2024 秋季高级机器学习
习题三

Qiankun Ji 221300066

2024.12.12

一. (40 points) **概率图模型**

1. (20 points) 图 1是一个贝叶斯网络结构，请仿照教材 14.4.1 变量消去部分内容，推断图中边际概率 $P(x_5)$.

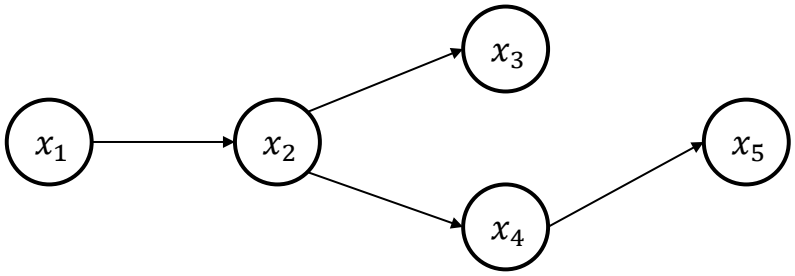

图 1: 贝叶斯网络结构

2. (20 points) 本题探究变分推断相关内容. 我们利用教材中相同的设定，假设当前有 N 个变量 $\{x_1, x_2, \ldots, x_N\}\}$ 均依赖于其他变量 z，所有能观察到的变量" 的联合分布的概率密度函数是:

$$p(\mathbf{x} \mid \Theta) = \prod_{i=1}^{N} \sum_{\mathbf{z}} p(x_i, \mathbf{z} \mid \Theta), \tag{1}$$

而所对应的对数似然函数为:

$$\ln p(\mathbf{x} \mid \Theta) = \sum_{i=1}^{N} \ln \left\{ \sum_{\mathbf{z}} p(x_i, \mathbf{z} \mid \Theta) \right\}, \tag{2}$$

其中 $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}, \Theta$ 是 $\mathbf{x}$ 与 $\mathbf{z}$ 服从的分布参数.

我们的推断任务是求解 $p(\mathbf{z} \mid \mathbf{x}, \Theta)$ 和 Θ. 一种有效手段是基于最大化对数似然函数，对 (2) 式使用 EM 算法: 在 E 步，根据 t 时刻的参数 Θ^t 对 $p(\mathbf{z} \mid \mathbf{x}, \Theta^t)$ 进行推断，并计算联合似然函数 $p(\mathbf{x}, \mathbf{z} \mid \Theta)$; 在 M 步，基于 E 步的结果进行最大化寻优，即对关于变量 Θ 的函数 $Q(\Theta; \Theta^t)$

进行最大化从而求取：

$$\Theta^{t+1} = \arg\max_{\Theta} Q(\Theta; \Theta^t) \tag{3}$$

$$= \arg\max_{\Theta} \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x}, \Theta^t) \ln p(\mathbf{x}, \mathbf{z} \mid \Theta). \tag{4}$$

（1）（10 points）$p(\mathbf{z} \mid \mathbf{x}, \Theta^t)$ 未必是隐变量 $\mathbf{z}$ 服从的真实分布，而只是一个近似分布. 现在将这个近似分布用 $q(\mathbf{z})$ 表示，请尝试验证

$$\ln p(\mathbf{x}) = \mathcal{L}(q) + KL(q \parallel p), \tag{5}$$

其中

$$\mathcal{L}(q) = \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} d\mathbf{z}, \tag{6}$$

$$KL(q \parallel p) = -\int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right\} d\mathbf{z}. \tag{7}$$

（2）（10 points）假设复杂的多变量 Z 可拆解为一系列相互独立的多变量 Z_i, 即 Z 服从分布：

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i), \tag{8}$$

尝试从最大化 $\mathcal{L}(q)$ 的角度说明变量子集 $\mathbf{z}_j$ 所服从的最优分布 q_j^* 应满足

$$\ln q_j^*(\mathbf{z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})] + \text{const}. \tag{9}$$

解：

1.

$$P(x_5) = \sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1} P(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1} P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_4)$$

$$= \sum_{x_4} P(x_5|x_4) \sum_{x_2} P(x_4|x_2) \sum_{x_3} P(x_3|x_2) \sum_{x_1} P(x_1)P(x_2|x_1)$$

$$= \sum_{x_4} P(x_5|x_4) \sum_{x_2} P(x_4|x_2) \sum_{x_3} P(x_3|x_2) m_{12}(x_2)$$

$$= \sum_{x_4} P(x_5|x_4) \sum_{x_2} P(x_4|x_2) \sum_{x_3} P(x_3|x_2)$$

$$= \sum_{x_4} P(x_5|x_4) \sum_{x_2} P(x_4|x_2) m_{12}(x_2) m_{32}(x_2)$$

$$= \sum_{x_4} P(x_5|x_4) m_{24}(x_4)$$

$$= m_{45}(x_5)$$

在以上的计算推断中我们采用 $\{x_1, x_3, x_2, x_4\}$ 的顺序计算加法，其中 $m_{ij}(x_j)$ 是求加过程的中间结果，下标 i 表示此项是对 x_i 的求加结果，下标 j 表示此项中剩下的其他变量。

2. (1) 验证公式 $\ln p(\mathbf{x}) = \mathcal{L}(q) + KL(q \parallel p)$

1. 定义项的展开：- 对数边际概率：$\ln p(\mathbf{x})$。- 变分下界：

$$\mathcal{L}(q) = \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} d\mathbf{z}.$$

- Kullback-Leibler (KL) 散度：

$$KL(q \parallel p) = -\int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right\} d\mathbf{z}.$$

2. 将联合分布分解为条件分布和边际分布：$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x})$，带入 $\mathcal{L}(q)$：

$$\mathcal{L}(q) = \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x})}{q(\mathbf{z})} \right\} d\mathbf{z}.$$

3. 分解对数：

$$\mathcal{L}(q) = \int q(\mathbf{z}) \ln p(\mathbf{x}) d\mathbf{z} + \int q(\mathbf{z}) \ln \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} d\mathbf{z}.$$

- 第一项：$\ln p(\mathbf{x})$ 是常数，积分后为：

$$\int q(\mathbf{z}) \ln p(\mathbf{x}) d\mathbf{z} = \ln p(\mathbf{x}).$$

- 第二项：重写为负的 KL 散度：

$$\int q(\mathbf{z}) \ln \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} d\mathbf{z} = -KL(q \parallel p).$$

4. 最终结果：

$$\ln p(\mathbf{x}) = \mathcal{L}(q) + KL(q \parallel p).$$

此式说明 $\ln p(\mathbf{x})$ 可以分解为变分下界 $\mathcal{L}(q)$ 和 KL 散度。

(2) 在此处我们简化 $q_i(\boldsymbol{z}_i)$ 为 q_i

$$
\begin{aligned}
L(q) &= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \, d\boldsymbol{z} \\
&= \int \sum_{i=1}^{M} q_i \left[\ln p(\boldsymbol{x}, \boldsymbol{z}) - \sum_{i=1}^{M} \ln q_i \right] d\boldsymbol{z} \\
= \int q_j \left(\int \ln p(\boldsymbol{x}, \boldsymbol{z}) \prod_{i=j} q_i \, d\boldsymbol{z}_i \right) d\boldsymbol{z}_j &- \int q_j \left(\int \ln q_j \prod_{i=j} q_i \, d\boldsymbol{z}_i \right) d\boldsymbol{z}_j - \int q_j \left(\int \sum_{i=j} q_i \prod_{i=j} q_i \, d\boldsymbol{z}_i \right) d\boldsymbol{z}_j \\
&= \int q_j \left(\int \ln p(\boldsymbol{x}, \boldsymbol{z}) \prod_{i=j} q_i \, d\boldsymbol{z}_i \right) d\boldsymbol{z}_j - \int q_j \ln q_j \, d\boldsymbol{z}_j - \int \sum_{i=j} q_i \prod_{i=j} q_i \, d\boldsymbol{z}_i \\
&= \int q_j \ln \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j) \, d\boldsymbol{z}_j - \int q_j \ln q_j \, d\boldsymbol{z}_j + \text{const}
\end{aligned}
$$

我们只关心 q_j，所以上式中最后一项与 q_j 无关，直接使用常数 const 替代，而第一个项中的 $\ln \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)$ 表示：

$$
\ln \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j) = \mathbb{E}_{i=j}[\ln p(\boldsymbol{x}, \boldsymbol{z})] + \text{const} = \int \ln p(\boldsymbol{x}, \boldsymbol{z}) \prod_{i=j} q_i \, dz_i + \text{const}
$$

$\mathbb{E}_{i=j}[\ln p(\boldsymbol{x}, \boldsymbol{z})]$ 即为 $\ln p(\boldsymbol{x}, \boldsymbol{z})$ 对随机变量 $\boldsymbol{z}_i, i = j$ 求期望，最后得到的是关于随机变量 $\boldsymbol{z}_j$ 的函数，所以可以被表示为 $\ln \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)$。

所以由以上的变化，我们固定了 $q_i = j$，可以针对 q_j 对 $L(q)$ 进行最大化，观察变化后的式子，有：

$$
\begin{aligned}
L(q) &= \int q_j \ln \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j) \, dz_j - \int q_j \ln q_j \, dz_j + \text{const} \\
&= -\text{KL}(q_j || \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)) + \text{const}
\end{aligned}
$$

所以有 KL 散度 $\text{KL}(q_j || \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j))$ 为 0 时，即分布 $q_j = \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)$ 时 $L(q)$ 最大，所以可知 $\boldsymbol{z}_j$ 所服从的最优分布 q_j^* 应满足

$$
\ln q_j^*(\boldsymbol{z}_j) = \ln \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j) = \mathbb{E}_{i=j}[\ln p(\boldsymbol{x}, \boldsymbol{z})] + \text{const}.
$$

二. (60 points) **强化学习**

1. (25 points) 价值迭代的更新公式为:

$$
V^{k+1}(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P\left(s' \mid s, a\right) V^k\left(s'\right) \right\}, \tag{10}
$$

其中 s 表示 t 时刻的状态，s' 表示 $t+1$ 时刻的状态，a 表示 t 时刻的动作，γ 是折扣因子. 我们将其定义为一个贝尔曼最优算子 $\mathcal{T}$:

$$V^{k+1}(s) = \mathcal{T}V^k(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P\left(s' \mid s,a\right) V^k\left(s'\right) \right\} \tag{11}$$

若 O 是一个算子，如果满足 $\|OV - OV'\|_q \leq \|V - V'\|_q$ 条件，则我们称 O 是一个压缩算子，其中 $\|x\|_q$ 表示 x 的 L_q 范数.

（1）（15 points）请证明，当 $\gamma < 1$ 时，贝尔曼最优算子 $\mathcal{T}$ 是一个 γ-压缩算子.（提示：证明 $\|\mathcal{T}V - \mathcal{T}V'\|_\infty \leq \gamma \|V - V'\|_\infty$ 即可）

（2）（10 points）在（1）的基础上，请说明价值迭代的收敛性.（提示：可以设最优价值函数为 V^*，考虑 $\left\|V^k - V^*\right\|_\infty$ 与迭代次数 k 的联系）

2. (15 points) 本题探究蒙特卡罗强化学习算法中的策略.

（1）（8 points）请你描述**重要性采样**的过程. 具体来说，我们希望估计某个函数 $f(x)$ 在概率分布 $p(x)$ 下的期望，但是 $p(x)$ 采样困难. 如何引入一个更容易采样的分布 $q(x)$ 来协助估计?

（2）（7 points）同策略蒙特卡罗强化学习算法和异策略蒙特卡罗强化学习算法有何差异? 请你结合上一问中提到的方法进行讨论.

3. (20 points) 时序差分学习（TD 学习）是一种在强化学习中广泛应用的核心技术，结合了动态规划和蒙特卡洛方法的优点，用于估计策略的价值函数. 它通过直接从与环境的交互中学习，既不需要完整的模型，也无需等待整条轨迹结束即可更新估计. 这种特性使 TD 学习在在线学习和实时决策任务中非常高效. 教材中介绍了一种属于 TD 学习的经典算法-Sarsa 算法，下方为完整算法流程：

输入： 环境 E;
动作空间 A;
起始状态 x_0;
奖赏折扣 γ;
更新步长 α.

过程：

$Q(x,a) = 0$, $\pi(x,a) = \frac{1}{|A(x)|}$;
$x = x_0$, $a = \pi(x)$;
for $t = 1, 2, \ldots$ **do**
  $r, x' =$ 在 E 中执行动作 a 产生的奖赏与转移的状态;
  $a' = \pi^\epsilon(x')$;
  $Q(x,a) = Q(x,a) + \alpha\big(r + \gamma Q(x',a') - Q(x,a)\big)$;
  $\pi(x) = \arg\max_{a''} Q(x,a'')$;
  $x = x', a = a'$
end for

输出： 策略 π

图 2: Sarsa 算法

结合状态值函数与状态-动作值函数的关系以及动态规划的特点，我们可以得到：

$$Q^{\pi}(x,a) = \sum_{x' \in X} P^{a}_{x \to x'} \left(R^{a}_{x \to x'} + \gamma V^{\pi}(x') \right) \tag{12}$$

$$= \sum_{x' \in X} P^{a}_{x \to x'} \left(R^{a}_{x \to x'} + \gamma \sum_{a' \in A} \pi(x', a') Q^{\pi}(x', a') \right). \tag{13}$$

请你根据式 (15),(16), 尝试推理出 Sarsa 算法的更新公式，即图二中的步骤 6.

解：

1. (1) 考虑我们的价值向量 $V, V' \in \mathbb{R}^{|S|}$，设 $\pi_1^* = \arg\max_{\pi}(r_\pi + \gamma P_\pi V)$, $\pi_2^* = \arg\max_{\pi}(r_\pi + \gamma P_\pi V')$，其中根据强化学习的定义 $\pi = P(a|s)$, $P_\pi = [P(s_j|s_i, a_i)]_{ij}$, $r_\pi = [r(s_i, a_i)]_i$。所以我们有：

$$TV = \max_{\pi}(r_\pi + \gamma P_\pi V) = r_{\pi_1^*} + \gamma P_{\pi_1^*} V \geq r_{\pi_2^*} + \gamma P_{\pi_2^*} V$$
$$TV' = \max_{\pi}(r_\pi + \gamma P_\pi V') = r_{\pi_2^*} + \gamma P_{\pi_2^*} V' \geq r_{\pi_1^*} + \gamma P_{\pi_1^*} V'$$

由上式可得：

$$\begin{aligned} TV - TV' &= r_{\pi_1^*} + \gamma P_{\pi_1^*} V - (r_{\pi_2^*} + \gamma P_{\pi_2^*} V') \\ &\leq r_{\pi_1^*} + \gamma P_{\pi_1^*} V - (r_{\pi_1^*} + \gamma P_{\pi_1^*} V') \\ &\leq \gamma P_{\pi_1^*}(V - V') \end{aligned}$$

同理有：

$$TV' - TV \leq \gamma P_{\pi_2^*}(V' - V)$$

所以结合上述两式可得：

$$\gamma P_{\pi_2^*}(V - V') \leq TV - TV' \leq \gamma P_{\pi_1^*}(V - V')$$

定义向量 z，其中 z 中的每个元素值，都是定义内两个向量中对应项绝对值的最大值

$$z = \max\{|\gamma P_{\pi_2^*}(V - V')|, |\gamma P_{\pi_1^*}(V - V')|\}$$

那么有

$$-z \leq TV - TV' \leq z$$

$$|TV - TV'| \leq z$$

$$||TV - TV'||_\infty \leq ||z||_\infty$$

假设 p_i^T, q_i^T 分别为 $P_{\pi_1^*}$, $P_{\pi_2^*}$ 的第 i 行，考虑 z 的第 i 个分量

$$z_i = \max\{\gamma|p_i^T(V - V')|, \gamma|q_i^T(V - V')|\}$$

由于 $P_{\pi_1^*}$, $P_{\pi_2^*}$ 的行向量中的所有分量的和为 1 且非负，所以有：

$$\gamma|p_i^T(V - V')| \leq p_i^T|V - V'| \leq \gamma||V - V'||_\infty$$
$$\gamma|q_i^T(V - V')| \leq q_i^T|V - V'| \leq \gamma||V - V'||_\infty$$

$$z_i = \max\{\gamma|p_i^T(V - V')|, \gamma|q_i^T(V - V')|\} \leq \gamma||V - V'||_\infty$$

所以有

$$||TV - TV'||_\infty \leq ||z||_\infty = \max\{z_i\} \leq \gamma||V - V'||_\infty$$

当 $\gamma < 1$ 时，有 $||TV - TV'||_\infty \leq ||V - V'||_\infty$，贝尔曼最优算子 T 是一个 γ-压缩算子。

(2) 设价值函数最终收敛到的最优价值函数为 V^*，每次的迭代我们都取 $V^{k+1} = TV^k$，所以有

$$\begin{aligned}||V^{k+1} - V^k||_\infty &= ||TV^k - TV^{k-1}||_\infty \\ &\leq \gamma||V^k - V^{k-1}||_\infty \\ &\leq \gamma^k||V_1 - V_0||_\infty\end{aligned}$$

那么 $\forall m, n > N$ 对于 V_m, V_n 中的每一个分量 V_i^m, V_i^n 都有：

$$\begin{aligned}|V_i^m - V_i^n| &= |V_i^m - V_i^{m-1} + V_i^{m-1} - \cdots - V_i^{n+1} + V_i^{n+1} - V_i^n| \\ &\leq |V_i^m - V_i^{m-1}| + |V_i^{m-1} - V_i^{m-2}| + \cdots + |V_i^{n+1} - V_i^n| \\ &\leq \gamma^{m-1}|V_i^1 - V_i^0| + \cdots + \gamma^n|V_i^1 - V_i^0| \\ &= \gamma^n(1 + \cdots + \gamma^{m-n-1})|V_i^1 - V_i^0| \\ &\leq \gamma^n(1 + \cdots + \gamma^{m-n-1} + \gamma^{m-n} + \cdots)|V_i^1 - V_i^0| \\ &= \frac{\gamma^n}{1-\gamma}|V_i^1 - V_i^0|\end{aligned}$$

所以 $|V_i^m - V_i^n|$ 在 n 趋于无穷时，收敛至 0。根据柯西收敛准则，V 的各个分量在该迭代下都收敛到值 V_i^*，最终 V 在该迭代下收敛到 V^*，所以价值迭代收敛。

总结来说：通过证明 $\mathcal{T}$ 是 γ-压缩算子，表明每次迭代的误差会按比例缩小；价值迭代利用上述误差收缩性质，保证了迭代过程的全局收敛性；具体收敛速度由 γ 和初始差距 $\|V^0 - V^*\|_\infty$ 决定。

2. (1) 重要性采样：估计函数 $f(x)$ 在概率分布下 $p(x)$ 下的期望，可以引入更容易采样的分布 $q(x)$ 来进行协助估计，有如下证明，所求期望 $E[f]$：

$$E[f] = \int p(x)f(x)\,dx$$

$$= \int q(x)\frac{p(x)}{q(x)}f(x)\,dx$$

函数 $f(x)$ 在概率分布下 $p(x)$ 下的期望，可以看作函数 $\frac{p(x)}{q(x)}f(x)$ 在更容易采样的分布 $q(x)$ 下的期望，所以原先的采样估计

$$\hat{E}[f] = \frac{1}{m}\sum_{i=1}^{m} f(x_i)$$

可以化为在 q 上的采样 $\{x'_1, \ldots, x'_m\}$ 估计：

$$\hat{E}[f] = \frac{1}{m}\sum_{i=1}^{m} \frac{p(x'_i)}{q(x'_i)}f(x'_i)$$

(2) 同策略蒙特卡罗强化学习算法和异策略蒙特卡罗强化学习算法的差异在于策略的采样方式和对累积奖赏期望的估计以及对上策略的更新改进上：

1. 同策略蒙特卡罗： - 直接从当前的目标策略 $\pi(a|s)$ 采样轨迹。
- 估计策略 π 本身的期望回报，无需权重修正。
- 公式：

$$G_t = \sum_{k=t}^{T} r_k,$$

在采样过后的则是生成 ϵ-贪心策略来作为更新后的策略。
- 优点：实现简单，无需计算重要性权重。
- 缺点：无法直接利用其他策略的经验数据。

2. 异策略蒙特卡罗：

- 采样时引入 ϵ-贪心，从行为策略 $b(a|s)$ 采样轨迹，而目标是估计另一策略 $\pi(a|s)$ 的期望回报。

- 需要使用重要性采样校正不同策略间的差异。
- 公式：

$$G_t^{\pi} = \frac{\pi(a_t|s_t)\pi(a_{t+1}|s_{t+1})\dots}{b(a_t|s_t)b(a_{t+1}|s_{t+1})\dots}\sum_{k=t}^{T} r_k.$$

- 优点：可以重用不同策略生成的轨迹。
- 缺点：重要性权重的方差可能较大，影响估计的稳定性。

3. 状态-动作值函数的贝尔曼期望方程为：

$$Q_{\pi}(x,a) = \sum_{x' \in X} P^{a}_{x \to x'}(R^{a}_{x \to x'} + \gamma \sum_{a' \in A} \pi(x',a')Q^{\pi}(x',a')$$

)

$$= \mathbf{E}_{x',a'}(R + \gamma Q(x',a'))$$

可知，$Q_{\pi}(x,a)$ 的取值依赖于该状态下到下一状态 x'，以及后续所做的动作 a'。根据 SARSA 算法每次的动作执行，状态转换我们对 $r + \gamma Q(x',a')$ 进行采样，由于每次的 $r + \gamma Q(x',a')$ 采样都是对 $R + \gamma Q(x',a')$ 的无偏估计，根据大数定律，在长期多次采样中，经验均值会收敛于真实的期望值。而根据均值增量式计算，我们分配一个较小的正数 α，作为更新步长，最终得到 SARSA 算法的更新公式：

$$Q(x,a) \leftarrow Q(x,a) + \alpha(r + \gamma Q(x',a') - Q(x,a))$$

这个更新公式与图中步骤 6 的公式是一致的，它反映了 Sarsa 算法如何通过实际获得的奖励和未来状态的估计值来更新当前状态-动作对的 Q 值。通过这种方式，Sarsa 算法能够逐步学习到一个最优策略，使得长期累积奖励最大化。