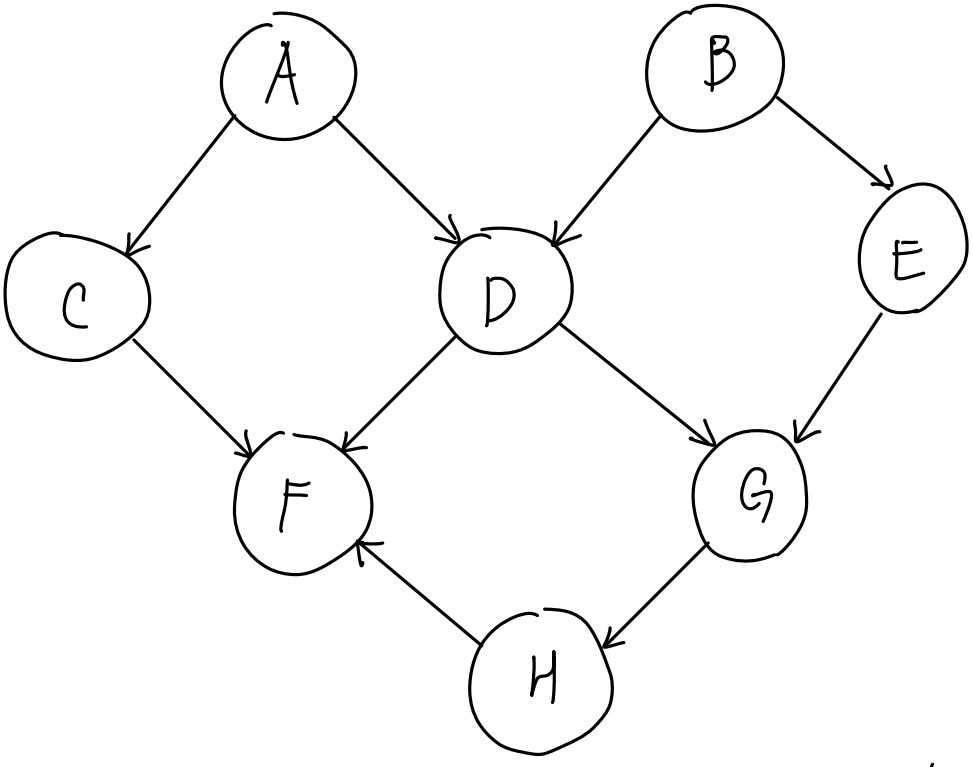
[1] [Bayesian Network]

20170243 강재윤

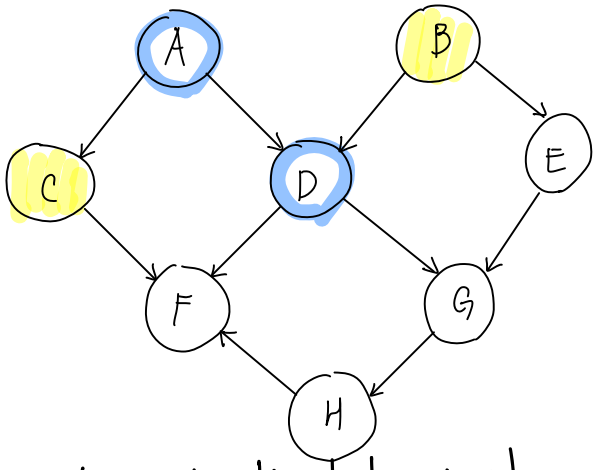


$(X \perp Y \mid Z)$: if X and Y are (conditionally) independent given Z.

(a) every pair of (marginally) independent variables.

(A, B) (A, E) (B, C) (C, E)

(b) $(B \perp C \mid A, D)$



To check the conditional independence in directed graph, we should consider all paths from B to C and determine whether the path is blocked by A and D according to D-Seperation.

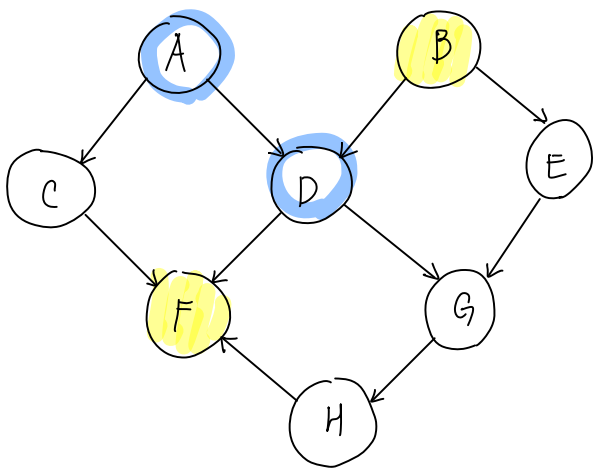
Path that passes through A: A is tail-to-tail case → block.
ex> B-D-A-C

Path that passes through D: D is head-to-tail case → block
ex> B-D-F-C

Path that passes through F: F is head-to-head case & not observed → block
ex> B-E-G-H-F-C

∴ independent

(c) $(B \not\perp F \mid A, D)$



Path that passes through D: D is head-to-tail case → block
ex> B-D-F

Path that passes E, G, H: we can reach from B to F
ex> B-E-G-H-F

∴ not independent

(d)

① We assume that each variable can take one of 10 values. Then, if we do not use the knowledge encoded in graph, the number of parameters we need to model the full joint distribution is $10^8 - 1$. The reason why we subtract 1 is that the sum of all probabilities is 1, so we can get it by subtracting the rest from 1. So we will skip the last entry, so we don't need it.

② We should first joint probability with the factorization. Then, we can write the factorization form as follow. Then, the number of parameters we have to need is decreasing compared to ①.

$$P(A,B,C,D,E,F,G,H)$$
$$= P(A)P(B)P(C|A)P(D|A,B)P(E|B)P(F|C,D,H)P(G|D,E)P(H|G)$$

10^0	10^0	10^1	10^2	10^1	10^3	10^2	10^1
=	=	=	=	=	=	=	=
1	1	10	100	10	1000	100	10

$$1+1+10+100+10+1000+100+10 = 1232$$

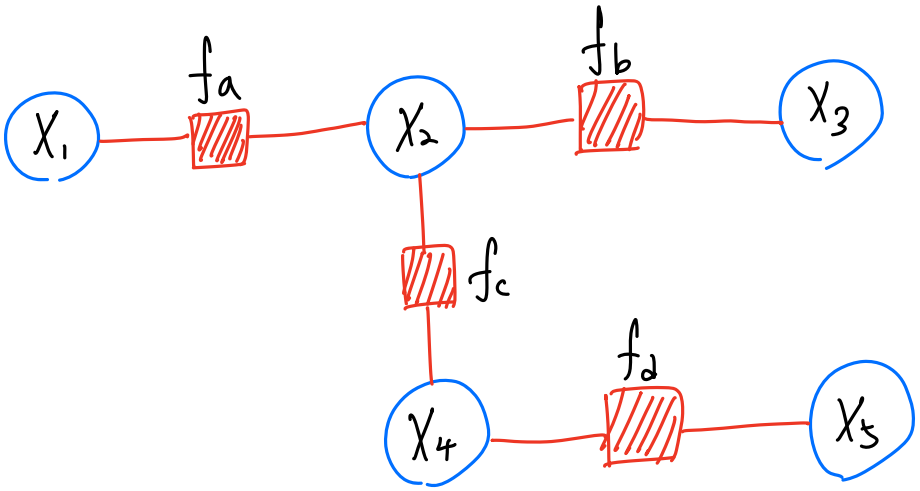
$$\therefore (10-1) \times 1232 = 11088$$

2. [Belief Propagation]

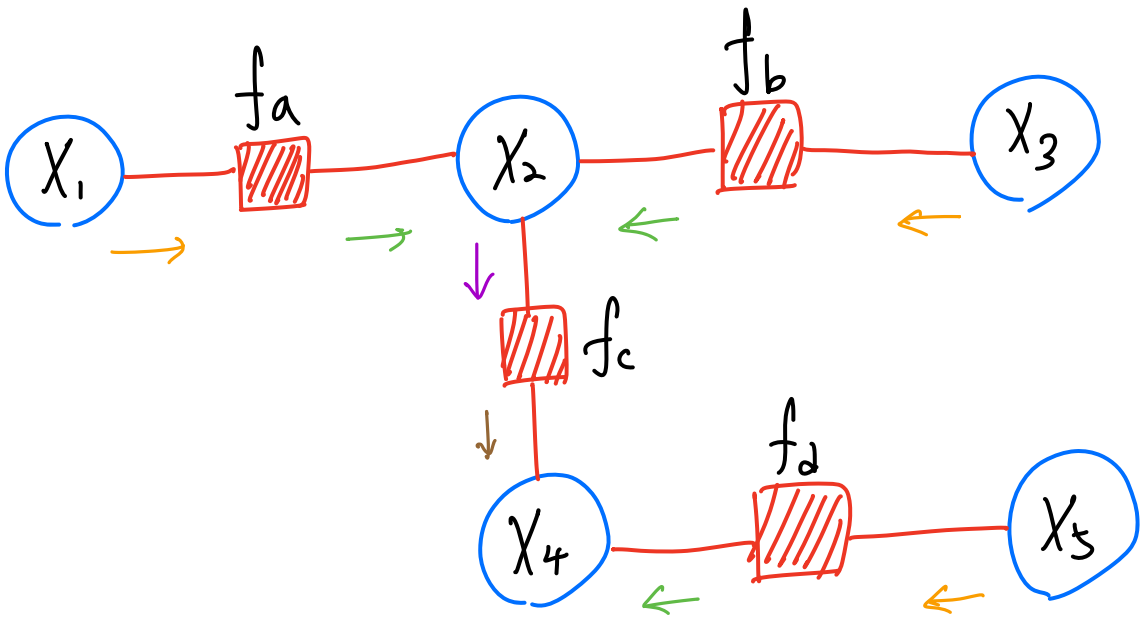
binary random variables $X_1, \ldots, X_5 \in \{0, 1\}$

$$p(X) = p(X_1, X_2, X_3, X_4, X_5) = \frac{1}{Z} f_a(X_1, X_2)\, f_b(X_2, X_3)\, f_c(X_2, X_4)\, f_d(X_4, X_5)$$

(a) corresponding factor graph.



(b) sum-product belief propagation → compute $p(X_4)$



$$\nu_{1\to a}(x_1) = 1$$

$$\nu_{3\to b}(x_3) = 1$$

$$\nu_{5\to d}(x_5) = 1$$

$$\mu_{a\to 2}(x_2) = \sum_{x_1} f_a(x_1,x_2)\cdot \nu_{1\to a} = \sum_{x_1} f_a(x_1,x_2)$$

$$\mu_{b\to 2}(x_2) = \sum_{x_3} f_b(x_2,x_3)\ \nu_{3\to b} = \sum_{x_3} f_b(x_2,x_3)$$

$$\mu_{d\to 4}(x_4) = \sum_{x_5} f_d(x_4,x_5)\ \nu_{5\to b} = \sum_{x_5} f_d(x_4,x_5)$$

$$\nu_{2\to c}(x_2) = \mu_{a\to 2}\cdot \mu_{b\to 2} = \left(\sum_{x_1} f_a(x_1,x_2)\right)\left(\sum_{x_3} f_b(x_2,x_3)\right)$$

$$\mu_{c\to 4}(x_4) = \sum_{x_2} f_c(x_2,x_4)\ (\nu_{2\to c})$$

$$= \sum_{x_2} f_c(x_2,x_4)\left(\sum_{x_1} f_a(x_1,x_2)\ \sum_{x_3} f_b(x_2,x_3)\right)$$

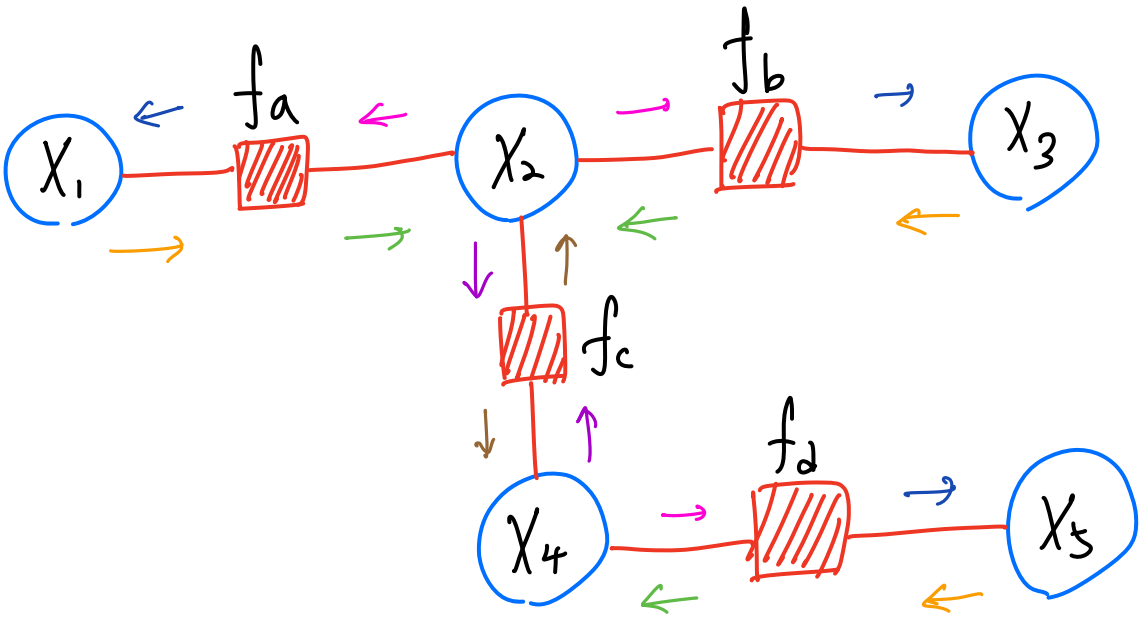
$$\therefore p(x_4) \propto \mu_{d\to 4}\cdot \mu_{c\to 4}$$

$$= \frac{1}{Z}\underbrace{\sum_{x_5} f_d(x_4,x_5)}_{\mu_{d\to 4}}\ \underbrace{\sum_{x_2} f_c(x_2,x_4)\left(\sum_{x_1} f_a(x_1,x_2)\ \sum_{x_3} f_b(x_2,x_3)\right)}_{\mu_{c\to 4}}$$

(c) $f_u(x,y) = x+y$ for every $u \in \{a,b,c,d\}$

max-sum BP algorithm → (i) optimal configuration X^* maximizing the joint distribution

(ii) maximum of the joint distribution.



$$V_{1\to a}(x_1) = 0$$

$$V_{3\to b}(x_3) = 0$$

$$V_{5\to d}(x_5) = 0$$

$$\mu_{a\to 2}(x_2) = \max_{x_1} \{\log f_a(x_1,x_2) + V_{1\to a}\}$$

$$= \max_{x_1} \{\log(x_1+x_2)\} = \log(1+x_2)$$

$$\mu_{b\to 2}(x_2) = \max_{x_3}\{\log f_b(x_2,x_3) + V_{3\to b}\}$$

$$= \max_{x_3}\{\log(x_2+x_3)\} = \log(x_2+1)$$

$$\mu_{d\to 4}(x_4) = \max_{x_5}\{\log f_d(x_4,x_5) + V_{5\to d}\}$$

$$= \max_{x_5}\{\log(x_4+x_5)\} = \log(x_4+1)$$

$$V_{2\to c}(x_2) = \mu_{a\to 2} + \mu_{b\to 2}$$

$$= \log(1+x_2) + \log(x_2+1)$$

$$= \log(1+x_2)(x_2+1)$$

$$V_{4\to c}(x_4) = \mu_{d\to 4}$$

$$= \log(x_4+1)$$

$$\mu_{c\to 4}(x_4) = \max_{x_2}\{\log f_c(x_2,x_4) + V_{2\to c}(x_2)\}$$

$$= \max_{x_2}\{\log(x_2+x_4) + \log(1+x_2)(x_2+1)\}$$

$$= \max_{x_2}\{\log(x_2+x_4)(1+x_2)(x_2+1)\}$$

$$= \log(1+x_4)(1+1)(1+1)$$

$$\mu_{c\to 2}(x_2) = \max_{x_4}\{\log f_c(x_2,x_4) + V_{4\to c}(x_4)\}$$

$$= \max_{x_4}\{\log(x_2+x_4) + \log(x_4+1)\}$$

$$= \max_{x_4}\{\log(x_2+x_4)(x_4+1)\}$$

$$= \log(x_2+1)(1+1)$$

$$\nu_{4\to d}(x_4) = \mu_{c\to 4}(x_4)$$

$$= \log(1+x_4)(1+1)(1+1)$$

$$\nu_{2\to a}(x_2) = \mu_{c\to 2}(x_2)$$

$$= \log(x_2+1)(1+1)$$

$$\nu_{2\to b}(x_2) = \mu_{c\to 2}(x_2)$$

$$= \log(x_2+1)(1+1)$$

$$\mu_{d\to 5}(x_5) = \max_{x_4}\{\log f_d(x_4,x_5) + \nu_{4\to d}\}$$

$$= \max_{x_4}\{\log(x_4+x_5) + \log(1+x_4)(1+1)(1+1)\}$$

$$= \log(1+x_5)(1+1))(1+1)(1+1)$$

$$\mu_{a\to 1}(x_1) = \max_{x_2}\{\log f_a(x_1,x_2) + \nu_{2\to a}\}$$

$$= \max_{x_2}\{\log(x_1+x_2) + \log(x_2+1)(1+1)\}$$

$$= \log(x_1+1)(1+1)(1+1)$$

$$\mu_{b\to 3}(x_3) = \max_{x_2}\{\log f_b(x_2,x_3) + \nu_{2\to b}\}$$

$$= \max_{x_2}\{\log(x_2+x_3) + \log(x_2+1)(1+1)\}$$

$$= \log(1+x_3)(1+1)(1+1)$$

$$\log p(X_4) = \max_{X_4} \left(\mu_{c \to 4} + \mu_{d \to 4} \right)$$

$$= \max_{X_4} \left(\log (1 + X_4)(1+1)(1+1) + \log(X_4 + 1) \right)$$

$$= \log 2^4 = \log 16$$

And, we need normalization term $\left(\frac{1}{Z}\right)$. then

$$P_{max} = \frac{16}{Z}, \quad X^* = (X_1^*, X_2^*, X_3^*, X_4^*, X_5^*) = (1, 1, 1, 1, 1)$$

3 [Gaussian Mixture Model & Expectation - Maximization]

Gaussian Mixture Model (GMM) with K components.

each component $k \in [K] := \{1, 2, \dots, K\}$

mean μ_k, variance σ_k^2, mixing parameter $\pi_k \rightarrow \theta = (\mu_k, \sigma_k, \pi_k)_{k \in [K]}$

dataset $X = \{x_i\}_{i \in [n]}$, $x_i \in \mathbb{R}$

latent variables $Z = \{z_i\}_{i \in [n]}$, $z_i = k$

$$\mathcal{N}(x_i \mid \mu_k, \sigma_k) := \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - x_k)^2}{2\sigma_k^2}\right)$$

(a)

① log-likelihood of the data

$$\log p(X \mid \theta) = \log \prod_{i=1}^{n} p(x_i \mid \theta) = \sum_{i=1}^{n} \log p(x_i \mid \theta)$$

$$= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)$$

② complete-data log-likelihood

$$\log p(X, Z \mid \theta) = \sum_{i \in [n]} \log p(x_i, z_i \mid \theta) = \sum_{i \in [n]} \log \left(p(x_i \mid z_i, \theta)\, p(z_i \mid \theta) \right)$$

$$= \sum_{i \in [n]} \log \left(\prod_{k \in [K]} \left(\pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2) \right)^{z_{ki}} \right)$$

$$= \sum_{i \in [n]} \sum_{k \in [K]} z_{ki} \log \left(\pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2) \right)$$

(b) $\theta^{(t)} = (\mu_k^{(t)}, \sigma_k^{(t)}, \pi_k^{(t)})_{k \in [K]}$

posterior probability

$$p(z_i = k \mid x_i, \theta^{(t)})$$

$$= \mathbb{E}_{z_i \mid x_i, \theta^{(t)}} [z_{ik}]$$

$$= \frac{\pi_k p(x_i \mid z_i = k, \mu_k^{(t)}, \sigma_k^{(t)2})}{\sum_{l \in [K]} \pi_k p(x_i \mid z_i = l, \mu_l^{(t)}, \sigma_l^{(t)2})}$$

$$= r_{ik} = z_{ik}^{(t)}$$

(c)

$$\mathbb{E}_{z_i \mid x_i, \theta^{(t)}} [\log p(x_i, z_i \mid \theta)]$$

$$= \sum_{k \in [K]} \mathbb{E}_{z_i \mid x_i, \theta^{(t)}} [z_{ki} \log (\pi_k N(x_i \mid \mu_k^{(t)}, \sigma_k^{(t)2})]$$

$$= \sum_{k \in [K]} r_{ik} \log (\pi_k N(x_i \mid \mu_k^{(t)}, \sigma_k^{(t)2})]$$

$$= \sum_{k \in [K]} z_{ik}^{(t)} \log (\pi_k N(x_i \mid \mu_k^{(t)}, \sigma_k^{(t)2})]$$

(d) M-step to compute $\theta^{(t+1)}$ from $\theta^{(t)}$ and $z_{ik}^{(t)}$'s

$$\theta^{(t+1)} = \left(\mu_k^{(t+1)}, \sigma_k^{(t+1)}, \pi_k^{(t+1)}\right)_{k\in[K]} = \arg\max_{\theta} \sum_{i\in[n]} \mathbb{E}_{z_i|x_i,\theta^{(t)}} \left[\log p(x_i, z_i|\theta)\right]$$

$$= \arg\max_{\theta} Q(\theta; \theta^{(t)})$$

① mean

$$\frac{\partial Q}{\partial \mu_k} = -\frac{1}{\sigma_k^{(t)^2}} \sum_{i\in[n]} z_{ik}^{(t)} (x_i - \mu_k^{(t)}) = 0$$

$$\Rightarrow \mu_k^{(t+1)} = \frac{\sum_{i\in[n]} z_{ik}^{(t)} x_i}{\sum_{i\in[n]} z_{ik}^{(t)}}$$

② variance

$$\frac{\partial Q}{\partial \sigma_k^2} = \sum_{i\in[n]} z_{ik}^{(t)} \left[-\frac{D}{\sigma_k^{(t)}} + \frac{1}{\sigma_k^{(t)^3}} \|x_i - \mu_k^{(t)}\|^2\right] = 0$$

$$\Rightarrow \sigma_k^{2(t+1)} = \frac{1}{D} \frac{\sum_{i\in[n]} z_{ik}^{(t)} \|x_i - \mu_k^{(t+1)}\|^2}{\sum_{i\in[n]} z_{ik}^{(t)}}$$

③ mixing parameter

$\{\pi_k\}$ must verify $\sum_{k\in[K]} \pi_k = 1$. Hence, reading the theory of constrained optimization, consider the Lagrangian

$$Q'(\theta, \lambda) = Q(\theta) + \lambda\left(1 - \sum_{k\in[K]} \pi_k\right)$$

$$\frac{\partial Q'(\theta,\lambda)}{\partial \pi_k} = \sum_{i\in[n]} \frac{z_{ik}^{(t)}}{\pi_k^{(t)}} - \lambda = 0$$

$$\Rightarrow \pi_k^{(t+1)} = \frac{1}{n} \sum_{i\in[n]} z_{ik}^{(t)}$$

(e) For every cluster $k \in [K]$, when we choose

$\pi_k = \frac{1}{K}$, $\sigma_k^2 = \sigma^2$ and σ^2 goes to 0,

then EM algorithm for GMM will be same with K-means.