2024 秋季高级机器学习
习题三参考答案

2025.1.17

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 $P(x_5)$.

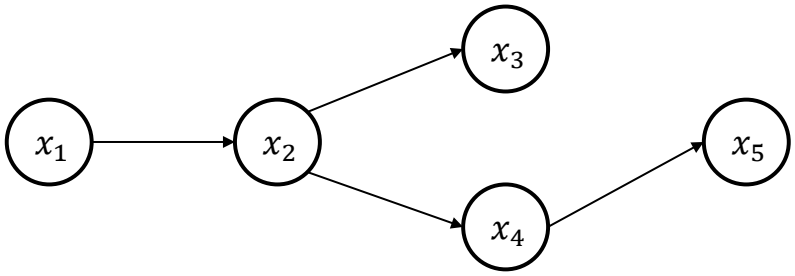

图 1: 贝叶斯网络结构

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.

$$\begin{aligned}
P(x_5) &= \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4} P(x_1, x_2, x_3, x_4, x_5) \\
&= \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4} P(x_1) P(x_2 \mid x_1) P(x_3 \mid x_2) P(x_4 \mid x_2) P(x_5 \mid x_4) \\
&= \sum_{x_4} P(x_5 \mid x_4) \sum_{x_2} P(x_4 \mid x_2) \sum_{x_3} P(x_3 \mid x_2) \sum_{x_1} P(x_1) P(x_2 \mid x_1) \\
&= \sum_{x_4} P(x_5 \mid x_4) \sum_{x_2} P(x_4 \mid x_2) \sum_{x_3} P(x_3 \mid x_2) m_{12}(x_2) \\
&= \sum_{x_4} P(x_5 \mid x_4) \sum_{x_2} P(x_4 \mid x_2) m_{32}(x_2) \\
&= \sum_{x_4} P(x_5 \mid x_4) m_{24}(x_4) \\
&= m_{45}(x_5)
\end{aligned}$$

2. (1)

$$\begin{aligned}
\ln p(\mathbf{x}) &= \int q(\mathbf{z}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{z} \\
&= \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{z} \mid \mathbf{x})} \right\} \mathrm{d}\mathbf{z} \\
&= \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z}, \mathbf{x}) q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x}) q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z} \\
&= \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z} - \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z} \\
&= \mathcal{L}(q) + \mathrm{KL}(q \| p).
\end{aligned}$$

(2) 代入 $q(\mathbf{z})$ 可得

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z} \\
&= \int \prod_{i=1}^{M} q_i\left(\mathbf{z}_i\right) \left\{ \ln p(\mathbf{x}, \mathbf{z}) - \ln \prod_{k=1}^{M} q_k\left(\mathbf{z}_k\right) \right\} \mathrm{d}\mathbf{z} \\
&= \int \prod_{i=1}^{M} q_i\left(\mathbf{z}_i\right) \ln p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} - \int \prod_{i=1}^{M} q_i\left(\mathbf{z}_i\right) \ln \prod_{k=1}^{M} q_k\left(\mathbf{z}_k\right) \mathrm{d}\mathbf{z} \\
&= \int q_j\left(\mathbf{z}_j\right) \left\{ \int \prod_{i \neq j} q_i\left(\mathbf{z}_i\right) \ln p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}_{i \neq j} \right\} \mathrm{d}\mathbf{z}_j - \int \prod_{i=1}^{M} q_i\left(\mathbf{z}_i\right) \sum_{k=1}^{M} \ln q_k\left(\mathbf{z}_k\right) \mathrm{d}\mathbf{z} \\
&= \int q_j\left(\mathbf{z}_j\right) \left(\ln \tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right) - \text{const} \right) \mathrm{d}\mathbf{z}_j - \int \prod_{i=1}^{M} q_i\left(\mathbf{z}_i\right) \ln q_j\left(\mathbf{z}_j\right) \mathrm{d}\mathbf{z} - \int \prod_{i=1}^{M} q_i\left(\mathbf{z}_i\right) \sum_{k \neq j} \ln q_k\left(\mathbf{z}_k\right) \mathrm{d}\mathbf{z} \\
&= \int q_j\left(\mathbf{z}_j\right) \left(\ln \tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right) - \text{const} \right) \mathrm{d}\mathbf{z}_j - \int q_j\left(\mathbf{z}_j\right) \ln q_j\left(\mathbf{z}_j\right) \mathrm{d}\mathbf{z}_j - \int \prod_{i \neq j} q_i\left(\mathbf{z}_i\right) \sum_{k \neq j} \ln q_k\left(\mathbf{z}_k\right) \mathrm{d}\mathbf{z}_{i \neq j} \\
&= \int q_j\left(\mathbf{z}_j\right) \left(\ln \tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right) - \text{const} \right) \mathrm{d}\mathbf{z}_j - \int q_j\left(\mathbf{z}_j\right) \ln q_j\left(\mathbf{z}_j\right) \mathrm{d}\mathbf{z}_j - \text{const}, \qquad (*)
\end{aligned}$$

其中

$$\ln \tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right) = \int \prod_{i \neq j} q_i\left(\mathbf{z}_i\right) \ln p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}_{i \neq j} + \text{const}.$$

等式 (*) 中第一个常数 const 用于归一化 $\tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right)$ 成一个概率，第二个常数是因为只考虑变量子集 $\mathbf{z}_j$ 的优化，因此最后一项可以看成是常数。

最终整理可得

$$\begin{aligned}
\mathcal{L}(q) &= \int q_j\left(\mathbf{z}_j\right) \ln \tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right) \mathrm{d}\mathbf{z}_j - \int q_j\left(\mathbf{z}_j\right) \ln q_j\left(\mathbf{z}_j\right) \mathrm{d}\mathbf{z}_j + \text{const} \\
&= -\mathrm{KL}\left(q_j\left(\mathbf{z}_j\right) \| \ln \tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right) \right) + \text{const} \leq \text{const},
\end{aligned}$$

当 $q_j(\mathbf{z}_j)$ 与 $\ln \tilde{p}(\mathbf{x}, \mathbf{z}_j)$ 分布相同时取等号，此时变量子集 $\mathbf{z}_j$ 最优，所服从的最优分布 q_j^* 应满足

$$\ln q_j^*(\mathbf{z}_j) = \ln \tilde{p}(\mathbf{x}, \mathbf{z}_j) = \int \prod_{i \neq j} q_i(\mathbf{z}_i) \ln p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}_{i \neq j} + \text{const} = \mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})] + \text{const},$$

获证。

二. (60 points) **强化学习**

1. (25 points) 价值迭代的更新公式为:

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

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 $\mathcal{T}$:

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

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：

1: $Q(x,a)=0,\ \pi(x,a)=\frac{1}{|A(x)|}$;
2: $x=x_0,\ a=\pi(x)$;
3: **for** $t=1,2,\ldots$ **do**
4: $r,x'=$ 在 E 中执行动作 a 产生的奖赏与转移的状态;
5: $a'=\pi^{\epsilon}(x')$;
6: $Q(x,a)=Q(x,a)+\alpha\big(r+\gamma Q(x',a')-Q(x,a)\big)$;
7: $\pi(x)=\arg\max_{a''} Q(x,a'')$;
8: $x=x',a=a'$
9: **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 (12),(13), 尝试推理出 Sarsa 算法的更新公式，即图二中的步骤 6.

解：

1. (1) 取 $\|\cdot\|_q$ 为 $\|\cdot\|_\infty$，则有

$$
\begin{aligned}
\|\mathcal{T}V - \mathcal{T}V'\|_\infty =& \max_{s\in\mathcal{S}} \left\{ \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P\left(s' \mid s,a\right) V\left(s'\right) \right\} \right. \\
& \left. - \max_{a'\in\mathcal{A}} \left\{ r\left(s,a'\right) + \gamma \sum_{s'\in\mathcal{S}} P\left(s' \mid s,a'\right) V'\left(s'\right) \right\} \right\} \\
\leq & \max_{s,a} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P\left(s' \mid s,a\right) V\left(s'\right) - r(s,a) - \gamma \sum_{s'\in\mathcal{S}} P\left(s' \mid s,a\right) V'\left(s'\right) \right\} \\
=& \gamma \max_{s,a} \left\{ \sum_{s'\in\mathcal{S}} P\left(s' \mid s,a\right) \left(V\left(s'\right) - V'\left(s'\right)\right) \right\} \\
\leq & \gamma \max_{s'} \left\{ V\left(s'\right) - V'\left(s'\right) \right\} \\
=& \gamma \left\| V - V' \right\|_\infty ,
\end{aligned}
$$

因此贝尔曼最优算子 $\mathcal{T}$ 是一个 γ-压缩算子。

（2）记最优价值函数为 V^*，由（1）可得

$$
\begin{aligned}
\left\| V^k - V^* \right\|_\infty &= \left\| \mathcal{T}V^{k-1} - \mathcal{T}V^* \right\|_\infty \\
&\leq \gamma \left\| V^{k-1} - V^* \right\|_\infty \\
&\leq \cdots \\
&\leq \gamma^k \left\| V^0 - V^* \right\|_\infty
\end{aligned}
$$

由于 $0 < \gamma < 1$，因此有

$$
\lim_{k\to\infty} \left\| V^k - V^* \right\|_\infty \leq \lim_{k\to\infty} \gamma^k \left\| V^0 - V^* \right\|_\infty = 0,
$$

因此价值迭代中当前价值函数与最优价值函数的无穷范数距离会随迭代次数指数收敛。

2.（1）我们希望估计的是期望

$$
\mathbb{E}_{x\sim p(x)}[f(x)] = \int p(x) f(x) dx
$$

由于 $p(x)$ 采样困难，因此可以引入一个分布 $q(x)$ 来协助估计，引入的分布 $q(x)$ 应尽量满足以下性质：1. 引入的分布 $q(x)$ 更容易采样；2. 对于 $p(x)$ 定义域内的任意 x，若 $p(x) > 0$，则应有 $q(x) > 0$，从而实现覆盖；3. 分布 $q(x)$ 应与分布 $p(x)$ 接近。引入后有

$$
\begin{aligned}
\mathbb{E}_{x\sim p(x)}[f(x)] &= \int p(x) f(x) dx \\
&= \int q(x) \frac{p(x)}{q(x)} f(x) dx \\
&= \mathbb{E}_{x\sim q(x)} \left[\frac{p(x)}{q(x)} f(x) \right]
\end{aligned}
$$

从而可以在 $q(x)$ 分布中采样，并使用权重 $\frac{p(x)}{q(x)}$ 修正结果，得到希望估计的期望。

（2）同策略蒙特卡罗强化学习算法中被评估与被改进的是同一个策略，因此采样得到的轨迹可以直接用于更新策略，而不需要使用重要性采样。异策略蒙特卡罗算法中采样策略与待更新的策略存在偏差，利用（1）中重要性采样，可以对不同策略的采样结果进行修正，从而使用异策略的轨迹更新策略。

3. Sarsa 算法的值函数跟新本质上是一种增量求和思想。注意式 (13) 的形式，我们可以考虑把基于 t 个采样已估计出值函数记为 $Q_t^\pi(x,a) = \frac{1}{t}\sum_{i=1}^{t} r_i$, 则在得到第 t + 1 个采样 r_{t+1} 时, 值函数的形式为:

$$Q_{t+1}^\pi(x,a) = Q_t^\pi(x,a) + \frac{1}{t+1}\left(r_{t+1} - Q_t^\pi(x,a)\right) \qquad \text{(增量求和式)}$$

令 x' 表示前一次在状态 x 执行动作 a 后转移到的状态，a' 表示策略 π 在 x' 上选择的动作。则我们会发现 t+1 步的奖赏即是状态 x 变化到 x' 的奖赏加上前面 t 步奖赏总和 $Q_t^\pi(x',a')$ 的 γ 折扣。我们可以写出等式:

$$r_{t+1} = R_{x\to x'}^a + \gamma Q_t^\pi(x',a')$$

再令

$$\frac{1}{t+1} = \alpha$$

则将上方两式带入增量求和式，我们将得到:

$$Q_{t+1}^\pi(x,a) = Q_t^\pi(x,a) + \alpha\left(R_{x\to x'}^a + \gamma Q_t^\pi(x',a') - Q_t^\pi(x,a)\right)$$

即为算法中步骤 6 的形式。