

机器学习

第 8 章概率图模型-Probabilistic Graphical Models

欧阳毅

浙江工商大学
管理工程与电子商务学院

2023

基础问题

- 表示 (Representation)
 - 如何获取模型的不确定度
 - 如何对我们的领域知识进行编码
- 学习 (Learning)
 - 怎样的模型对于我们的数据是有效的?

$$M = \arg \max_{M \in \mathcal{M}} F(D; M)$$

- 推断 (Inference)
 - 在给定一定的数据后，如何根据已有的知识对问题进行解答

$$P(X_i | D; M)$$

- 如何对我们的领域知识进行编码

基础问题

- 表示 (Representation)

- 什么是联合概率分布?

$$P(X_1, X_2, X_3, X_4, X_5, X_6)$$

- 若 X_i 是 0-1 取值的离散随机变量, 他们有多少种状态?
- 它们是否都需要被表示?

- 学习 (Learning)

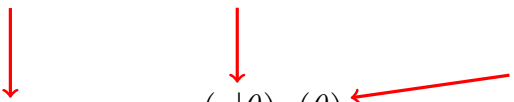
- 我们如何获取这些概率?
- 根据这些概率值和变量之间的关系, 我们如何建立领域知识

- 推断 (Inference)

- 若有不可见观察变量, 如何计算隐变量的条件概率分布?

贝叶斯理论 I

后验概率 或然率 Likelihood 先验概率 Prior


$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\sum_{\theta \in \Theta} p(x|\theta)p(\theta)}$$

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

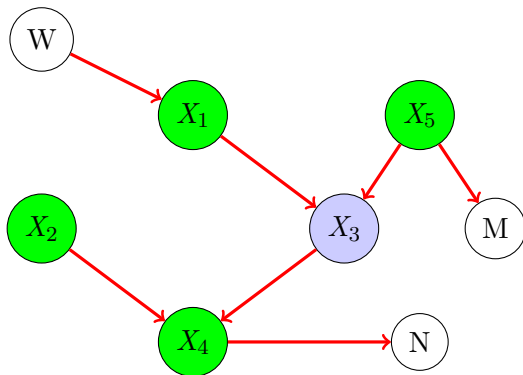
两类图模型 I

GM = Multivariate Statistics + Structure

- 有向图给出了因果关系（如：贝叶斯网络）
- 无向图给出了随机变量之间的联系（如：MRF）

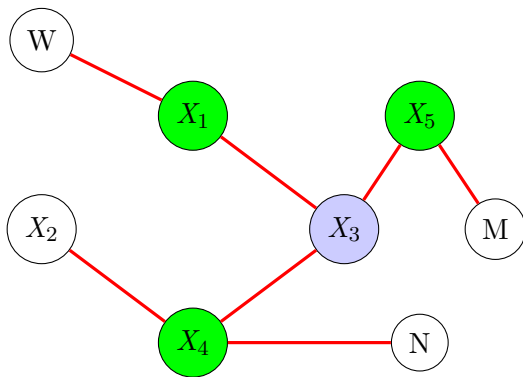
有向图 I

- 一个节点相对于马尔可夫毯 (Markov blanket) 之外的节点是条件独立
 - Parent
 - Child
 - Children's Co-parent



无向图 I

- 一个节点相对于直接相邻之外的节点是条件独立
- 通常使用势能函数 (potential) 和团 (cliques) 对联合概率进行建模



例子：给出无向图 Markov blanket

```
model = MarkovModel(  
    ("x", "y"),  
    ("z", "y"),  
    ("y", "w"),  
    ("y", "v"),  
    ("u", "w"),  
    ("s", "v"),  
    ("w", "t"),  
    ("w", "m"),  
    ("v", "n"),  
    ("v", "q"),  
)  
a1=model.markov_blanket('y')  
print(list(a1))
```


贝叶斯网络 I

- BN 是一种有向图结构，它的节点表示随机变量，边表示一个变量对另一个的影响。
- BN 提供了一种对联合概率进行因式分解的计算方法
- 它利用一组条件独立假设提供了一种完备的分布表示

因式分解定理 I

- 给定一个有向无环图 G ，概率分布的形式与 G 中给定节点双亲的因子保持一致

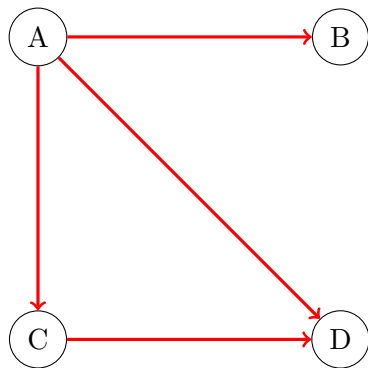
$$P(X) = \prod_{i=1, n} P(X_i | X_{pi})$$

- 也就是说联合概率可写成如下形式：

$$P(x_1, \dots, x_m) = \prod_{i=1}^m p(x_i | x_{pi})$$

我们说概率分布 P 可关于图 G 因式分解

因式分解定理 I



G1

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

我们说概率分布 P 可以关于 $G1$ 因式分解

表示 Representation

定义 (无向图模型)

一个无向图模型由一个定义在无向图 H 上的概率分布 $P(X_1, \dots, X_n)$, 和一组与 H 中团 (maximal clique) 相连的正势能函数 (positive potential functions) Ψ_c 表示。

s.t.

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in C} \Psi_c(X_c) \quad (\text{A Gibbs distribution})$$

其中 Z 为划分函数:

$$Z = \sum_{X_1, \dots, X_n} \prod_{c \in C} \Psi_c(X_c)$$

势能函数可以理解作为一种连续函数, 它的参数是与随机变量在图结构中关联的反映

目录

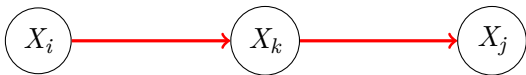
- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

贝叶斯网络模型 I

在一个贝叶斯网络中，任意一条由三个变量构成的图 X_i, X_k, X_j ，可能存在于下面三种连接方式：

- 1) 串行连接 (serial connection) 或链 (chain)，如图所示。
根据公式，图相应的联合分布为

$$P(X_i, X_k, X_j) = P(X_i)P(X_k|X_i)P(X_j|X_k) \quad (1)$$



贝叶斯网络模型 II

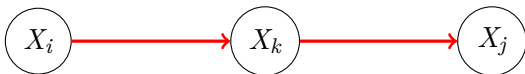
根据“条件局部独立性”可以得出这样一个结论：给定节点 k 时，节点 j 和其非后代节点 i 关于节点 j 的父节点 k 条件独立。证明：

$$P(X_i, X_j | X_k) = \frac{P(X_i, X_j, X_k)}{P(X_k)} \quad (2)$$

$$= \frac{P(X_i)P(X_k | X_i)P(X_j | X_k)}{P(X_k)} \quad (3)$$

$$= \frac{P(X_i, X_k)P(X_j | X_k)}{P(X_k)} \quad (4)$$

$$= P(X_i | X_k)P(X_j | X_k) \quad (5)$$

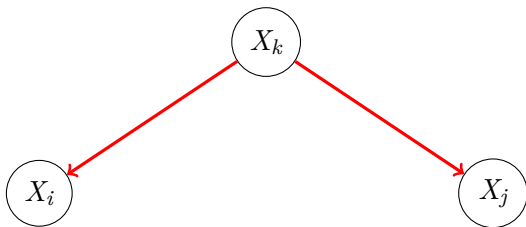


即：在给定 X_k 的条件下， X_i, X_j 被阻断是独立的:head-to-tail

贝叶斯网络模型 III

- 2) 发散连接 (diverging connection) 或叉口 (fork), 表示 X_i 和 X_j 有共同的原因, 相应的联合分布为

$$P(X_i, X_k, X_j) = P(X_k)P(X_i|X_k)P(X_j|X_k) \quad (6)$$



贝叶斯网络模型 IV

给定节点 k 时, 节点 j 和其非后代节点 i 关于节点 i 的父节点 k 条件独立。

$$P(X_i, X_j | X_k) = \frac{P(X_i, X_j, X_k)}{P(X_k)} \quad (7)$$

$$= \frac{P(X_k)P(X_i | X_k)P(X_j | X_k)}{P(X_k)} \quad (8)$$

$$= \frac{P(X_i, X_k)P(X_j | X_k)}{P(X_k)} \quad (9)$$

$$= P(X_i | X_k)P(X_j | X_k) \quad (10)$$

这说明, 在发散连接的情况下, $X_i \perp X_j | X_k$ 。

即: 在给定 X_k 的条件下, X_i, X_j 被阻断是独立的: tail-to-tail

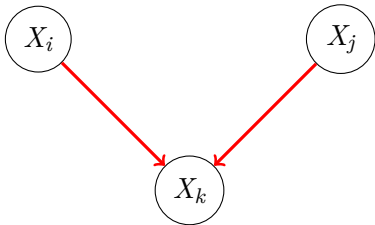
贝叶斯网络模型 V

- 3) 收敛连接 (v-structure) , 相应的联合分布为

$$P(X_i, X_k, X_j) = P(X_i)P(X_j)P(X_k|X_i, X_j) \quad (11)$$

节点 i 和节点 j 是先验独立的

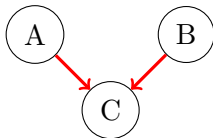
$$P(X_i, X_j) = P(X_i)P(X_j) \quad (12)$$



但给定 X_k 后, X_i, X_j 并不独立。在 X_k 未知的条件下, X_i, X_j 被阻断是独立的

V-structure

添加条件 C 可能损失独立性



A	B	C	Prob
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	1	0.25

- 这个图也可以用来解释上面这句话。在图中 A 和 B 相互独立，如果给定 C 的值，那么 A 和 B 之间不独立。
- 给定 $C=1$, $P(A=1|C)=2/3$
- 但是 $P(A=1|C)$ 的大小和 B 的取值有关，
如果 B 取 0，那么 $P(A=1|C)=1$ ，
如果 B 取 1， $P(A=1|C)=1/2$ ，此时 A 和 B 不独立。

目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

I-maps

定义 ($I(P)$)

令 P 是在随机变量 X 上的分布. 我们定义独立性集合 $I(P)$ 为具有 $(X \perp Y | Z)$ 形式的独立断言集合.

定义 (I-maps)

图中任意的对象 K 同一组与独立性集合 $I(K)$ 相联系. 若 $I(K) \subseteq I$ 我们说 K 是一个独立集 I 的 I-map.

若 G 是独立集 $I(P)$ 的一个 I-map, 我们可以说 G 是 P 的一个 I-map

关于 I-map 的性质

- G 是 P 的一个 I-map, 是一个 G 不会误导对 P 独立性分析的必要条件
- 任何 G 的断言必在 P 中保持, 相反不成立, P 可能会有另外的独立性, 其并未在 G 中反映。

Local Markov assumptions

定义 (G)

一个贝叶斯网络 (BN) 结构 G 是一个有向无环图 (DAG), 它的节点表示随机变量 X_1, \dots, X_n .

定义 (local conditional independence assumptions $I_l(G)$)

令 Pa_{X_i} 表示 X 在 G 中的双亲节点集合, $ND(X_i)$ 表示图中非 X_i 子孙的集合, 用 $I_l(G)$ 表示局部条件独立性假设集合

$$I_l(G) \triangleq \{X_i \perp ND(X_i) | Pa_{X_i} : \forall i\}$$

换句话说, 在给定 X_i 其双亲的前提, 节点 X_i 与其非子孙节点下独立

Active trail 有效迹

- Causal trail $X \rightarrow Z \rightarrow Y$
:active if and only if Z is not observed.
- Evidential trail $X \leftarrow Z \leftarrow Y$
: active if and only if Z is not observed.
- Common cause $X \leftarrow Z \rightarrow Y$
:active if and only if Z is not observed.
- Common effect $X \rightarrow Z \leftarrow Y$
:active if and only if either Z or one of Z 's descendants is observed

Active trail 有效迹

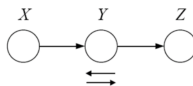
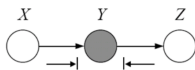
表: 给定证据 Z 的前提, 什么情况 X 和 Y 存在相互影响?

情况 i	$W \notin Z$	$W \in Z$
$X \rightarrow Y$	✓	✓
$Y \rightarrow X$	✓	✓
$X \rightarrow W \rightarrow Y$	✓	×
$X \leftarrow W \leftarrow Y$	✓	×
$X \leftarrow W \rightarrow Y$	✓	×
$X \rightarrow W \leftarrow Y$	×	✓

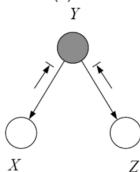
× 处表示 X, Y 是独立的

✓ 处表示 X, Y 不是独立的

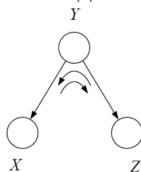
从信息流动的角度对三种结构进行分析



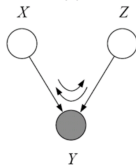
(a)



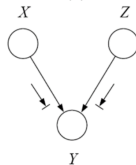
(b)



(a)



(b)



图分割准则

贝叶斯网络的 D-separation 图分割准则 (D 是指有向图)

定义 (D-separation)

若在 G 中, 给定 Z 条件下, X 和 Y 间不存在有效迹, X 和 Y 在给定 Z 下是 D-separated (条件独立)

定义 (G)

$I(G)$ = 包含了存在于 D-separation 中所有的独立性断言

$$I(G) = \{(X \perp Y | Z) : dsep_G(X, Y | Z)\}$$

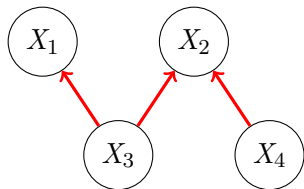
定义 (局部条件独立性假设 $I_l(G)$)

令 Pa_{X_i} 表示 X 在 G 中的双亲节点集合, $ND(X_i)$ 表示图中非 X_i 子孙的集合, 用 $I_l(G)$ 表示局部条件独立性假设集合

$$I_l(G) \triangleq \{X_i \perp ND(X_i) | Pa_{X_i} : \forall i\}$$

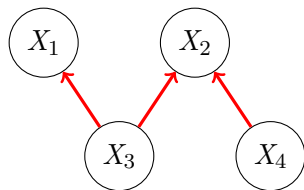
换句话说, 在给定 X_i 其双亲的前提, 节点 X_i 与其非子孙节点下独立

BN 练习



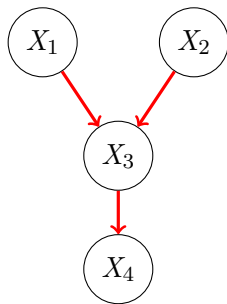
- 给出上图的 $I(G)$ 表示
- $I(G) = \{X_3 \perp X_4, X_1 \perp X_2 | X_3, X_1 \perp X_4 | X_3\}$

BN 练习



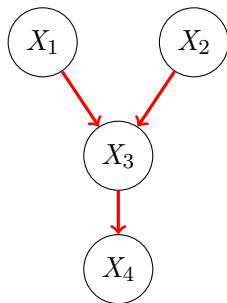
- 给出上图的 $I(G)$ 表示
- $I(G) = \{X_3 \perp X_4, X_1 \perp X_2 | X_3, X_1 \perp X_4 | X_3\}$

BN 练习



- 给出上图的 $I(G)$ 表示
- $I(G) = \{X_1 \perp X_2, X_1 \perp X_4 | X_3, X_2 \perp X_4 | X_3\}$

BN 练习



- 给出上图的 $I(G)$ 表示
- $I(G) = \{X_1 \perp X_2, X_1 \perp X_4 | X_3, X_2 \perp X_4 | X_3\}$

目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

等价定理

定义

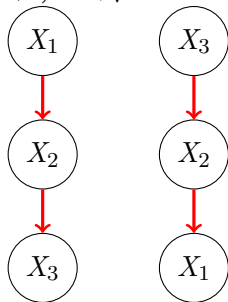
给定图 G ，令 D_1 表示满足 $I(G)$ 的所有分布族， D_2 表示根据 G 因式分解的所有分布族，

$$P(X) = \prod_{i=1:d} P(X_i | X_{\pi_i})$$

则有： $D_1 \equiv D_2$

等价定理

不同的 BN 图可能是等价的，只要他们具有相同的条件独立性断言，如图：



$$X_1 \perp X_3 | X_2$$

I-等价

定义 (I-equivalence)

在随机变量 X 上的, 两个 BN 图 G_1, G_2 , 若 $I(G_1) = I(G_2)$, 则称为 I-等价

这提供了对 P 进行因式分解的途径

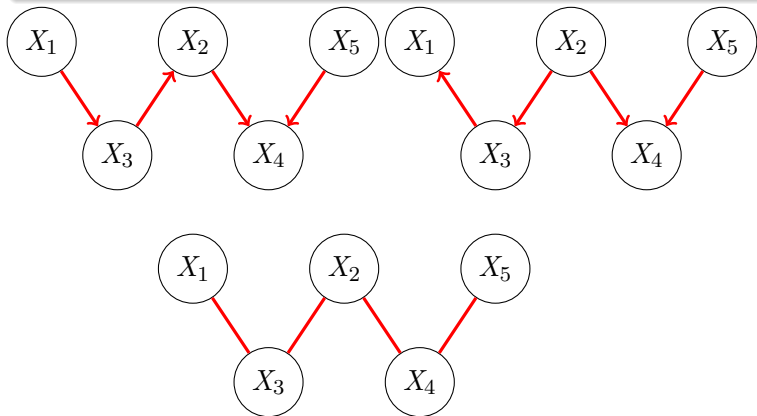
定义 (skeleton)

BN 图 G 的骨架为定义在 V 集上的一个无向图, 它包含了原 G 图中 $\langle X, Y \rangle$ 的边作为 (X, Y)

Detecting I-等价

定理

令 $G1$ 和 $G2$ 为两个在 V 上的图。若 $G1$ 和 $G2$ 有相同的骨架，并且 v -结构相同，则它们是 I -等价



极小 I-map

定义 (极小 I-map)

一个有向无环图 G 对于分布 P 是一个极小 I-map , 首先它是 P 的 I-map 映射, 并且若删除任意一个 G 中的边, 都不再是 I-map.

一个分布可能存在多个极小 I-maps

P-maps

定义 (P-maps)

在一个有向无环图 G 和分布 P 中, 若 $I(P) = I(G)$, 则 G 对于分布 P 是一个 P-maps 映射 (perfect map),

定理

并不是每个分布都有 P -映射

证明.

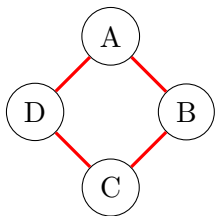
给个反例, 若模型具有有以下独立性:

$$A \perp C | \{B, D\}, \text{ and } B \perp D | \{A, C\}$$

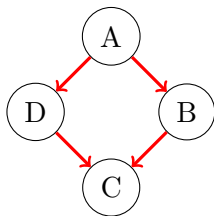
我们找不出贝叶斯网络的表现形式



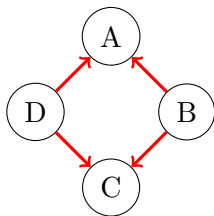
例子



Markov Network



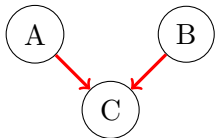
BN1



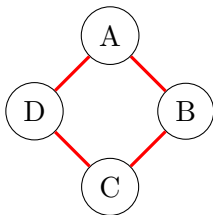
BN2

- MN: $\{A \perp C | \{B, D\}, B \perp D | \{A, C\}\}$
- BN1 中: $A \perp C | \{B, D\}, B \perp D | A$
- BN2 中: $A \perp C | \{B, D\}, B \perp D$
- This MN does not have a P-map as BN

例子



G1

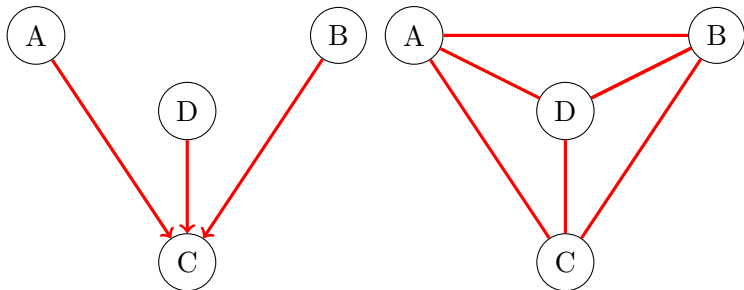


G2

- 若一个分布的条件独立性断言为: $\{A \perp B, \neg(A \perp B | C)\}$ 那么 G1 就是该分布的 P-map, 对于该分布, 没有无向图作为 P-map
- G2 中的条件独立性断言为: $\neg(A \perp B), C \perp D | \{A, B\}, A \perp B | \{C, D\}$, 没有有向图作为 P-map

有向图转换为无向图

- Moralization 方法，保证添加了最少的连接。
- 1. 若一个节点只有一个 parent，那么 parent 到它的有向箭头可去掉
- 2. 若一个节点有多个双亲节点，则要把每个直接双亲组两两连接



G2

目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

无向图模型

- 成对关系（非因果）
- 可构建模型，对图中结构进行评价，但不能显式产生样本

无向图模型

定义 (无向图模型)

一个无向图模型由一个定义在无向图 H 上的概率分布 $P(X_1, \dots, X_n)$, 和一组与 H 中团 (maximal clique) 相连的正势能函数 (positive potential functions) Ψ_c 表示。

s.t.

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in C} \Psi_c(X_c) \quad (A \text{ Gibbs distribution})$$

其中 Z 为划分函数:

$$Z = \sum_{X_1, \dots, X_n} \prod_{c \in C} \Psi_c(X_c)$$

势能函数可以理解作为一种连续函数, 它的参数是与随机变量在图结构中关联的反映

目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

Global Markov Independencies

- 令 H 为一个无向图

定义

$sep_H(A; C|B)$: 表示 B 分隔 A 和 C , 即从节点 A 到节点 C 的路径都需经过 B

- 一个概率分布满足全局马尔可夫属性指:
若任意的不相交集 A, B, C , B 分隔 A 和 C , 在给定 B 的条件下, A 和 C 相互独立, 即:

$$I(H) = \{A \perp C | B : sep_H(A; C|B)\}$$

Local Markov independencies I

- 对于每个节点 $X_i \in V$, 有唯一的马尔可夫毯 (Markov blanket), 记为 MB_{X_i} , 它由与 X_i 相邻的一组节点构成。(无向图有共享边存在)

定义

局部马尔可夫独立性是指:

$$I_L(H) \triangleq \{X_i \perp V - \{X_i\} - MB_{X_i} | MB_{X_i} : \forall i\}$$

目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

量化分析-Cliques 势团 I

定义

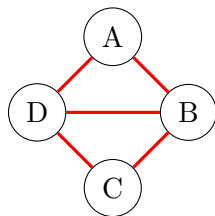
对于图 $G = \{V, E\}$, 一个完全子图 (clique) 是一个节点全连接
的子图即: $G' = \{V' \subseteq V, E' \subseteq E\}$

定义

最大势团 (maximal clique) 是一个完全子图, 任何节点超集合都不在是完全图。

即: 加个节点就不是完全图了。

量化分析-Cliques 势团 I



子势团 (sub-cliques) 不必是最大势图

- $\text{max-cliques} = \{A, B, C\}, \{B, C, D\}$
- $\text{sub-cliques} = \{A, B\}, \{C, D\} \dots$
- 注意: A, C 之间没有直接边

Gibbs Distribution and Clique Potential I

定义

一个无向图模型由一个定义在无向图 H 上的概率分布 $P(X_1, \dots, X_n)$, 和一组与 H 中团 (maximal clique) 相连的正势能函数 (positive potential functions) Ψ_c 表示。

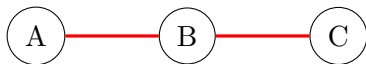
s.t.

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in C} \Psi_c(X_c) \quad (\text{A Gibbs distribution})$$

where Z is known as the partition function:

$$Z = \sum_{X_1, \dots, X_n} \prod_{c \in C} \Psi_c(X_c)$$

Clique Potential I



- 此模型蕴含 $A \perp C | B$. 则联合概率可因式分解为:

$$p(A, B, C) = p(B)p(A|B)p(C|B)$$

- 也可以写成:

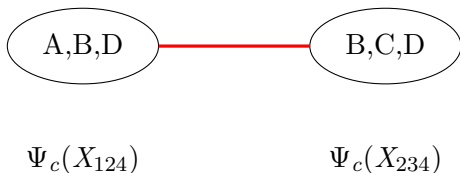
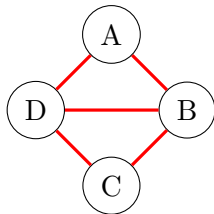
$$p(A, B, C) = p(A, B)(C|B)$$

或

$$p(A, B, C) = p(B, C)p(A|B)$$

- 但不是所有势能函数都能进行边缘化
- 不能让所有的势能函数都是条件形式

Clique Potential-使用最大势团

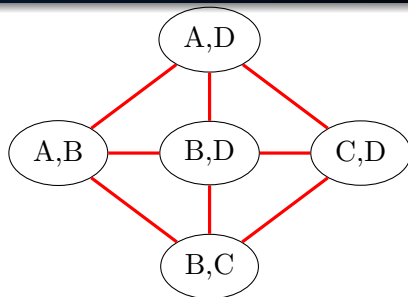
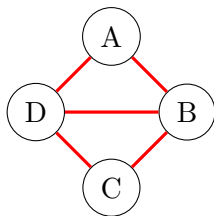


$$P(X_1, X_2, X_3, X_4) = \frac{1}{Z} \Psi_c(X_{124}) \times \Psi_c(X_{234})$$

$$Z = \sum_{X_1, X_2, X_3, X_4} \Psi_c(X_{124}) \times \Psi_c(X_{234})$$

- 对于离散节点，我们可以表示 $P(X_{1:4})$ 为 2 个三维表，而不是 1 个四维表

Clique Potential-使用 subcliques



$$P(X_1, X_2, X_3, X_4) = \frac{1}{Z} \prod_{ij} \Psi_{ij}(X_{ij})$$

$$= \frac{1}{Z} \Psi_{12}(X_{12}) \Psi_{23}(X_{23}) \Psi_{24}(X_{24}) \Psi_{34}(X_{34})$$

- 对于离散节点，我们可以表示 $P(X_{1:4})$ 为 5 个二维表，而不是 1 个四维表

Hammersley-Clifford Theorem I

定理 (Hammersley-Clifford Theorem)

A strictly positive distribution $p(x)$ (i.e. $p(x) > 0$ for all x) satisfies the global Markov property (G) with respect to $G(V, E)$ if and only if it can be factorized according to G :

$$p(x) = \frac{1}{Z} \prod_{c \in C(G)} \psi_c(x_c)$$

若一个分布是严格正的且满足全局马尔可夫属性，当且仅当它对应的图模型可被因式分解为上式。

P-map 映射 (Perfect maps)

定理 (P-map 映射 (Perfect maps))

一个 Markov 网络 H 是一个分布 P 的 P -map (perfect map), 则对任意的 $X; Y; Z$ 我们有:

$$sep_H(X, Z | Y) \Leftrightarrow P \models (X \perp Z | Y)$$

定理 (Perfect maps)

在无向图模型中, 并不是每个分布都存在 P -map 映射

证明.

举个反例: 没有一个无向图可以描述所有存在 v-structure 中的独立性断言. $X \rightarrow Z \leftarrow Y$. □

指数形式

定义

我们采用一种非约束形式表示团势能 $\Psi_c(X_c)$ ，使用一个实数函数 $\phi_c(X_c)$ 表示能量函数

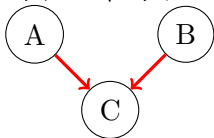
$$\Psi_c(X_c) = \exp(-\phi_c(X_c))$$

- 这样的定义使得联合概率函数可写为一种累加的结构

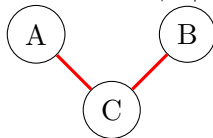
$$P(X) = \frac{1}{Z} \exp\left\{-\sum_{c \in C} \phi_c(X_c)\right\} = \frac{1}{Z} \exp\{-H(X)\}$$

- $H(X)$ 被称为自由能量

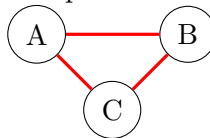
问题：存在一个马尔可夫网络是给定贝叶斯网络的 P-map 吗？



$$A \perp B, \neg(A \perp B | C)$$



$$A \perp B | C, \neg(A \perp B)$$



$$\neg(A \perp B | C), \neg(A \perp B)$$

- V-structure 在马尔可夫网络中并没有对应的结构

概率推断和学习

- 一个 GM 模型描述了一个唯一的概率分布 P
- 典型任务
 - 1. 如何回答如 $P_M(X|Y)$ 的查询
我们用推断描述这一类查询过程
 - 2. 从数据 D 中如何估计可能的模型 M
我们用学习描述这一类估计 M 的处理过程
并不是所有变量都是可观察的，我们需要推断来计算缺失信息。

Likelihood

- 多数查询会涉及到证据 evidence
- Evidence e 是一个 E 集合在其定义域中的一种取值情况。
- 简单查询：计算证据概率

$$P(e) = \sum_{x_1} \dots \sum_{x_k} P(x_1, \dots, x_k, e)$$

- 这也被称为计算 e 的 likelihood

条件概率

- 我们通常对给定证据的条件下，变量的条件概率感兴趣

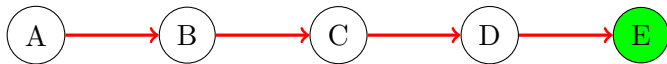
$$P(X|e) = \frac{P(X, e)}{P(e)} = \frac{P(X, e)}{\sum_x P(X = x, e)}$$

- 我们会查询 Y 的一个子集，而不关心其它随机变量 Z 的取值：

$$P(Y|e) = \sum_z P(Y, Z = z|e)$$

这个过程称为 z 的边缘化

链式消解



- 查询 $P(e)$

$$P(e) = \sum_d \sum_c \sum_b \sum_a P(a, b, c, d, e)$$

- 通过链式分解，我们有：

$$= \sum_d \sum_c \sum_b \sum_a P(a) P(b|a) P(c|b) P(d|c) P(e|d)$$

- 通过重新排列

$$= \sum_d \sum_c \sum_b P(c|b) P(d|c) P(e|d) \sum_a P(a) P(b|a)$$

目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

链式消解



- 通过重新排列, 消除了 A 变量

$$= \sum_d \sum_c \sum_b P(c|b)P(d|c)P(e|d)P(b)$$

- 通过重新排列, 消除了 B 变量

$$= \sum_d \sum_c P(d|c)P(e|d)P(c)$$

- 通过重新排列, 消除了 C 变量

$$= \sum_d P(e|d)P(d)$$

- 通过重新排列, 消除了 D 变量: $P(e)$
- 计算复杂度为 $O(kN^2)$, 而原来是 $O(N^k)$

目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

采用变量消解方法对 GM 进行推理

思路:

- 给出查询形式:

$$P(X_1, e) = \sum_{x_n} \dots \sum_{x_3} \sum_{x_2} \prod_i P(x_i | Pa_i)$$

- 这实际是给出了对隐变量进行消解的顺序
- 迭代进行以下步骤:
 - 将所有不相关的项移动到内部累加之外
 - 执行内部累加, 产生一个新项
 - 将新项插入到乘积项中
- wrap-up

$$P(X_1 | e) = \frac{\phi(X_1, e)}{\sum_{x_1} \phi(X_1, e)}$$

变量消解的输出

- 令 X 为某组随机变量
 F 是一因子集合, 其中 $\phi \in F$
 $Y \subset X$ 为查询变量, $T = X - Y$ 为被消解变量集合
- T 被消解后的结果是一个因子:

$$\tau(Y) = \sum_T \prod_{\phi \in F} \phi$$

这个因子不必对应于网络中的任何概率或条件概率

对于证据 (Evidence) 的处理

根据条件使用 Sum-Product 操作

- 证据的势能:

$$\delta(E_i, e_i) = \begin{cases} 1 & \text{if } E_i \equiv e_i \\ 0 & \text{if } E_i \neq e_i \end{cases}$$

- 总体证据的势能:

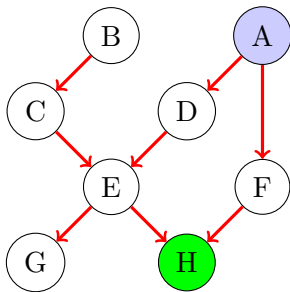
$$\delta(E, e) = \prod_{i \in I_E} \delta(E_i, e_i)$$

- 引入证据

$$\tau(E, e) = \sum_{T, e} \prod_{i \in I_E} \phi \times \delta(E, e)$$

变量消解-例子 I

- 查询: $P(A|h)$
 - 需要消解变量 B,C,D,E,F,G,H

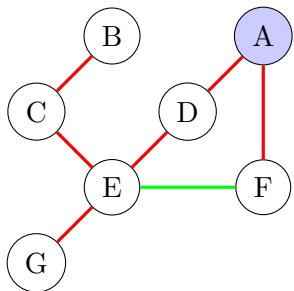


- 初始化因子:

$$P(A : H) = P(a)P(b)P(c|b)P(d|a)P(e|c, d)P(f|a)P(g|e)P(h|e, f)$$

- 选择消解次序:H,G,F,E,D,C,B

变量消解-例子 I



Step1 :Conditioning, 根据 h 的观察值 \hat{h} 固定证据节点 h

$$m_h(e, f) = p(h = \hat{h} | e, f)$$

$$P(A : H) = P(a)P(b)P(c|b)P(d|a)P(e|c, d)P(f|a)P(g|e)m_h(e, f)$$

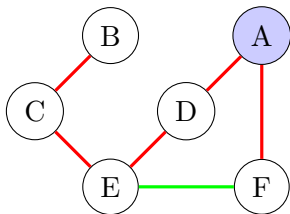
变量消解-例子 II

- 还需要消解:G,F,E,D,C,B

Step2 : 消除 G

$$m_g(e) = \sum_g P(g|e) = 1$$

$$P(A : H) = P(a)P(b)P(c|b)P(d|a)P(e|c, d)P(f|a)m_h(e, f)$$



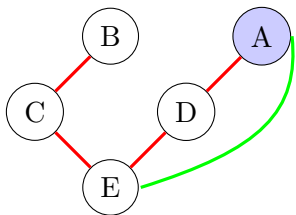
- 还需要消解:F,E,D,C,B

变量消解-例子 III

Step3 : 消除 F

$$m_f(e, a) = \sum_f P(f|a) m_h(e, f)$$

$$P(A : H) = P(a)P(b)P(c|b)P(d|a)P(e|c, d)m_f(a, e)$$



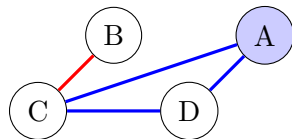
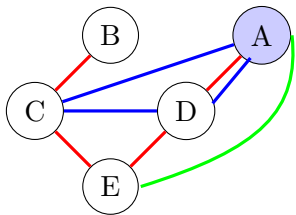
- 还需要消解:E,D,C,B

变量消解-例子 IV

Step4 : 消除 E

$$m_e(a, c, d) = \sum_e P(e|c, d) m_f(a, c)$$

$$P(A : H) = P(a)P(b)P(c|b)P(d|a)m_e(a, c, d)$$



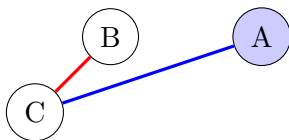
- 还需要消解:D,C,B

变量消解-例子 V

Step5 : 消除 D

$$m_d(a, c) = \sum_d P(d|a) m_e(a, c, d)$$

$$P(A : H) = P(a)P(b)P(c|b)m_d(a, c)$$



- 还需要消解:C,B

Step6 : 消除 C

$$m_c(a, b) = \sum_c P(c|b) m_d(a, c)$$

$$P(A : H) = P(a)P(b)m_c(a, b)$$



变量消解-例子 VI

- 还需要消解:B

Step7 : 消除 B

$$m_b(a) = \sum_c P(b) m_c(a, b)$$

$$P(A : H) = P(a) m_b(a)$$

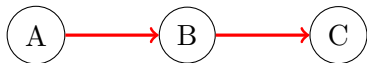


Step8 : Wrap-up

$$P(a, \hat{h}) = P(a) m_b(a), P(\hat{h}) = \sum_a P(a) m_b(a)$$

$$P(a|\hat{h}) = \frac{P(a) m_b(a)}{\sum_a P(a) m_b(a)}$$

变量消解-例子 I

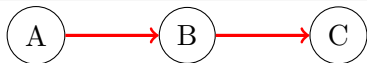


$$P(A, B, C) = P(A)P(B|A)P(C|B)$$

我们说概率分布 P 可以关于 $G1$ 因式分解

- A: 性别 (Male: 0.6; Female: 0.4)
- B: 身高
Male 大于 170cm=0.6 小于 170cm= 0.4
Female 大于 170cm=0.3 小于 170=0.7
- C: 体重
身高大于 170cm: 大于 80kg=0.65; 小于 80kg=0.35
身高小于 170cm: 大于 80kg=0.45; 小于 80kg=0.55

变量消解-例子 I



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

令：体重大于 80kg, $C=0$ ；体重小于 80kg, $C=1$

身高大于 170cm, $B=0$ ，身高小于 170cm, $B=1$

有位同学体重小于 80kg 预测身高？

$$\begin{aligned}
 P(B_1|C_1) &= \frac{\sum_A P(A, B=1, C=1)}{P(C=1)} \\
 &= \frac{\sum_A P(A, B=1, C=1)}{\sum_A \sum_B P(A, B, C=1)} \\
 &= \frac{0.6 * 0.4 * 0.55 + 0.4 * 0.7 * 0.55}{0.6 * 0.6 * 0.35 + 0.6 * 0.4 * 0.55 + 0.4 * 0.3 * 0.35 + 0.4 * 0.7 * 0.55} \\
 &= 0.63
 \end{aligned}$$

变量消解-例子 I

```
model = BayesianModel( [("A", "B"), ("B", "C")])
# add CPD to each edge
cpd_a = TabularCPD("A", 2, [[0.6], [0.4]])
cpd_b = TabularCPD(
    "B", 2,
    [[0.6, 0.3],
     [0.4, 0.7]],
    evidence=["A"],
    evidence_card=[2],)
cpd_c = TabularCPD(
    "C", 2,
    [[0.65, 0.45],
     [0.35, 0.55]],
    evidence=["B"],
    evidence_card=[2],)
model.add_cpds(cpd_a, cpd_b, cpd_c)
infer = VariableElimination(model)
print(infer.query(variables = ['B'], evidence={'C':1}))
```

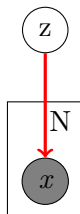
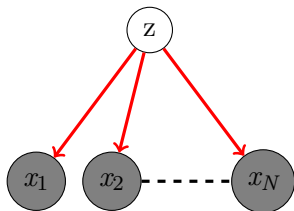
目录

- ① 引言
- ② 有向图模型分析
 - 模型结构
 - 性质分析
 - 等价定理
- ③ 无向图模型
 - 无向图模型
 - 独立性分析
 - 量化分析
- ④ GM 进行推理
 - 链式消解
 - 精确推断
 - 近似推断 *

变分推断 *

思路:

- 通过使用已知简单分布来逼近需推断的复杂分布，并通过限制近似分布的类型，从而得到一种局部最优，但具有确定解的近似后验分布。
- 概率图模型一种表示方法：盘式记法 (plate notation)
 - N 个变量 $\{x_1, \dots, x_N\}$ 均依赖于隐变量 z
 - 相互独立，由相同机制生成的多个变量被放在一个方框 (plate) 内，并在方框中标出个数 N
 - 方框可以嵌套，阴影表示已知（观测到）变量



变分推断 *

观察变量 x 的联合分布的概率密度函数:

$$p(x|\Theta) = \prod_{i=1}^N \sum_z p(x_i, z|\Theta)$$

$$\ln p(x|\Theta) = \sum_{i=1}^N \ln \left\{ \sum_z p(x_i, z|\Theta) \right\}$$

可使用 EM 算法求解:

- E Step: 根据 t 时刻的参数 Θ^t 对 $p(z|x, \Theta^t)$ 进行推断, 并计算联合似然函数 $p(x, z|\Theta^t)$. (已知 Θ^t , 求 $p(z|x, \Theta^t)$ 最大化)
- M Step: 基于 E 步的结果最大化寻找 Θ , 即对关于变量 Θ 的函数 $Q(\Theta; \Theta^t)$ 进行最大化

$$\Theta^{t+1} = \arg \max_{\Theta} Q(\Theta; \Theta^t) \quad (13)$$

$$= \arg \max_{\Theta} \sum_z p(z|x, \Theta^t) \ln p(x, z|\Theta) \quad (14)$$

变分推断 *

实际上 $Q(\Theta; \Theta^t)$ 是对 $\ln p(x, z|\Theta)$ 在分布 $p(z|x, \Theta^t)$ 下的期望, 当分布 $p(z|x, \Theta^t)$ 于变量 z 的真实后验分布相等时, $Q(\Theta; \Theta^t)$ 近似于对数似然函数。从而得到稳定的 Θ 参数。

$$\ln p(x) = L(q) + KL[q(z)||p(z|x)]$$

- $L(q) = \int q(z) \ln \frac{p(x, z)}{q(z)} dz$
- $KL[q(z)||p(z|x)] = - \int q(z) \ln \frac{p(z|x)}{q(z)} dz$
- $q(z) = \prod_{i=1}^M q_i(z_i)$ 指数族

$$KL[q||p] = \int q(x) \ln \frac{q(x)}{p(x)} dx = - \int q(x) \frac{p(x)}{q(x)}$$

$$\ln p(x) = \ln p(x, z) - \ln p(z|x)$$

$$\begin{aligned} KL[q(z)||p(z|x)] &= - \int dz (q(z) \ln \frac{p(x, z)}{q(z)} + \ln p(x)) \\ &= -L + \ln p(x) \end{aligned}$$

$$\begin{aligned}
 L(q) &= \int \prod_i q_i(z) \{ \ln p(x, z) - \sum_i \ln q_i(z) \} dz \\
 &= \int q_j(z_j) \{ \int \ln p(x, z) \prod_{i \neq j} q_i(z_i) dz_i \} dz_j - \int \prod_k q_k(z) \sum_i \ln q_i(z) dz
 \end{aligned}$$

$$\begin{aligned}
 \int \prod_k q_k(z) \sum_i \ln q_i(z) dz &= \sum_i \int \prod_k q_k(z) \ln q_i(z) dz \\
 &= \sum_i \int q_i(z_i) q_i(\hat{z}_i) \ln q_i(z_i) dz_i d\hat{z}_i \\
 &= \sum_i \int q_i(z_i) \ln q_i(z_i) dz_i
 \end{aligned}$$

- 求和可以拿到积分外面
- $z = \{z_i, \hat{z}_i\}$
- $\forall i, \int q_i(z_i) dz_i = 1$

$$L(q) = \int q_i(z_i) \left\{ \int \ln p(x, z) \prod_{j \neq i} q_j(z_j) dz_j \right\} dz_i - \sum_i \int q_i(z_i) \ln q_i(z_i) dz_i$$

$$\begin{aligned} \int q_i(z_i) \left\{ \int \ln p(x, z) \prod_{j \neq i} q_j(z_j) dz_j \right\} dz_i &= \int q_i(z_i) dz_i \int Q(\hat{z}_i) \ln p(x, z) d\hat{z}_i \\ &= \int q_i(z_i) \ln Q_i^*(z_i) dz_i + \ln Z \end{aligned}$$

- 令 $Q_i^*(z_i) = \frac{1}{Z} \exp \left\langle \int \ln p(x, z) q(\hat{z}) d\hat{z} \right\rangle$

$$\begin{aligned}
 L(q) &= \int q_i(z_i) \ln Q_i^*(z_i) dz_i + \ln Z - \sum_i \int q_i(z_i) \ln q_i(z_i) dz_i \\
 &= \left\{ \int q_i(z_i) \ln Q_i^*(z_i) dz_i - \int q_i(z_i) \ln q_i(z_i) dz_i \right\} + H[q(\hat{z}_i)] + \ln Z
 \end{aligned}$$

考虑花括号中部分

$$\begin{aligned}
 \int q_i(z_i) \ln Q_i^*(z_i) dz_i - \int q_i(z_i) \ln q_i(z_i) dz_i &= \int q_i(z_i) \ln \frac{Q_i^*(z_i)}{q_i(z_i)} dz_i \\
 &= -KL[q_i(z_i) || Q_i^*(z_i)]
 \end{aligned}$$

因此有：

$$L(q) = -KL[q_i(z_i) || Q_i^*(z_i)] + H[q(\hat{z}_i)] + \ln Z$$

我们希望最大化联合似然函数 $p(x, z | \Theta^t)$ ，就是最大化 $L(q)$ 。观察到 L 依赖于每个 q_i 仅通过 KL 项。

$$\frac{\partial L(q(z))}{\partial q_i(z_i)} = \frac{\partial -KL[q_i(z_i) || Q_i^*(z_i)] - \lambda_i (\int q_i(x_i) dx_i - 1)}{\partial q_i(z_i)} = 0$$

最大化 L, 实际上就是 KL 差异等于 0

$$q(z_i) = Q^*(z_i)$$

$Q_i^*(z_i) = \frac{1}{Z} \exp \left\langle \int \ln p(x, z) Q(\hat{z}) d\hat{z} \right\rangle$ 迭代更新公式:

$$q(z_i) \leftarrow \frac{1}{Z} \exp \left\langle \int \ln p(x, z) q(\hat{z}_i) d\hat{z}_i \right\rangle$$

Solution of $-KL[q_i(z_i)||Q_i^*(z_i)]$, Gaussian case I

令 J 为隐变量 z 的维度数, μ_j, σ_j 为均值和方差向量的第 j 个元素。

$$N(z; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \int q_\theta(z) \log p(z) dz &= \int N(z; \mu, \sigma^2) \log N(z; 0, I) dz \\ &= \int N(z; \mu, \sigma^2) \left(-\frac{1}{2}z^2 - \frac{1}{2}\log(2\pi)\right) dz \\ &= -\frac{1}{2} \int N(z; \mu, \sigma^2) z^2 dz - \frac{J}{2} \log(2\pi) \end{aligned}$$

Solution of $-KL[q_i(z_i)||Q_i^*(z_i)]$, Gaussian case I

$$\because \sigma(z_i)^2 = E(z_i^2) - [E(z_i)]^2 = E(z_i^2) - \mu_i^2$$

$$\int q_\theta(z) \log p(z) dz = -\frac{J}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^J (\mu_j^2 + \sigma_j^2)$$

$$\int q_\theta(z) \log q_\theta(z) dz = -H(z) = \int N(z; \mu, \sigma^2) \log N(z; \mu, \sigma^2) dz$$

$$\because \int e^{-x^2} dx = \sqrt{\pi}, \int u dv = uv - \int v du$$

$$\begin{aligned} \int x^2 e^{-x^2} dx &= \int x \cdot x e^{-x^2} dx \\ &= \left[x \frac{-e^{-x^2}}{2} \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \frac{-e^{-x^2}}{2} dx \\ &= 0 + \frac{1}{2} \int e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \end{aligned}$$

$$\begin{aligned}
-H(z) &= \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \left(-\log \sqrt{2\pi}\sigma - \frac{(z-\mu)^2}{2\sigma^2} \right) dz \\
&= \int -\frac{\log \sqrt{2\pi}\sigma}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz - \int \frac{(z-\mu)^2}{2\sigma^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz \\
&= -\frac{\log \sqrt{2\pi}\sigma * \sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int e^{-\frac{(z-\mu)^2}{2\sigma^2}} d\left(\frac{z-\mu}{\sqrt{2}\sigma}\right) \\
&\quad - \int \frac{(z-\mu)^2}{2\sigma^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz \\
&= -\frac{\log \sqrt{2\pi}\sigma * \sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int e^{-y^2} dy + \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int \frac{(z-\mu)^2}{2\sigma^2} e^{-\frac{(z-\mu)^2}{2\sigma^2}} d\left(\frac{z-\mu}{\sqrt{2}\sigma}\right) \\
&= -\frac{\log \sqrt{2\pi}\sigma}{\sqrt{\pi}} * \sqrt{\pi} - \frac{1}{\sqrt{\pi}} \int y^2 e^{-y^2} dy \\
&= -\frac{1}{2} \log(2\pi) - \frac{1}{2} (1 + \log \sigma_j^2)
\end{aligned}$$

Solution of $-KL[q_i(z_i)||Q_i^*(z_i)]$, Gaussian case I

$$\int q_{\theta}(z) \log p(z) dz = -\frac{J}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^J (\mu_j^2 + \sigma_j^2)$$

$$\begin{aligned} \int q_{\theta}(z) \log q_{\theta}(z) dz &= \int N(z; \mu, \sigma^2) \log N(z; \mu, \sigma^2) dz \\ &= -\frac{J}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^J (1 + \log \sigma_j^2) \end{aligned}$$

$$\begin{aligned} -KL[q_{\theta}(z)||p_{\theta}(z)] &= \int q_{\theta}(z) (\log p_{\theta}(z) - \log q_{\theta}(z)) dz \\ &= \frac{1}{2} \sum_{j=1}^J [(1 + \log \sigma_j^2) - \mu_j^2 - \sigma_j^2] \end{aligned}$$

Solution of $-KL[q_i(z_i)||Q_i^*(z_i)]$, Gaussian case

```
def kl_divergence(mu, logvar):  
    if mu.data.ndimension() == 4:  
        mu = mu.view(mu.size(0), mu.size(1))  
    if logvar.data.ndimension() == 4:  
        logvar = logvar.view(logvar.size(0), logvar.size  
                               (1))  
  
    kl_divergence = -0.5 * (1 + logvar - mu.pow(2) -  
                           logvar.exp())  
    sum_kl_divergence = kl_divergence.sum(1).mean(0,  
                                                  True)  
  
    return sum_kl_divergence
```