x)$$

- To **minimize** $R(\alpha_i|x)$, we **maximizes** $p(\omega_i|x)$

  Decide ω_i if $p(\omega_i|x) > p(\omega_j|x)$, for all $j \neq i$

  (same as Bayes' decision rule)

Maximum Likelihood Estimation (MLE)

- The samples are i.i.d.

$$j^{th}\ class\ set\ D_j = \{x_l | (x_l, \bar{\omega}_l) \in S_j\},\ S_j \subseteq S = \{(x_l, \bar{\omega}_l) | l = 1, \dots, N\}$$

- Maximum likelihood estimation: find $\hat{\theta}(D)$ to maximize $p(x|D)$

$$p(x|\omega_j) \approx p(x|D_j) \approx p(x|\hat{\theta}(D_j)), \qquad \hat{\theta}(D_j) = arg\max_{\theta} p(D_j|\theta)$$

- The i.i.d. assumption implies that

$$p(D_j|\theta_j) = \prod_{x \in D_j} p(x|\theta_j)$$

- Let D be a generic sample set of size $n = |D|$
- Log-likelihood function:

$$l(\theta; D) \equiv \ln p(D|\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta)$$

$$\nabla_\theta l(\theta; D) = 0$$

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암진단은 피검사를 하여 암표지자 값을 보고 판단을 한다. 이 암표지자를 확률 변수 x 로 했을 때, 정상인과 암환자 모두 다음의 매개변수 θ 로 표현되는 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$의 형태를 가지고 있다. 샘플은 i.i.d. 특성을 만족한다. 정상인의 암표지자 값을 평균하면 0.01이 되고 암환자의 암표지자를 평균하면 0.1이 된다. 정상인과 암환자의 분포를 가장 잘 나타내는 $\hat{\theta}_1, \hat{\theta}_2$ 를 MLE 방법으로 추정하시오.

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암진단은 피검사를 하여 암표지자 값을 보고 판단을 한다. 이 암표지자를 확률 변수 x 로 했을 때, 정상인과 암환자 모두 다음의 매개변수 θ 로 표현되는 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$의 형태를 가지고 각 샘플은 i.i.d. 라고 가정한다. 정상인의 암표지자 값을 평균하면 0.01이 되고 암환자의 암표지자를 평균하면 0.1이 된다. 정상인과 암환자의 분포를 가장 잘 나타내는 $\hat{\theta}_1, \hat{\theta}_2$ 를 MLE 방법으로 추정하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암진단은 피검사를 하여 암표지자 값을 보고 판단을 한다. 이 암표지자를 확률 변수 x 로 했을 때, 정상인과 암환자 모두 다음의 매개변수 θ 로 표현되는 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$의 형태를 가지고 각 샘플은 i.i.d. 라고 가정한다. 정상인의 암표지자 값을 평균하면 0.01이 되고 암환자의 암표지자를 평균하면 0.1이 된다. 정상인과 암환자의 분포를 가장 잘 나타내는 $\hat{\theta}_1, \hat{\theta}_2$ 를 MLE 방법으로 추정하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 정상인의 샘플 $n = 99000$ 개, 즉 $D_i = \{x_i | i = 1, \dots, n\}$ 를 가지고 MLE를 수행하자. Likelihood function 은 아래와 같이 정의한다.

$$p(D_i|\theta) = \prod_{i=1}^{n} \theta x_i e^{-\theta x_i}$$

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암진단은 피검사를 하여 암표지자 값을 보고 판단을 한다. 이 암표지자를 확률 변수 x 로 했을 때, 정상인과 암환자 모두 다음의 매개변수 θ 로 표현되는 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$의 형태를 가지고 각 샘플은 i.i.d. 라고 가정한다. 정상인의 암표지자 값을 평균하면 0.01이 되고 암환자의 암표지자를 평균하면 0.1이 된다. 정상인과 암환자의 분포를 가장 잘 나타내는 $\hat{\theta}_1, \hat{\theta}_2$ 를 MLE 방법으로 추정하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 정상인의 샘플 $n = 99000$ 개, 즉 $D_i = \{x_i | i = 1, \dots, n\}$ 를 가지고 MLE를 수행하자. Likelihood function 은 아래와 같이 정의한다.

$$p(D_i|\theta) = \prod_{i=1}^{n} \theta x_i e^{-\theta x_i}$$

  - ✓ 양변에 log를 위하여 log-likelihood 를 구하면 다음과 같다.

$$l(\theta) = n \log \theta + \sum_{i=1}^{n} \log x_i - \theta \sum_{i=1}^{n} x_i$$

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암진단은 피검사를 하여 암표지자 값을 보고 판단을 한다. 이 암표지자를 확률 변수 x 로 했을 때, 정상인과 암환자 모두 다음의 매개변수 θ 로 표현되는 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$의 형태를 가지고 각 샘플은 i.i.d. 라고 가정한다. 정상인의 암표지자 값을 평균하면 0.01이 되고 암환자의 암표지자를 평균하면 0.1이 된다. 정상인과 암환자의 분포를 가장 잘 나타내는 $\hat{\theta}_1, \hat{\theta}_2$ 를 MLE 방법으로 추정하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 정상인의 샘플 $n = 99000$ 개, 즉 $D_i = \{x_i | i = 1, \dots, n\}$ 를 가지고 MLE를 수행하자. Likelihood function 은 아래와 같이 정의한다.

$$p(D_i|\theta) = \prod_{i=1}^{n} \theta x_i e^{-\theta x_i}$$

  - ✓ 양변에 log를 위하여 log-likelihood 를 구하면 다음과 같다.

$$l(\theta) = n \log \theta + -\theta \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \log x_i$$

  - ✓ 양변에 θ 에 대해 미분하여 그 값이 0이 되도록 θ 를 구하면

$$\frac{d}{d\theta} l(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0, \quad \rightarrow \quad \hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암진단은 피검사를 하여 암표지자 값을 보고 판단을 한다. 이 암표지자를 확률 변수 x 로 했을 때, 정상인과 암환자 모두 다음의 매개변수 θ 로 표현되는 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$의 형태를 가지고 각 샘플은 i.i.d. 라고 가정한다. 정상인의 암표지자 값을 평균하면 0.01이 되고 암환자의 암표지자를 평균하면 0.1이 된다. 정상인과 암환자의 분포를 가장 잘 나타내는 $\hat{\theta}_1, \hat{\theta}_2$ 를 MLE 방법으로 추정하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 정상인의 샘플 $n = 99000$ 개, 즉 $D_i = \{x_i | i = 1, \dots, n\}$ 를 가지고 MLE를 수행하자. Likelihood function 은 아래와 같이 정의한다.

    $$p(D_i|\theta) = \prod_{i=1}^{n} \theta x_i e^{-\theta x_i}$$

  - ✓ 양변에 log를 위하여 log-likelihood 를 구하면 다음과 같다.

    $$l(\theta) = n \log \theta + -\theta \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \log x_i$$

  - ✓ 양변에 θ 에 대해 미분하여 그 값이 0이 되도록 θ 를 구하면

    $$\frac{d}{d\theta} l(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0, \quad \rightarrow \quad \hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i}$$

  - ✓ 여기서 정상인의 경우 암표지자 평균이 0.01 이므로 정상인 분포의 $\hat{\theta}_1$은 $\hat{\theta}_1 = 100$ 이 되고 암환자의 경우는 암표지자 평균이 0.1이므로 암환자 분포의 $\hat{\theta}_2$ 는 $\hat{\theta}_2 = 10$이 된다.

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암표지자 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$ 에서 환자의 검사결과로 부터 θ 를 추정하였더니 정상인인 경우$\hat{\theta}_1 = 100,$ 암환자의 경우$\hat{\theta}_2 = 10$ 으로 추정이 되었다. 암진단을 받으러 온 사람의 검사결과 $x = 0.06$ 으로 나왔다. 정상인을 암환자로 잘못 진단하였을 때 리스크를 1로 하고, 암환자를 정상인으로 잘 못 진단 하였을 때 리스크를 10으로 설정 하였다. 정확히 진단하였을 때 리스크는 0으로 한다. 이 리스크를 감안하여 암환자 여부를 진단하시오.

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암표지자 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$ 에서 환자의 검사결과로 부터 θ 를 추정하였더니 정상인인 경우$\hat{\theta}_1 = 100,$ 암환자의 경우$\hat{\theta}_2 = 10$ 으로 추정이 되었다. 암진단을 받으러 온 사람의 검사결과 $x = 0.06$ 으로 나왔다. 정상인을 암환자로 잘못 진단하였을 때 리스크를 1로 하고, 암환자를 정상인으로 잘 못 진단 하였을 때 리스크를 10으로 설정 하였다. 정확히 진단하였을 때 리스크는 0으로 한다. 이 리스크를 감안하여 암환자 여부를 진단하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암표지자 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$ 에서 환자의 검사결과로 부터 θ 를 추정하였더니 정상인인 경우$\hat{\theta}_1 = 100,$ 암환자의 경우$\hat{\theta}_2 = 10$ 으로 추정이 되었다. 암진단을 받으러 온 사람의 검사결과 $x = 0.06$ 으로 나왔다. 정상인을 암환자로 잘못 진단하였을 때 리스크를 1로 하고, 암환자를 정상인으로 잘 못 진단 하였을 때 리스크를 10으로 설정 하였다. 정확히 진단하였을 때 리스크는 0으로 한다. 이 리스크를 감안하여 암환자 여부를 진단하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.

  ✓ 정상인 확률은 $p(\hat{\theta}_1) = 0.99$ 이고 암환자의 확률은 $p(\hat{\theta}_2) = 0.01$ 이다.

  질문: $R(\alpha_1|x = 0.06) = \lambda_{11} p(\hat{\theta}_1|x = 0.06) + \lambda_{12} p(\hat{\theta}_2|x = 0.06) = ?$

  $R(\alpha_2|x = 0.06) = \lambda_{21} p(\hat{\theta}_1|x = 0.06) + \lambda_{22} p(\hat{\theta}_2|x = 0.06) = ?$

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암표지자 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$ 에서 환자의 검사결과로 부터 θ 를 추정하였더니 정상인인 경우$\hat{\theta}_1 = 100,$ 암환자의 경우$\hat{\theta}_2 = 10$ 으로 추정이 되었다. 암진단을 받으러 온 사람의 검사결과 $x = 0.06$ 으로 나왔다. 정상인을 암환자로 잘못 진단하였을 때 리스크를 1로 하고, 암환자를 정상인으로 잘 못 진단 하였을 때 리스크를 10으로 설정 하였다. 정확히 진단하였을 때 리스크는 0으로 한다. 이 리스크를 감안하여 암환자 여부를 진단하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 정상인 확률은 $p(\hat{\theta}_1) = 0.99$ 이고 암환자의 확률은 $p(\hat{\theta}_2) = 0.01$ 이다.

    질문: $R(\alpha_1|x = 0.06) = \lambda_{11} p(\hat{\theta}_1|x = 0.06) + \lambda_{12} p(\hat{\theta}_2|x = 0.06) = ?$

    $R(\alpha_2|x = 0.06) = \lambda_{21} p(\hat{\theta}_1|x = 0.06) + \lambda_{22} p(\hat{\theta}_2|x = 0.06) = ?$
  - ✓ 정상인 확률은 $p(\hat{\theta}_1) = 0.99$ 이고 암환자의 확률은 $p(\hat{\theta}_2) = 0.01$ 이다.

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암표지자 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$ 에서 환자의 검사결과로 부터 θ 를 추정하였더니 정상인인 경우$\hat{\theta}_1 = 100,$ 암환자의 경우$\hat{\theta}_2 = 10$ 으로 추정이 되었다. 암진단을 받으러 온 사람의 검사결과 $x = 0.06$ 으로 나왔다. 정상인을 암환자로 잘못 진단하였을 때 리스크를 1로 하고, 암환자를 정상인으로 잘 못 진단 하였을 때 리스크를 10으로 설정 하였다. 정확히 진단하였을 때 리스크는 0으로 한다. 이 리스크를 감안하여 암환자 여부를 진단하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.
  - ✓ 정상인 확률은 $p(\hat{\theta}_1) = 0.99$ 이고 암환자의 확률은 $p(\hat{\theta}_2) = 0.01$ 이다.

    질문: $R(\alpha_1|x = 0.06) = \lambda_{11} p(\hat{\theta}_1|x = 0.06) + \lambda_{12} p(\hat{\theta}_2|x = 0.06) = ?$

    $R(\alpha_2|x = 0.06) = \lambda_{21} p(\hat{\theta}_1|x = 0.06) + \lambda_{22} p(\hat{\theta}_2|x = 0.06) = ?$

  - ✓ $p(\hat{\theta}_1|x = 0.06) \propto p(x = 0.06|\hat{\theta}_1) p(\hat{\theta}_1) = 100 * 0.06 e^{-100*0.06} * 0.99 = 0.0147$
  - ✓ $p(\hat{\theta}_2|x = 0.06) \propto p(x = 0.06|\hat{\theta}_2) p(\hat{\theta}_2) = 10 * 0.06 e^{-10*0.06} * 0.01 = 0.00329$

Exercise

- 서울대학교병원에 암진단을 받으러 온 사람은 10만명이다. 그중 1000명이 암환자로 판명이 난다. 암표지자 확률 분포는 $p(x|\theta) = \theta x e^{-\theta x}, \quad for\ x > 0\ and\ \theta > 0$ 에서 환자의 검사결과로 부터 θ 를 추정하였더니 정상인인 경우$\hat{\theta}_1 = 100$, 암환자의 경우$\hat{\theta}_2 = 10$ 으로 추정이 되었다. 암진단을 받으러 온 사람의 검사결과 $x = 0.06$ 으로 나왔다. 정상인을 암환자로 잘못 진단하였을 때 리스크를 1로 하고, 암환자를 정상인으로 잘 못 진단 하였을 때 리스크를 10으로 설정 하였다. 정확히 진단하였을 때 리스크는 0으로 한다. 이 리스크를 감안하여 암환자 여부를 진단하시오.

- Sol. (힌트) 문장의 수치에 해당하는 내용과 질문을 수식으로 표현해 보세요.

  ✓ 정상인 확률은 $p(\hat{\theta}_1) = 0.99$ 이고 암환자의 확률은 $p(\hat{\theta}_2) = 0.01$ 이다.

  질문: $R(\alpha_1|x = 0.06) = \lambda_{11} p(\hat{\theta}_1|x = 0.06) + \lambda_{12} p(\hat{\theta}_2|x = 0.06) = ?$

  $R(\alpha_2|x = 0.06) = \lambda_{21} p(\hat{\theta}_1|x = 0.06) + \lambda_{22} p(\hat{\theta}_2|x = 0.06) = ?$

  ✓ 정상인 확률은 $p(\hat{\theta}_1) = 0.99$ 이고 암환자의 확률은 $p(\hat{\theta}_2) = 0.01$ 이다.

  ✓ $p(\hat{\theta}_1|x = 0.06) \propto p(x = 0.06|\hat{\theta}_1) p(\hat{\theta}_1) = 100 * 0.06 e^{-100*0.06} * 0.99 = 0.0147$

  ✓ $p(\hat{\theta}_2|x = 0.06) \propto p(x = 0.06|\hat{\theta}_2) p(\hat{\theta}_2) = 10 * 0.06 e^{-10*0.06} * 0.01 = 0.00329$

  ✓ $R(\alpha_1|x = 0.06) = 0 * 0.0147 + 10 * 0.00329 = 0.0329$

  $R(\alpha_2|x = 0.06) = 1 * 0.0147 + 0 * 0.00329 = 0.0147$

Error Probabilities and Integrals

- Consider the 2-class problem and suppose that the feature space is divided into 2 regions R_1 and R_2. There are 2 ways in which a classification error can occur.
  - An observation x falls in R_2, and the true state is ω_1.
  - An observation x falls in R_1, and the true state is ω_2.

- The error probability

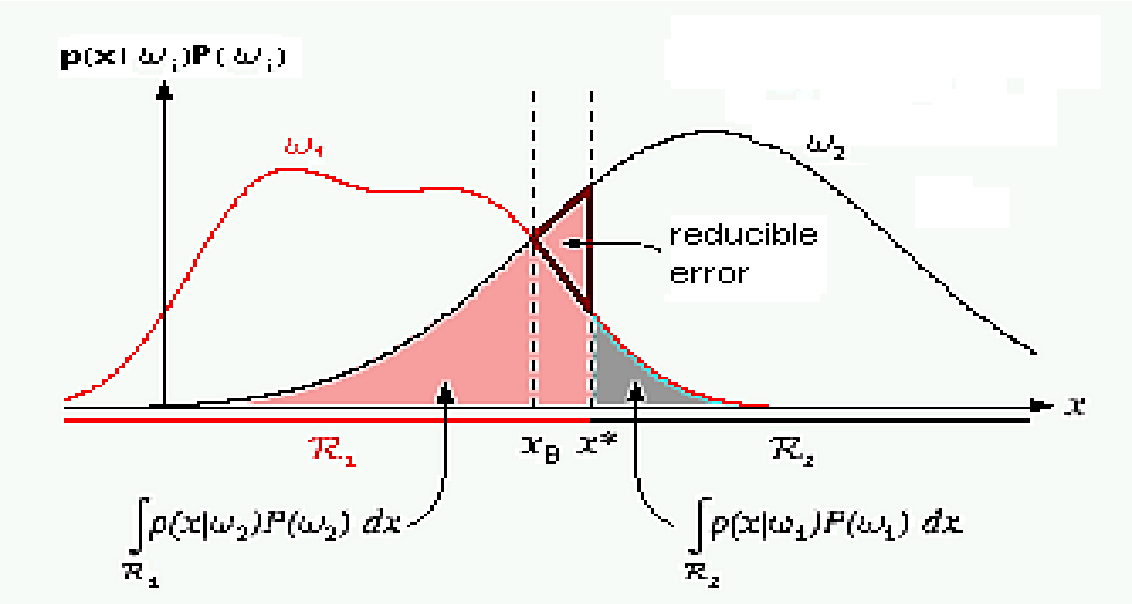
$$P(error) = P(x \in R_2|\omega_1)p(\omega_1) + P(x \in R_1|\omega_2)p(\omega_2)$$

$$= \int_{R_2} p(x|\omega_1)p(\omega_1)dx + \int_{R_1} p(x|\omega_2)p(\omega_2)dx$$

Error Probabilities and Integrals

- Because x^* is chosen arbitrarily, the probability of error is not as small as it might be.
- x_B = Bayes optimal decision boundary , and gives the lowest probability of error.
- Bayes classifier maximizes the correct probability.

$$P(correct) = \sum_{i=1}^{C} P(\mathbf{x} \in \Re_i \mid \omega_i) p(\omega_i) = \sum_{i=1}^{C} \int_{\Re_i} p(\mathbf{x} \mid \omega_i)\, p(\omega_i) d\mathbf{x}$$



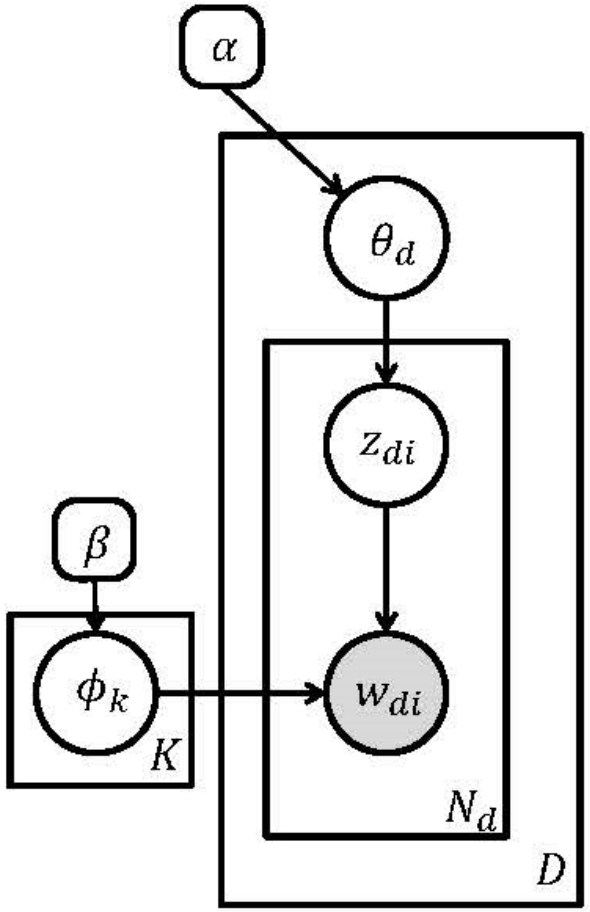
Summary

- Bayes Formula
  - Priori probability
  - Likelihood
  - Posterior Probability
  - Bayes Decision
- Risk Formulation
  - Conditional Risk
  - Likelihood Ratio Test
  - Zero-one Loss Function (Bayes Decision)
- Maximum Likelihood Estimation (MLE)
- Error Probability

Bayesian Networks

Jin Young Choi

Seoul National University

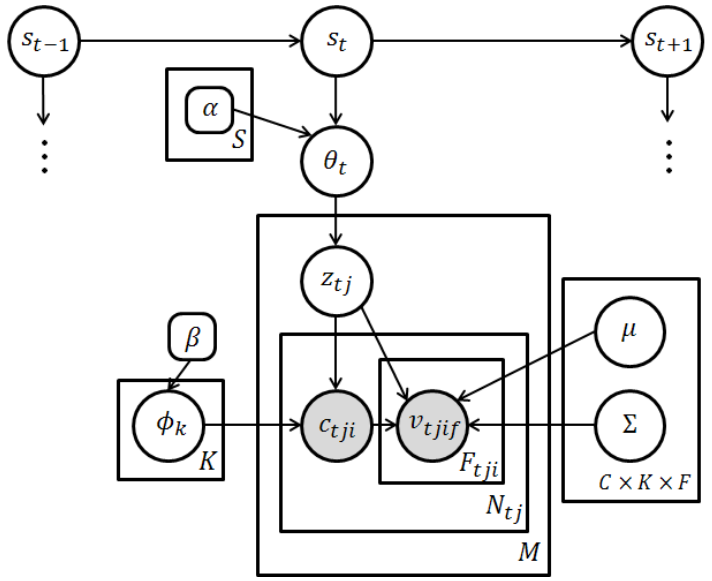


Outline

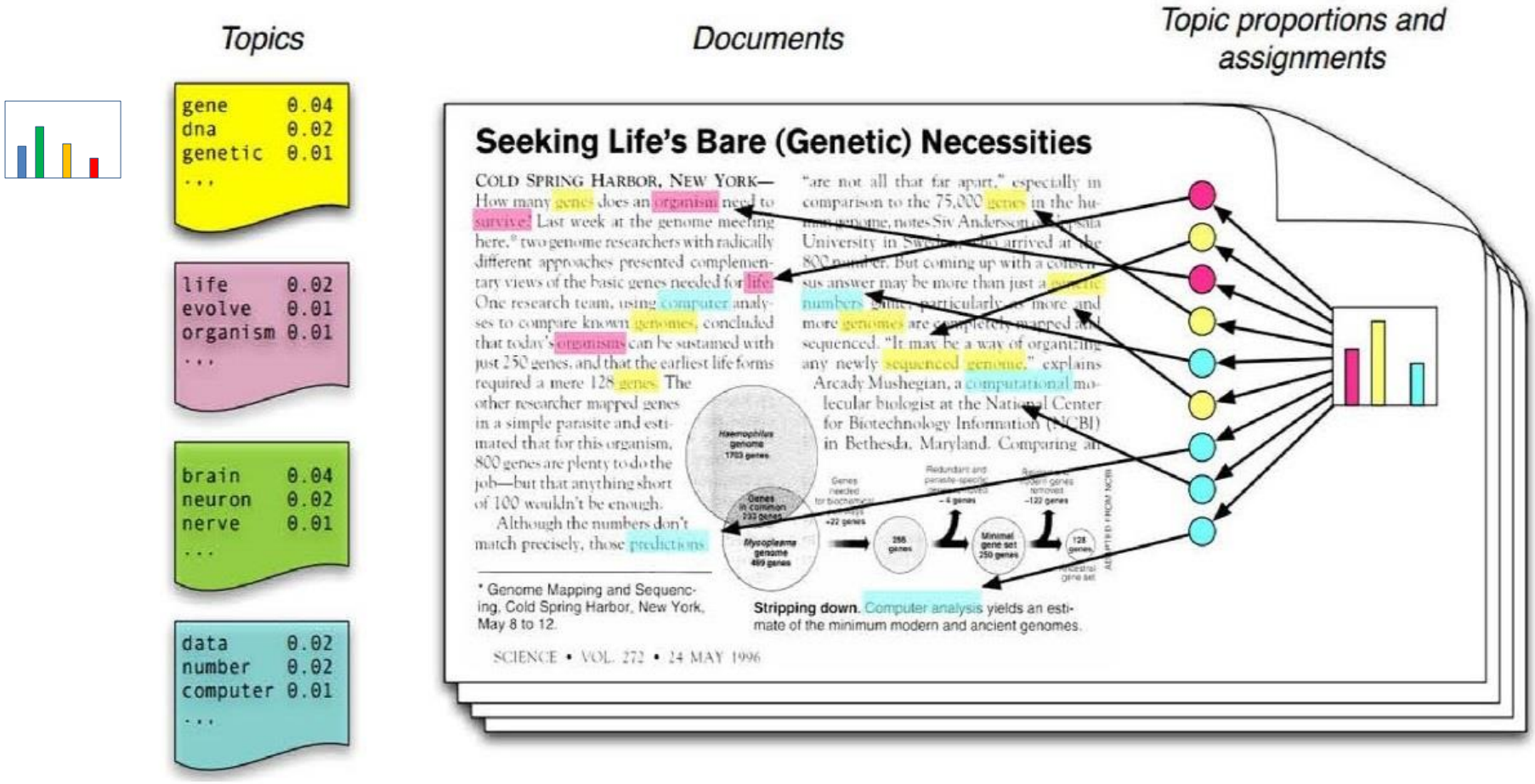
- Application Examples of Bayesian Networks
  - Traffic Pattern Analysis
  - Topic Model (Document Analysis)
- Directed Acyclic Graph
- Conditional Independence
- D-separation
- Bayesian Parameters
- Parameterized Conditional Distributions
- Multinomial, Dirichlet Distribution, Conjugate Prior
- Markov Blanket

Application: Traffic Pattern Analysis

- Surveillance in crowded scenes

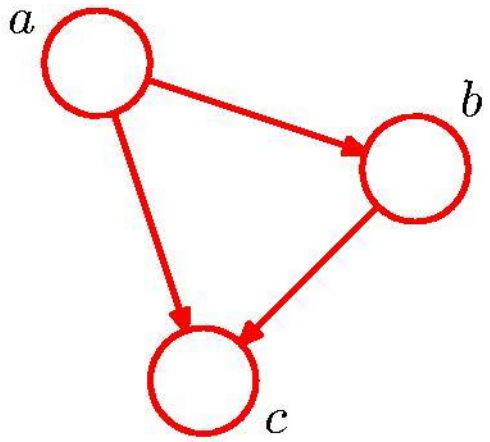


LDA Model (Topic Modelling)



Bayesian Networks

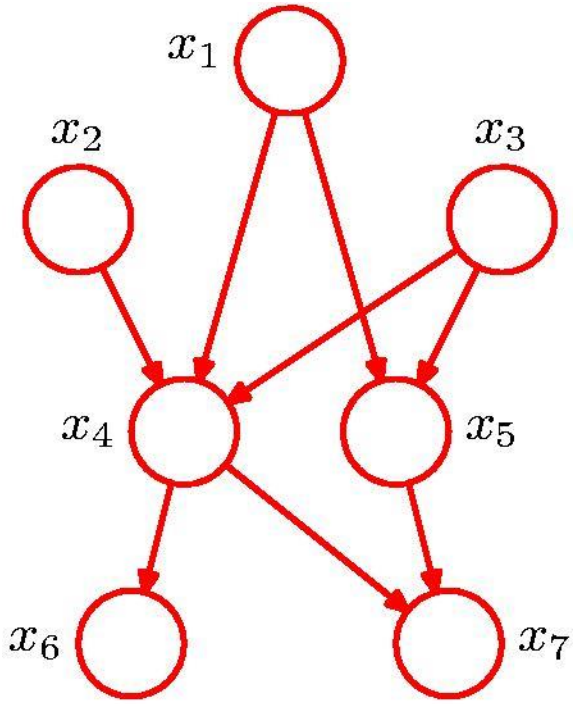
- Directed Acyclic Graph (DAG)



$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$$

Bayesian Networks



$$p(x_1,\ldots,x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)$$
$$p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

General Factorization

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Conditional Independence

- a is independent of b given c

$$p(a|b,c) = p(a|c)$$

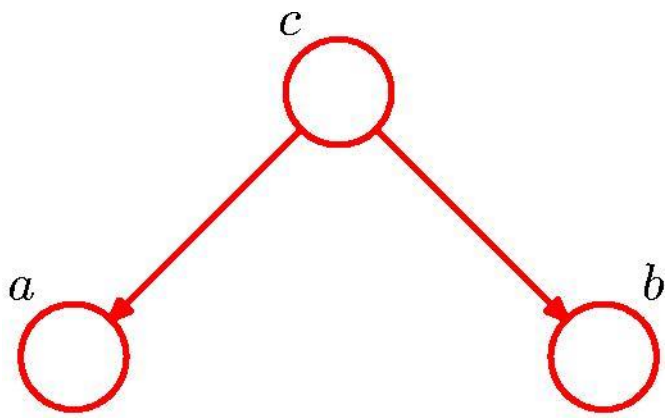
- Equivalently

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}$$

- Notation

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 1



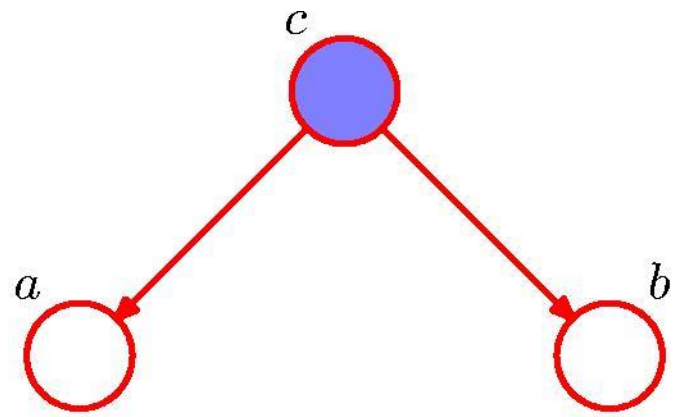
$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

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

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

U, V , and c are independent. $a = U + c, b = V + c$; a, b independent?

Conditional Independence: Example 1



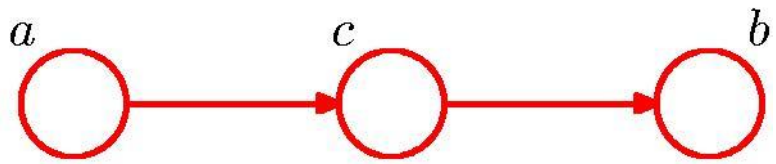
$$p(a,b,c) = p(a|c)p(b|c)p(c)$$

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

$$a \perp\!\!\!\perp b \mid c$$

U, V , and c are independent. $a = U + c, b = V + c, c = 1$; a, b independent?

Conditional Independence: Example 2



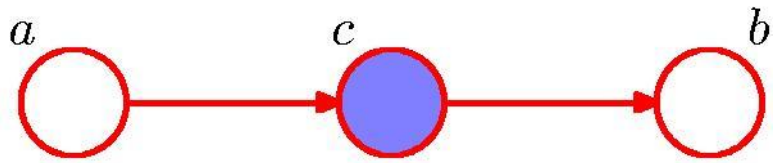
$$p(a,b,c) = p(a)p(c|a)p(b|c)$$

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

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

$$p(b,c|a) = p(c|a)p(b|a,c) = p(c|a)p(b|c)$$

Conditional Independence: Example 2

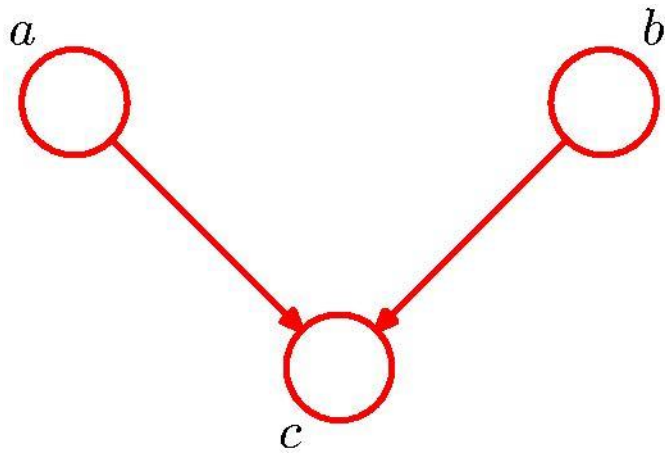


$$
\begin{aligned}
p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\
&= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

$$p(a|c) = \frac{p(c|a)p(a)}{p(c)}$$

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 3



$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

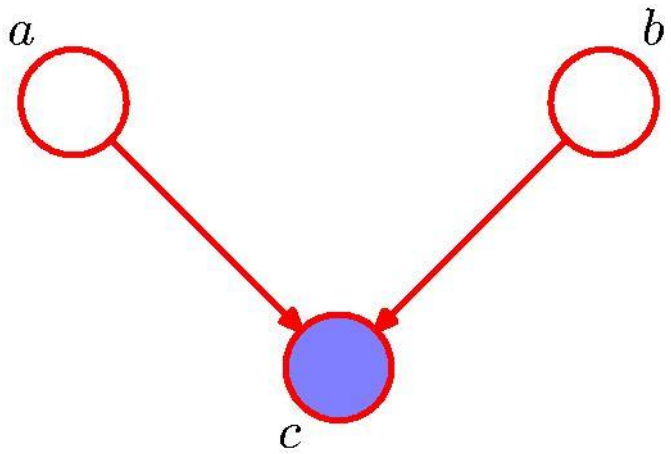
$$p(a, b) = p(a)p(b)$$

$$a \perp\!\!\!\perp b \mid \emptyset$$

Note: this is the opposite of Example 1, with c unobserved.

a and b are independent Bernoulli rvs. $c = a + b$

Conditional Independence: Example 3



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

$$a \not\perp\!\!\!\perp b \mid c$$

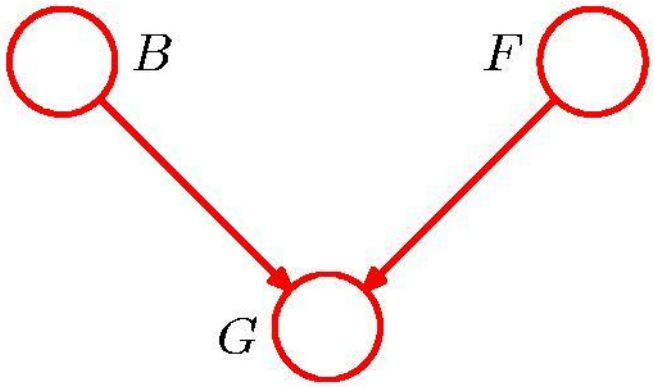
Note: this is the opposite of Example 1, with C observed.

a and b are independent Bernoulli rvs. $c = a + b$

“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

- G is dependent to B and F



$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$

- F is independent to B

B = Battery (0=flat, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading (0=empty, 1=full)

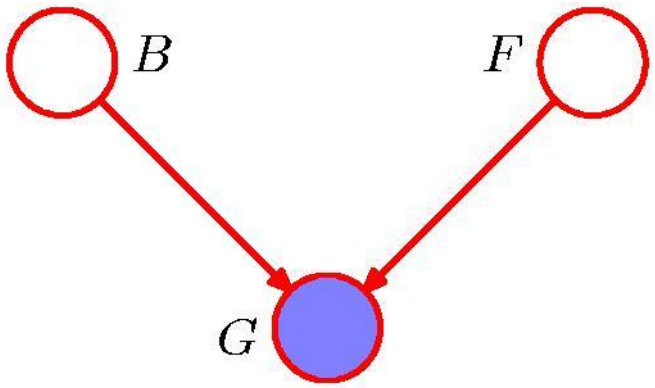
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$



Probability of an empty tank increased by observing $G = 0$.

수식화?

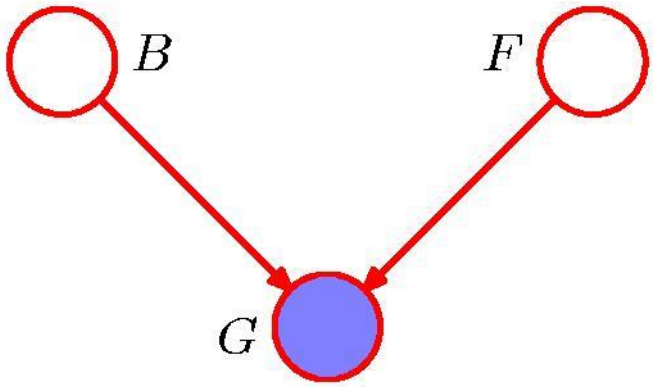
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$



Probability of an empty tank increased by observing $G = 0$.

$$
p(F=0|G=0) =
$$

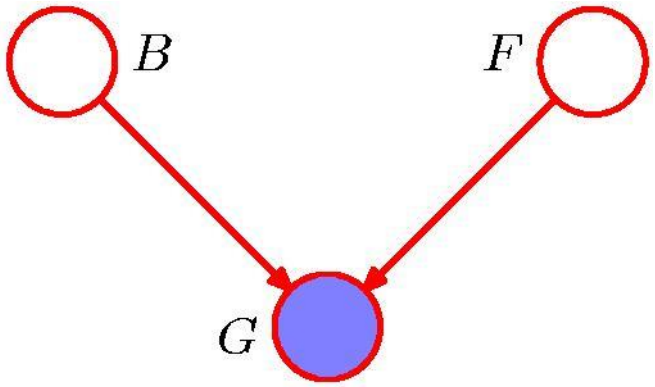
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$



Probability of an empty tank increased by observing $G = 0$.

$$
p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)}
$$

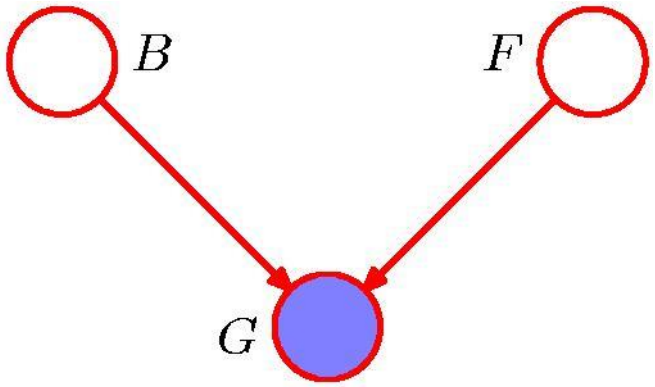
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8\\
p(G=1|B=1,F=0) &= 0.2\\
p(G=1|B=0,F=1) &= 0.2\\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9\\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$p(F=0) = 0.1$$



Probability of an empty tank increased by observing $G = 0$.

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)}$$

$P(G=0|F=0) = 1 - P(G=1|F=0) = ?$

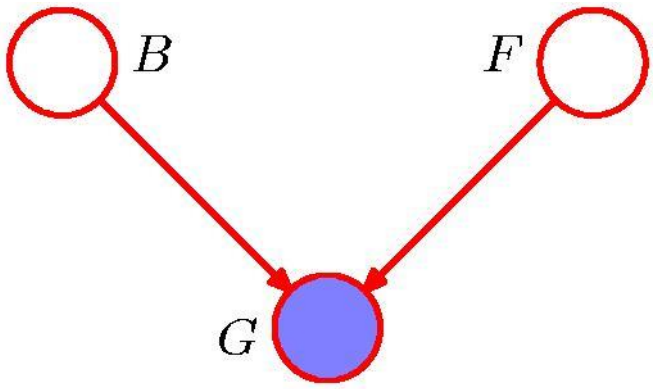
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8\\
p(G=1|B=1,F=0) &= 0.2\\
p(G=1|B=0,F=1) &= 0.2\\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9\\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$



Probability of an empty tank increased by observing $G = 0$.

$$
p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)}
$$

$$
P(G=0|F=0) = 1 - P(G=1|F=0) = 1 - \sum_B P(G=1|B,F=0)P(B) = 1 - 0.2 \times 0.9 - 0.1 \times 0.1 = 0.81
$$

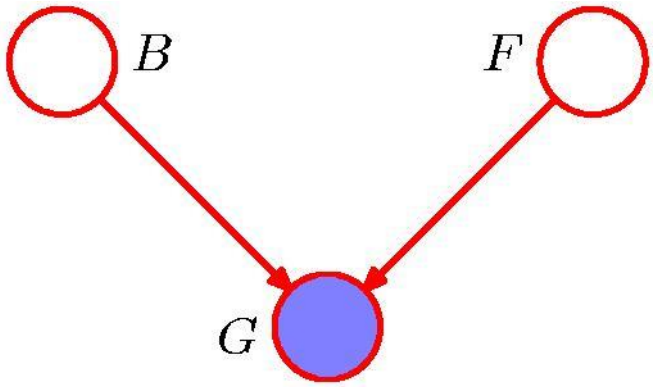
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8\\
p(G=1|B=1,F=0) &= 0.2\\
p(G=1|B=0,F=1) &= 0.2\\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9\\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$p(F=0) = 0.1$$



Probability of an empty tank increased by observing $G = 0$.

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)}$$

$$P(G=0|F=0) = 1 - P(G=1|F=0) = 1 - \sum_B P(G=1|B,F=0)P(B) = 1 - 0.2 \times 0.9 - 0.1 \times 0.1 = 0.81$$

$$P(G=0) = 1 - P(G=1) = 1 - \sum_{B,F} P(G=1|B,F)P(B,F) = 1 - 0.8 \times 0.81 - 0.2 \times 0.09 - 0.2 \times 0.09 - 0.1 \times 0.01 = 0.315$$

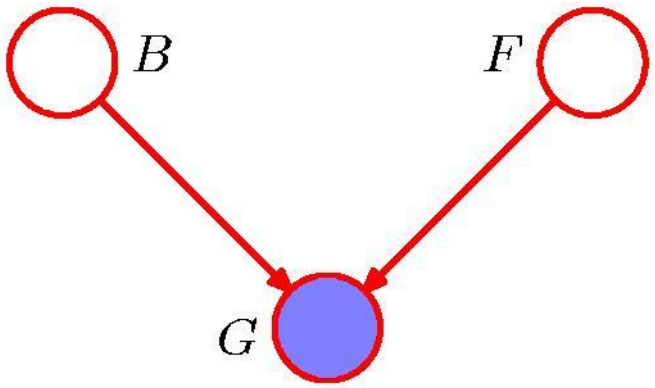
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8\\
p(G=1|B=1,F=0) &= 0.2\\
p(G=1|B=0,F=1) &= 0.2\\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9\\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$p(F=0) = 0.1$$



Probability of an empty tank increased by observing $G = 0$.

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)} = \frac{0.81 \times 0.1}{0.315} = 0.257$$

$$P(G=0|F=0) = 1 - P(G=1|F=0) = 1 - \sum_B P(G=1|B,F=0)P(B) = 1 - 0.2 \times 0.9 - 0.1 \times 0.1 = 0.81$$

$$P(G=0) = 1 - P(G=1) = 1 - \sum_{B,F} P(G=1|B,F)P(B,F) = 1 - 0.8 \times 0.81 - 0.2 \times 0.09 - 0.2 \times 0.09 - 0.1 \times 0.01 = 0.315$$

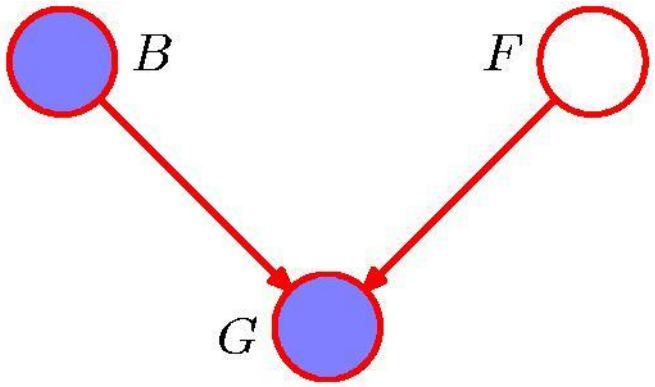
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$



Probability of an empty tank reduced by observing $B = 0 \,\&\, G = 0$
This referred to as “explaining away”.

어떤 영향을 받을 까요? 1) 영향 받지 않는다. 2) Empty 확률이 높아진다. 3) Empty 확률이 낮아진다.

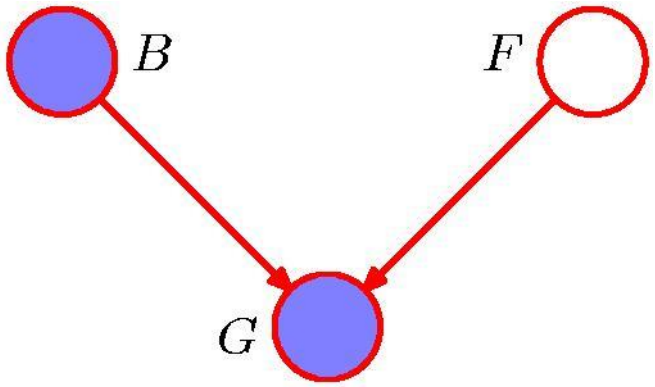
“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$



Probability of an empty tank reduced by observing $B = 0 \,\&\, G = 0$

This referred to as “explaining away”.

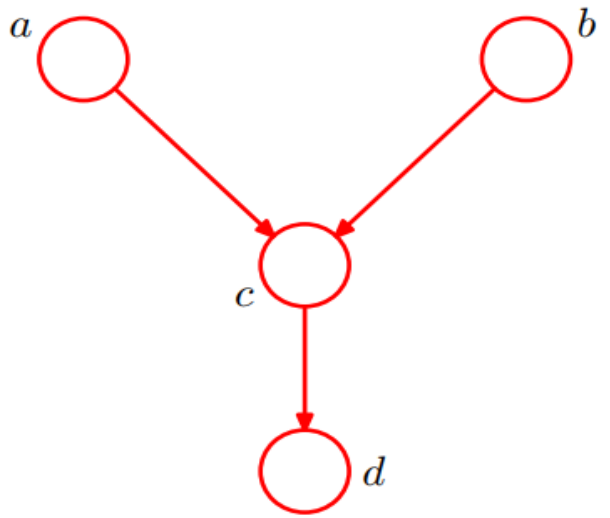
$$
\begin{aligned}
p(F=0|G=0,B=0) &= \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)} \\
&\simeq 0.111
\end{aligned}
$$

- F is dependent to B given G and the empty probability is reduced because the gage becomes less reliable.

Exercise

Answer the following questions for the right-hand Bayesian network.

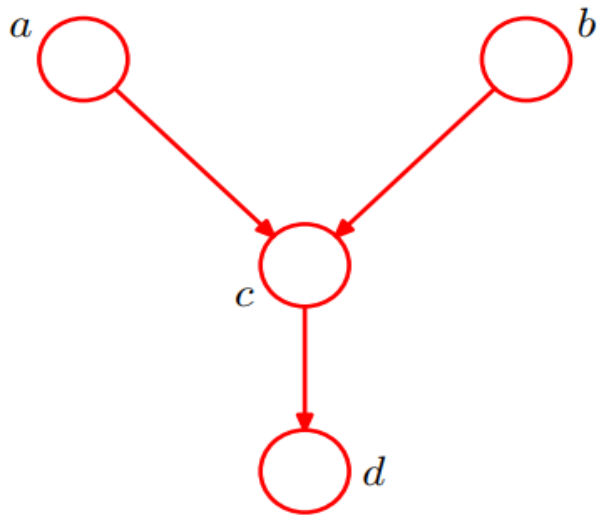
1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



1) nothing observed

$$p(a,b) = p(a)p(b) \ ?$$

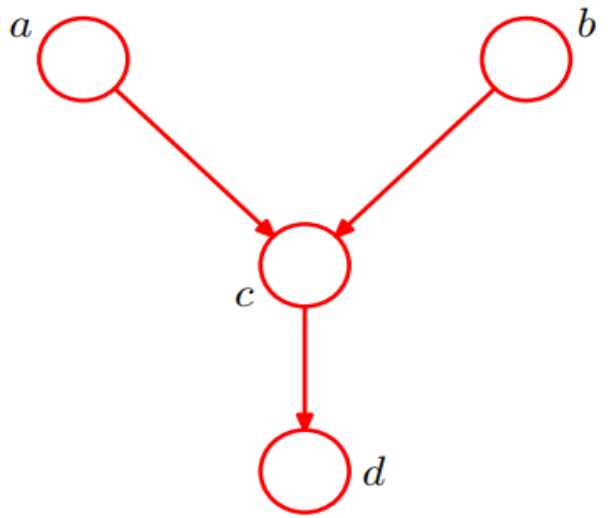
2) d is observed

$$p(a,b|d) \neq p(a|d)p(b|d) \ ?$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



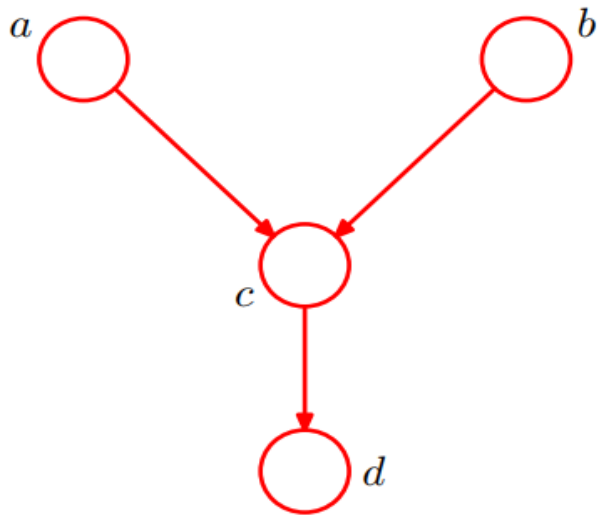
1) Nothing observed

$$p(a,b) = \Sigma_c \Sigma_d p(a,b,c,d) =$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



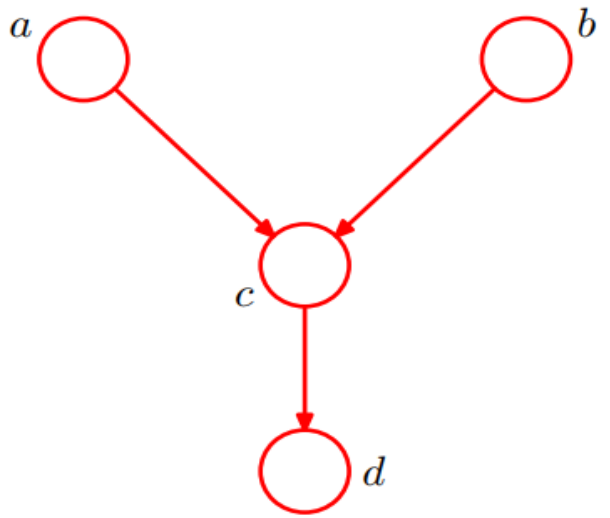
1) Nothing observed

$$p(a,b) = \Sigma_c \Sigma_d p(a,b,c,d) = \Sigma_c \Sigma_d \, p(a)p(b)p(c|a,b)p(d|c)$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



1) Nothing observed

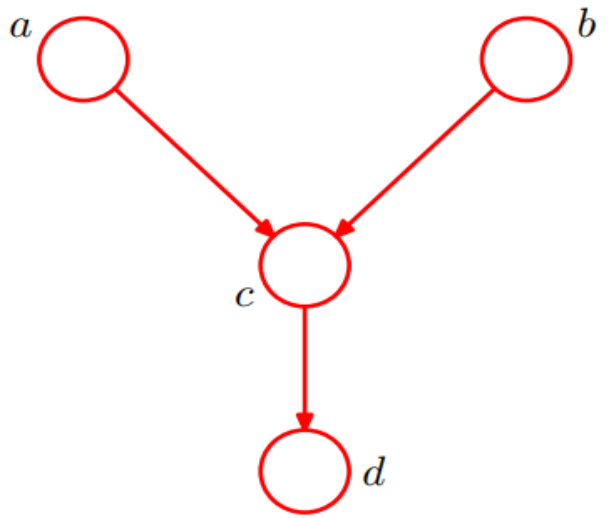
$$p(a,b) = \Sigma_c \Sigma_d p(a,b,c,d) = \Sigma_c \Sigma_d \, p(a)p(b)p(c|a,b)p(d|c)$$

$$= p(a)p(b)\Sigma_c \Sigma_d \, p(c|a,b)p(d|c)$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



1) Nothing observed

$$p(a,b) = \Sigma_c \Sigma_d p(a,b,c,d) = \Sigma_c \Sigma_d \, p(a)p(b)p(c|a,b)p(d|c)$$

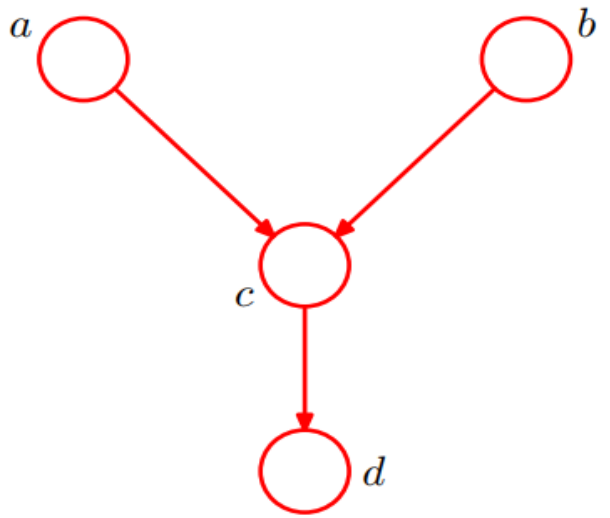
$$= p(a)p(b)\Sigma_c \Sigma_d \, p(c|a,b)p(d|c)$$

$$= p(a)p(b)\Sigma_c p(c|a,b)\Sigma_d \, p(d|c)$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



1) Nothing observed

$$p(a,b) = \Sigma_c \Sigma_d p(a,b,c,d) = \Sigma_c \Sigma_d \, p(a)p(b)p(c|a,b)p(d|c)$$

$$= p(a)p(b)\Sigma_c \Sigma_d \, p(c|a,b)p(d|c)$$

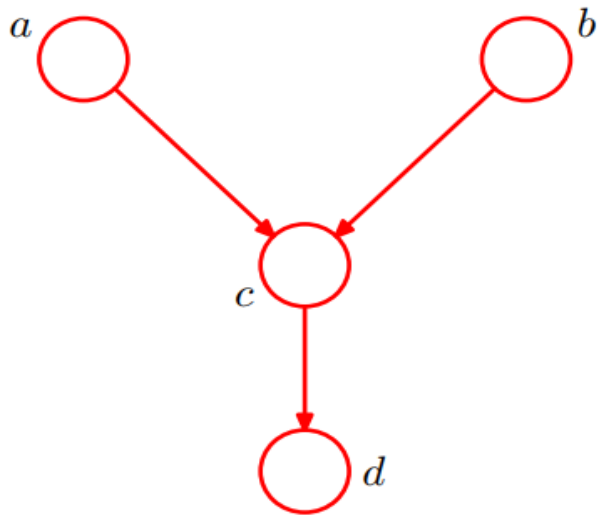
$$= p(a)p(b)\Sigma_c p(c|a,b)\Sigma_d \, p(d|c)$$

$$= p(a)p(b)$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



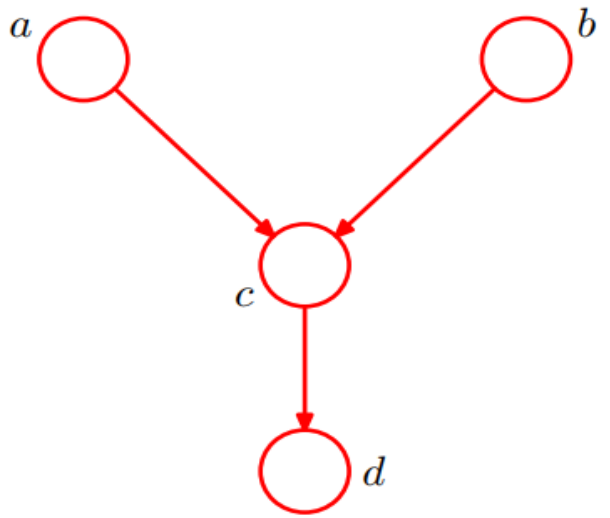
2) d is observed

$$p(a, b|d) = \frac{p(a,b,d)}{p(d)}$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



2) d is observed

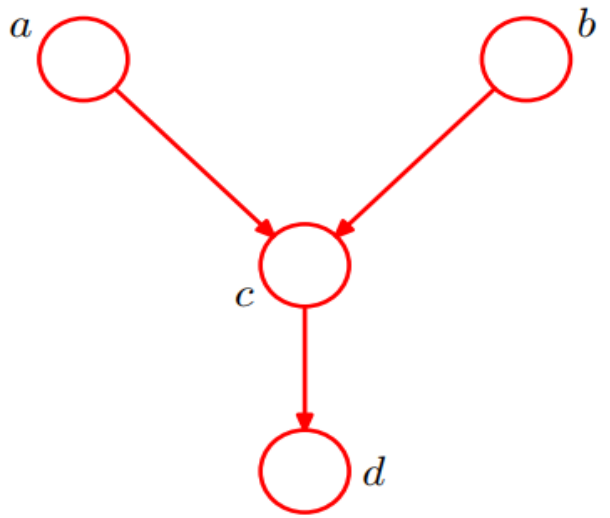
$$p(a,b|d) = \frac{p(a,b,d)}{p(d)}$$

$$= \Sigma_c \left\{ \frac{p(a,b,c,d)}{p(d)} \right\}$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



2) d is observed

$$p(a,b|d) = \frac{p(a,b,d)}{p(d)}$$

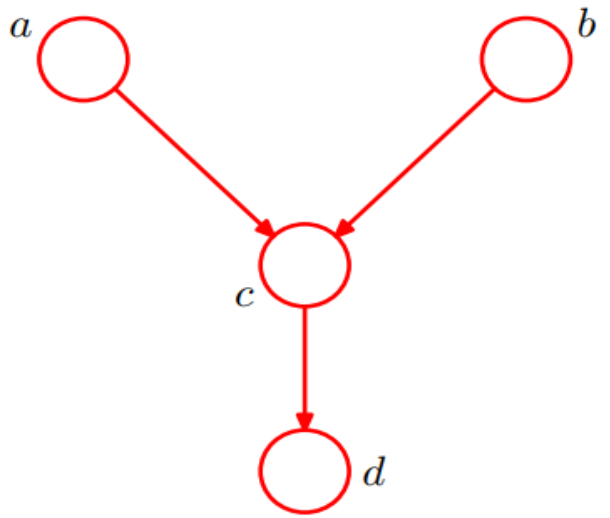
$$= \Sigma_c \left\{ \frac{p(a,b,c,d)}{p(d)} \right\}$$

$$= \Sigma_c \left\{ \frac{p(a)p(b)p(c|a,b)p(d|c)}{p(d)} \right\}$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



2) d is observed

$$p(a,b|d) = \frac{p(a,b,d)}{p(d)}$$

$$= \Sigma_c \left\{ \frac{p(a,b,c,d)}{p(d)} \right\}$$

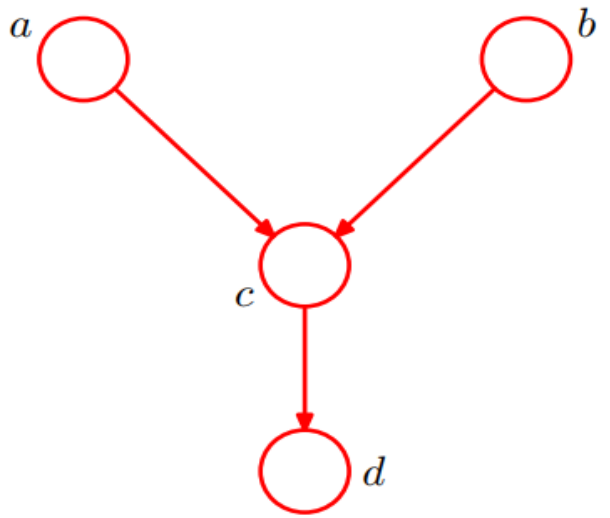
$$= \Sigma_c \left\{ \frac{p(a)p(b)p(c|a,b)p(d|c)}{p(d)} \right\}$$

$$= \frac{p(a)p(b)}{p(d)} \Sigma_c \{ p(c|a,b)p(d|c) \}$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



2) d is observed

$$p(a,b|d) = \frac{p(a,b,d)}{p(d)}$$

$$= \Sigma_c \left\{ \frac{p(a,b,c,d)}{p(d)} \right\}$$

$$= \Sigma_c \left\{ \frac{p(a)p(b)p(c|a,b)p(d|c)}{p(d)} \right\}$$

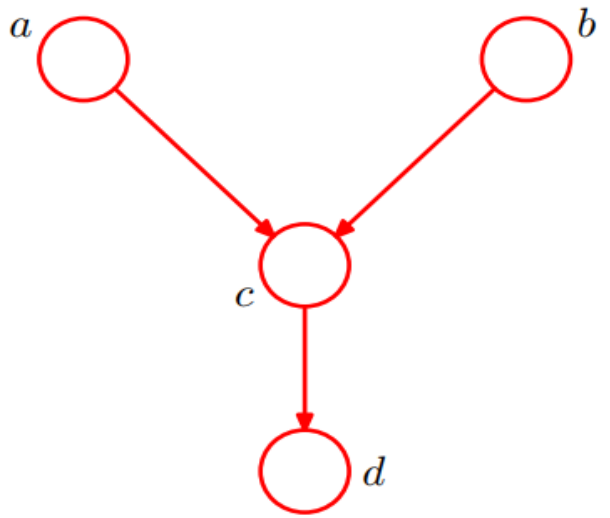
$$= \frac{p(a)p(b)}{p(d)} \Sigma_c \{ p(c|a,b)p(d|c) \}$$

$$= \frac{p(a)p(b)p(d|a,b)}{p(d)}$$

Exercise

Answer the following questions for the right-hand Bayesian network.

1) When any random variables are not observed, show that a and b are independent to each other.
2) When d is observed, show that a and b are dependent to each other.



2) d is observed

$$p(a,b|d) = \frac{p(a,b,d)}{p(d)}$$

$$= \Sigma_c \left\{ \frac{p(a,b,c,d)}{p(d)} \right\}$$

$$= \Sigma_c \left\{ \frac{p(a)p(b)p(c|a,b)p(d|c)}{p(d)} \right\}$$

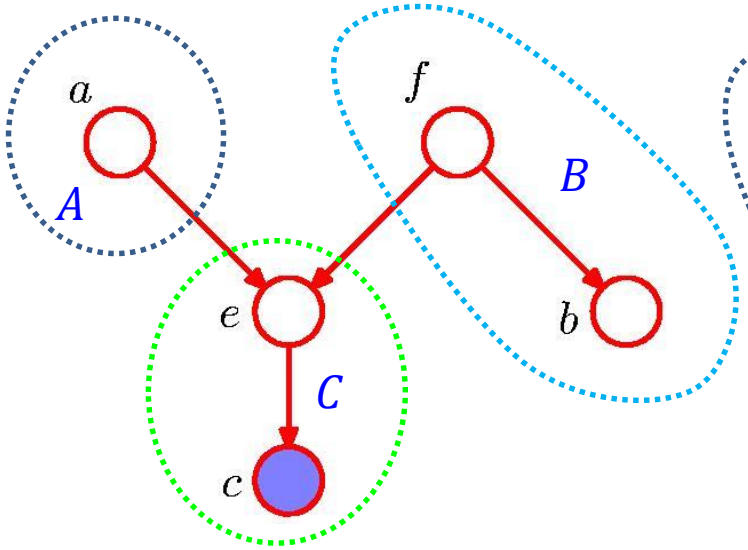
$$= \frac{p(a)p(b)}{p(d)} \Sigma_c \{ p(c|a,b)p(d|c) \}$$

$$= \frac{p(a)p(b)p(d|a,b)}{p(d)} \neq p(a|d)p(b|d)$$

D-separation

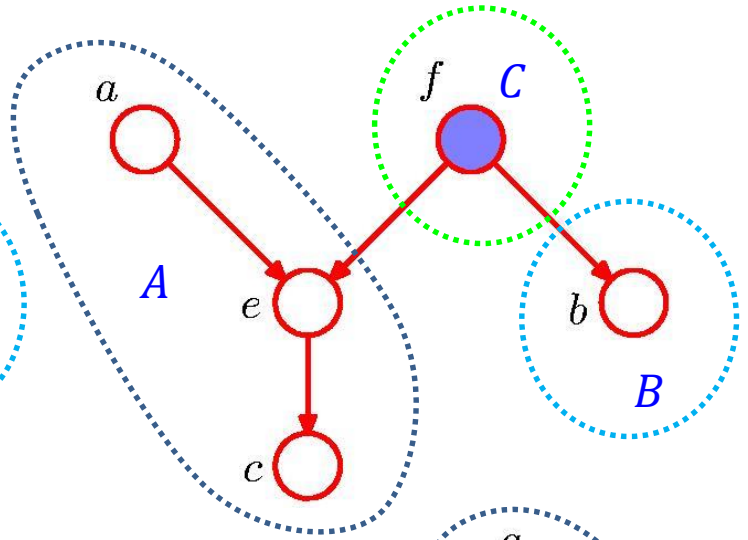
- A, B, and C are non-intersecting subsets of nodes in a directed graph.
- A path from A to B is blocked if it contains a node such that either
  - the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or
  - the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set C.
- If all paths from A to B are blocked, A is said to be d-separated from B by C.
- If A is d-separated from B by C, the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$.

D-separation: Example



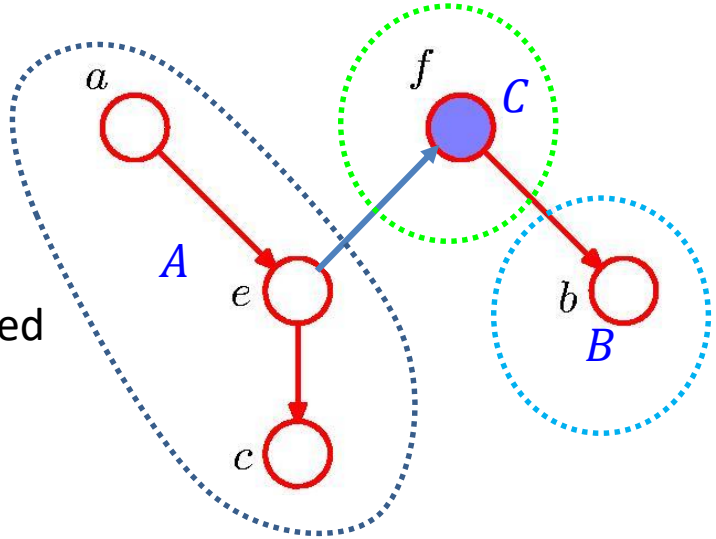
A to B is unblocked

$$a \not\perp\!\!\!\perp b \mid c$$



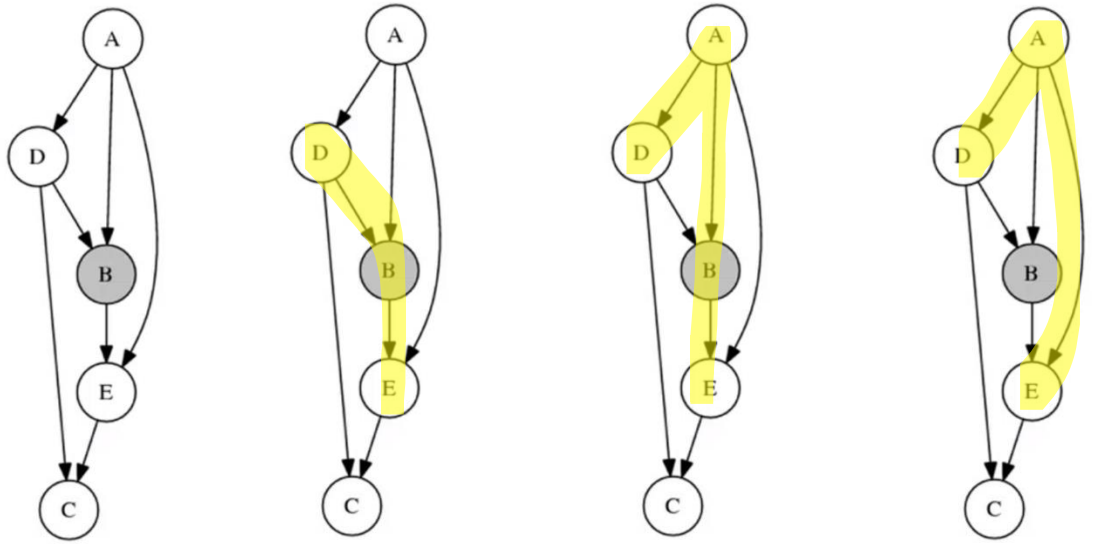
$$a \perp\!\!\!\perp b \mid f$$

A to B is blocked
(d-separated)



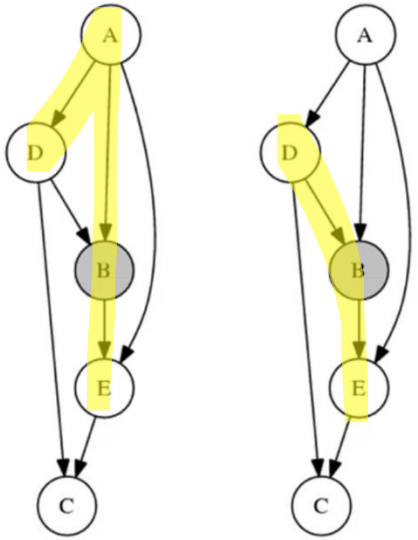
Exercise

When B is observed in the following Bayesian network, decide whether every path from D to E is blocked (d-separated) or not and determine the dependency between D and E.



Exercise

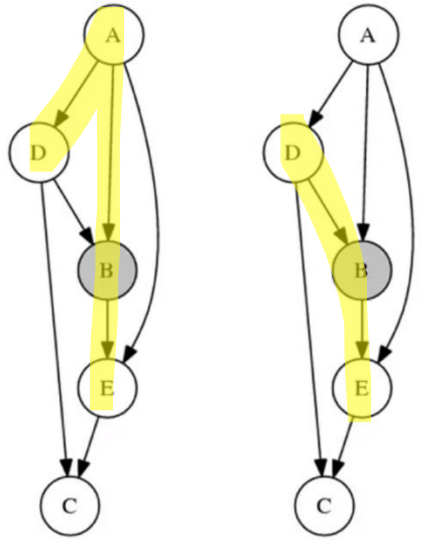
a. path 1: (D←A→B→E) or (D → B → E)
   Is blocked (d-separated) or not?



Exercise

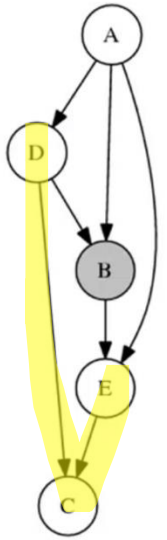
a. path 1: path via B, i.e., (D←A→B→E) or (D → B → E)
Is blocked (d-separated) or not?

Answer: blocked (d-separated),
since the connection in B is head to tail and B is observed.



Exercise

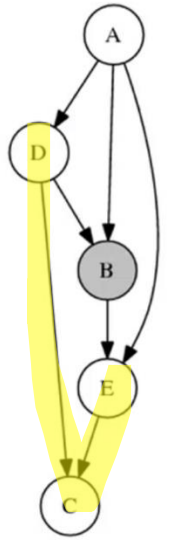
b. path 2 (D → C ← E)
   Is blocked (d-separated) or not?



Exercise

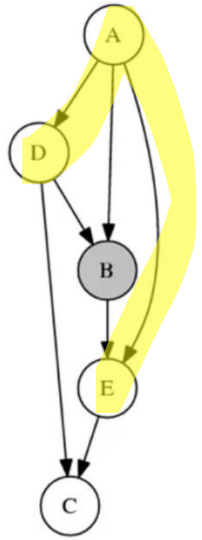
b. path 2 (D → C ← E)
Is blocked (d-separated) or not?

Answer: blocked (d-separated),
since the connection in C is head to head and C is not observed.



Exercise

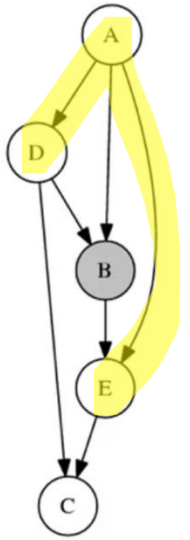
c. path 3 (D ← A→ E) :
Is blocked (d-separated) or not?



Exercise

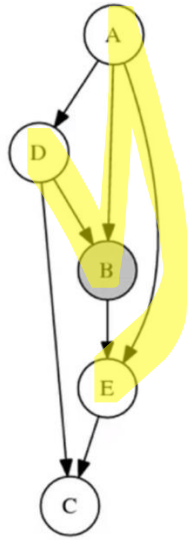
c. path 3 (D ← A→ E) :
Is blocked (d-separated) or not?

Answer: not blocked,
since the connection in A is tail to tail and A is not observed.



Exercise

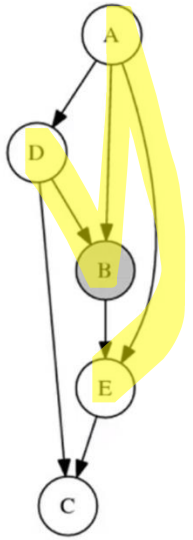
d. path 4 (D → B ← A → E) :
Is blocked (d-separated) or not?



Exercise

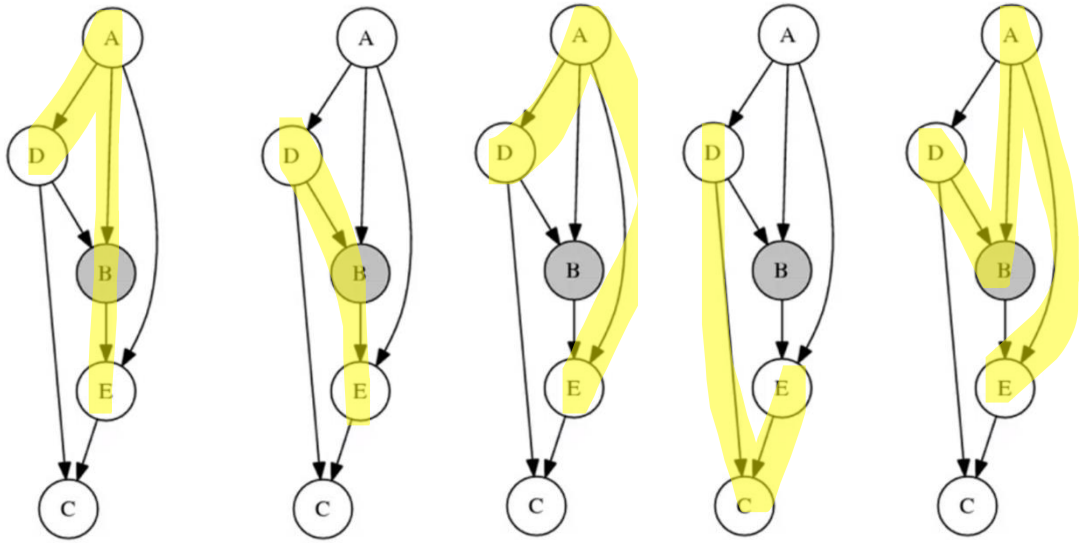
d. path 4 (D → B ← A → E) :
Is blocked (d-separated) or not?

Answer: not blocked,
since the connection in B is head to head and B is observed, the path D → B ← A becomes non-blocking by B, and
since the connection in A is tail to tail and A is not observed, the path B ← A → E becomes non-blocking .



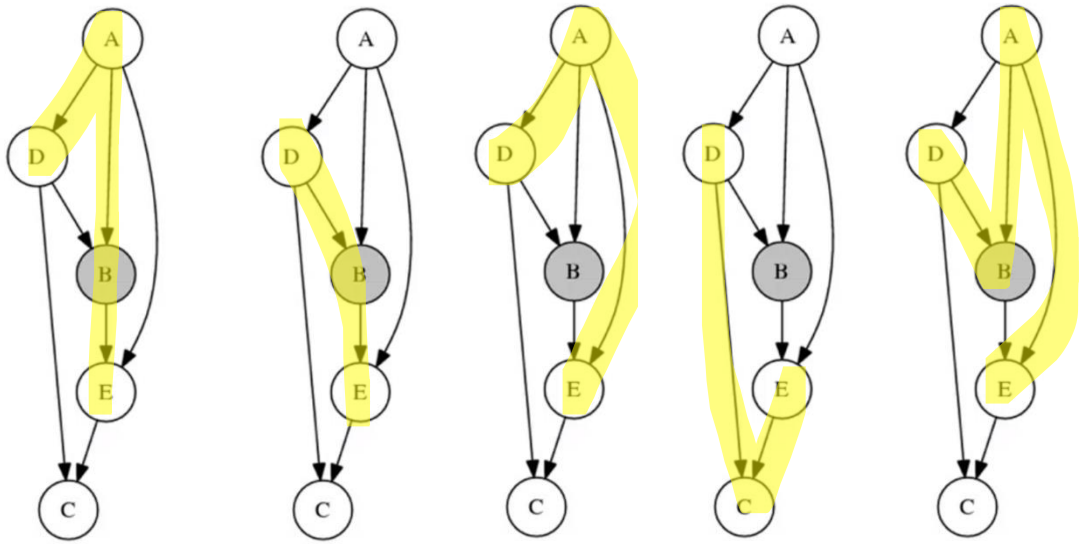
Exercise

e. There are 4 blocked paths and 1 non-blocked path.
Is blocked (d-separated) or not?



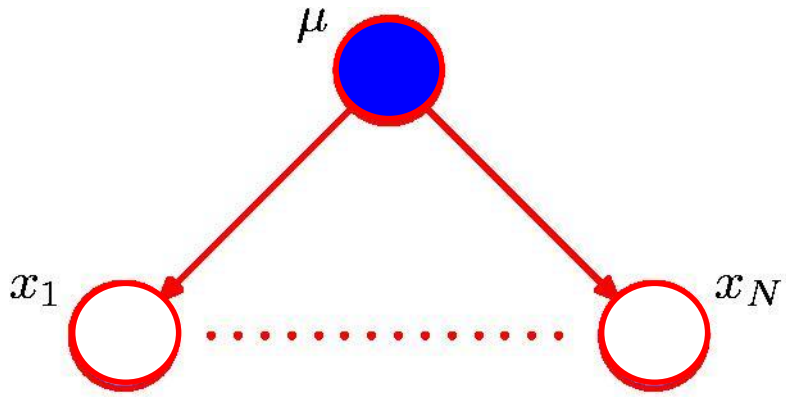
Exercise

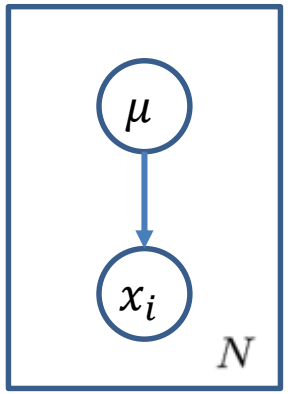
e. There are 4 blocked paths and 1 non-blocked path.
Is blocked (d-separated) or not?



Answer: not blocked (d-separated),
since there exists at least one non-blocking path.
Thus D and E are dependent to each other.

D-separation: I.I.D. Data





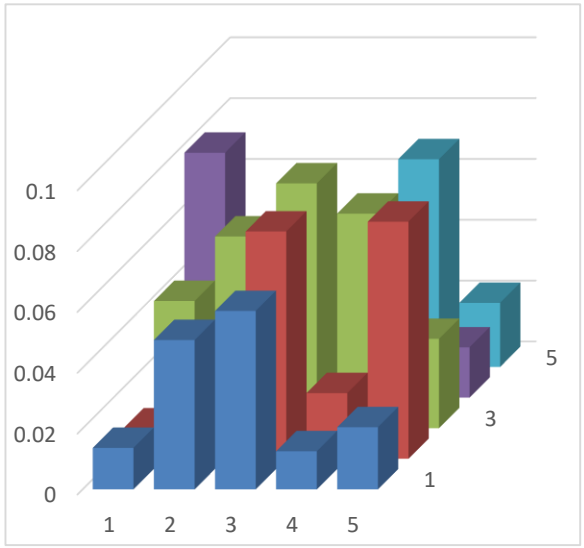
$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu)\,\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n)$$

Discrete Variables, Multinomial

$$p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$$

$$p(x_{11}, \dots, x_{1K}, x_{21}, \dots, x_{2K} | \mu_{11}, \dots, \mu_{KK}) = \frac{n!}{x_{11}! \dots x_{1K}!} \frac{n!}{x_{21}! \dots x_{2K}!} \mu_{11}^{x_{11} x_{21}} \dots \mu_{KK}^{x_{1K} x_{2K}}$$



Discrete Variables (1), Multinomial

- General joint distribution: $K^2 - 1$ parameters

$$p(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\mu}) = \prod_{k=1}^{K}\prod_{l=1}^{K} \mu_{kl}^{x_{1k}x_{2l}}$$

- Independent joint distribution: $2(K-1)$ parameters

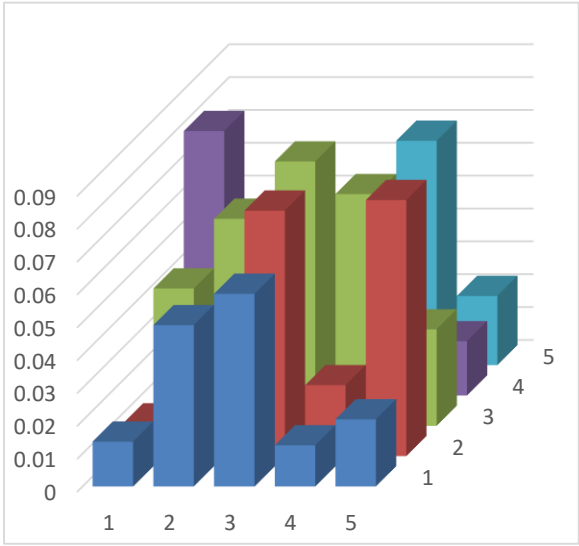
$$\hat{p}(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_{1k}^{x_{1k}} \prod_{l=1}^{K} \mu_{2l}^{x_{2l}}$$

- $p(x_1, x_2) = p(x_1|x_2)p(x_2)$

  $K - 1 + K(K-1) = K^2 - 1$ parameters

- $p(x_1, x_2) = p(x_1)p(x_2)$

  $K - 1 + K - 1 = 2(K-1)$ parameters

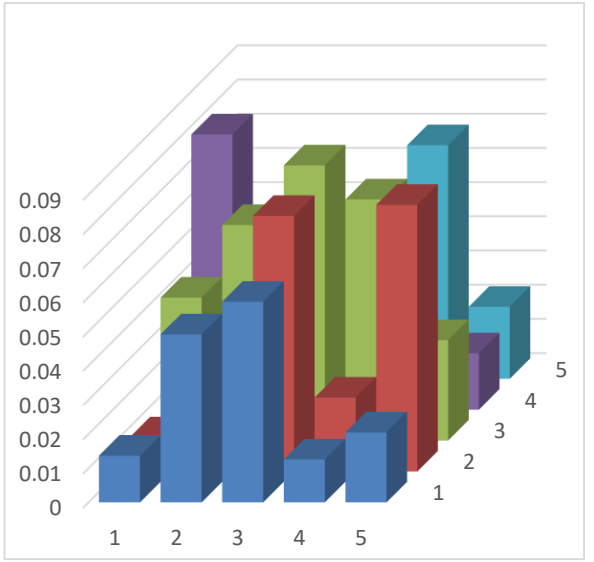


Discrete Variables, Dirichlet

- The posterior distributions are in the same family as the prior probability distribution.

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

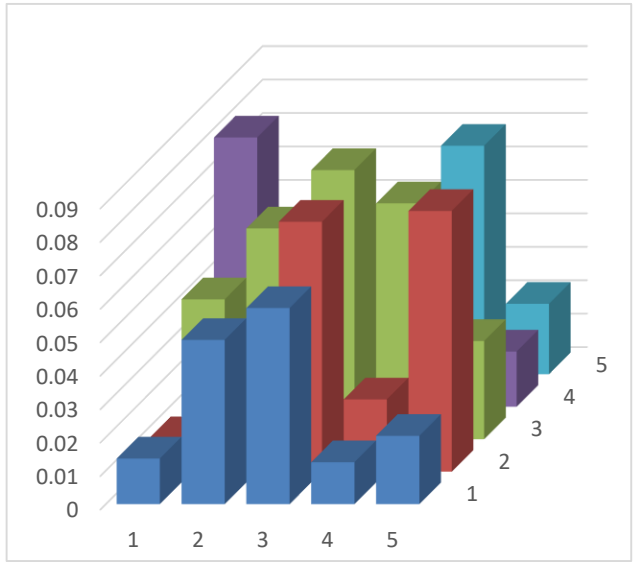
- The prior and posterior are then called **conjugate distributions,** and the prior is called a **conjugate prior** for the likelihood function.
- Dirichlet distribution is a conjugate (prior) distribution to the multinomial distribution.
- Gaussian is a conjugate prior of Gaussian.



Discrete Variables, Dirichlet

- Posteriori: $p(\mu|x,\alpha) \propto p(x|\mu)p(\mu|\alpha)$
- Mul(K, μ): $p(x_1,\dots,x_K|\mu_1,\dots,\mu_K)=\frac{n!}{x_1!\dots x_K!}\mu_1^{x_1} \dots \mu_K^{x_K}$
- Dir(K, α): $p(\mu_1,\dots,\mu_K|\alpha_1,\dots,\alpha_K)=\frac{\Gamma(\sum_{i=1}^{K}(\alpha_i-1))}{\prod_{i=1}^{K}\Gamma(\alpha_i-1)}\mu_1^{\alpha_1} \dots \mu_K^{\alpha_K}$
- Parameters: $\alpha_1,\dots,\alpha_K > 0$ (hyper-parameters)
- Support: $\mu_1,\dots,\mu_K \in (0,1)$ where $\sum_{i=1}^{K}\mu_i = 1$
- Dir(K, c + α): $p(\mu|x,\alpha) \propto p(x|\mu)p(\mu|\alpha)$

  where $c = (c_1,\dots,c_K)$ is number of occurrences
- $E[\mu_k] = \frac{c_k+\alpha_k}{\sum_{i=1}^{K}(c_i+\alpha_i)}$

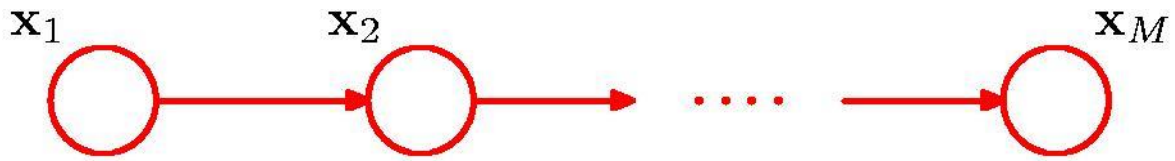


Discrete Variables (2)

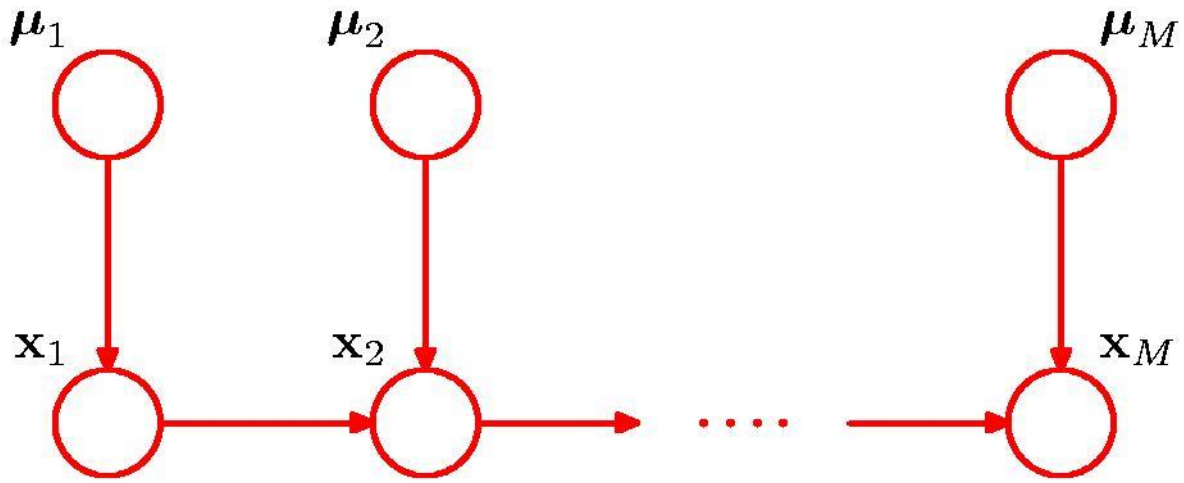
- General joint distribution over M variables: $K^M - 1$ parameters

- M -node Markov chain: $K - 1 + (M - 1)K(K - 1)$ parameters

$$p(x_1, x_2) = p(x_1)p(x_2|x_1)p(x_3|x_2) \ldots p(x_M|x_{M-1})$$



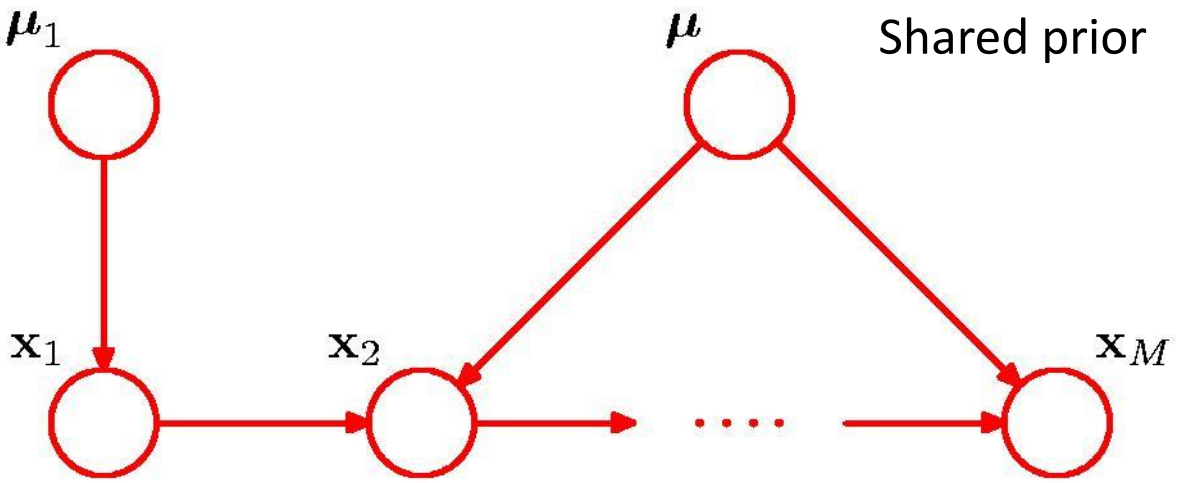
Discrete Variables: Bayesian Parameters (1)



$$p\left(\{\mathbf{x}_m, \boldsymbol{\mu}_m\}\right) = p\left(\mathbf{x}_1 \,|\boldsymbol{\mu}_1\right) p\left(\boldsymbol{\mu}_1\right) \prod_{m=2}^{M} p\left(\mathbf{x}_m|\mathbf{x}_{m-1}, \boldsymbol{\mu}_m\right) p\left(\boldsymbol{\mu}_m\right)$$

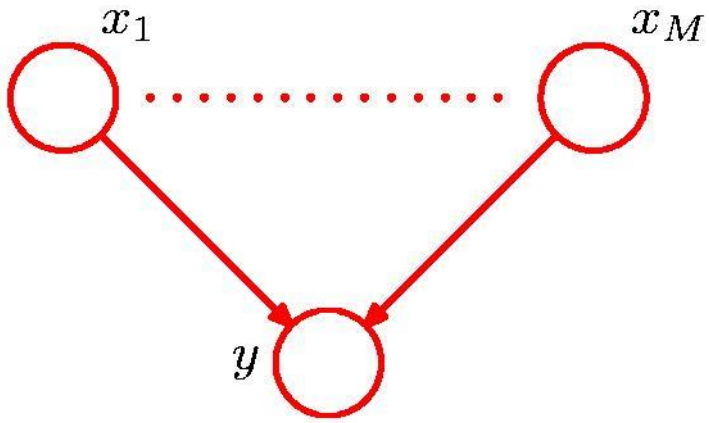
$$p(\boldsymbol{\mu}_m) = \mathrm{Dir}(\boldsymbol{\mu}_m|\boldsymbol{\alpha}_m)$$

Discrete Variables: Bayesian Parameters (2)



$$p\left(\{\mathbf{x}_m\}, \boldsymbol{\mu}_1, \boldsymbol{\mu}\right) = p\left(\mathbf{x}_1 \,|\boldsymbol{\mu}_1\right) p\left(\boldsymbol{\mu}_1\right) \prod_{m=2}^{M} p\left(\mathbf{x}_m|\mathbf{x}_{m-1}, \boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}\right)$$

Parameterized Conditional Distributions



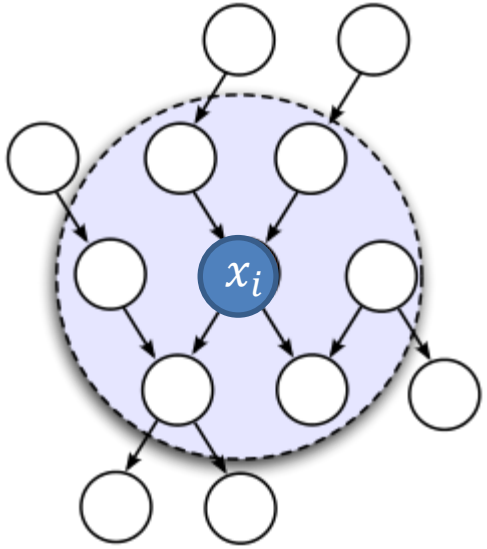
If $x_1, \ldots, x_M$ are discrete, K-state variables, $p(y = 1|x_1, \ldots, x_M)$ in general has $O(K^M)$ parameters because $p(x_1, \ldots, x_M|y = 1)$ requires $K^M - 1$ parameters.

The parameterized form

$$p(y = 1|x_1, \ldots, x_M) = \sigma\left(w_0 + \sum_{i=1}^{M} w_i x_i\right) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

requires only $M + 1$ parameters (actually this can not model a probability distribution).

The Markov Blanket



$$p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_M)}{\int p(\mathbf{x}_1,\ldots,\mathbf{x}_M)\,\mathrm{d}\mathbf{x}_i}$$

$$= \frac{\prod_k p(\mathbf{x}_k|\mathrm{pa}_k)}{\int \prod_k p(\mathbf{x}_k|\mathrm{pa}_k)\,\mathrm{d}\mathbf{x}_i}$$

$$= \prod_{k\in MB} p(x_k|pa_k)$$

Any factor $p(x_k|pa_k)$ that does not have any functional dependence on x_i can be taken outside the integral over x_i, and will therefore cancel between numerator and denominator.

Summary

- Bayesian Networks
- Directed Acyclic Graph
- Conditional Independence
- D-separation
- Bayesian Parameters
- Parameterized Conditional Distributions
- Multinomial, Dirichlet Distribution, Conjugate Prior
- Markov Blanket