2023 秋季高级机器学习
习题三

2023.11.30

一. (30 points) **高斯混合模型**

一个由 K 个组分 (component) 构成的多维高斯混合模型的概率密度函数如下:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} P(z=k)\, p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

其中 z 是隐变量，$P(z)$ 表示 K 维离散分布，其参数为 $\boldsymbol{\pi}$，即 $p(z=k) = \pi_k$。$p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 表示参数为 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ 的多维高斯分布。

1. (10 points) 请使用盘式记法表示高斯混合模型。
2. (10 points) 考虑高斯混合模型的一个具体的情形，其中各个分量的协方差矩阵 $\boldsymbol{\Sigma}_k$ 全部被限制为一个共同的值 $\boldsymbol{\Sigma}$。求 EM 算法下参数 $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}$ 的更新公式。
3. (10 points) 考虑一个由下面的混合概率分布给出的概率密度模型:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|k) \tag{2}$$

并且假设我们将 $\boldsymbol{x}$ 划分为两部分，即 $\boldsymbol{x} = (\boldsymbol{x}_a, \boldsymbol{x}_b)$。证明条件概率分布 $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ 本身是一个混合概率分布。求混合系数以及分量概率密度的表达式。(注意此题没有规定 $p(\boldsymbol{x}|k)$ 的具体形式)

解:

1. 高斯混合分布可以理解为两次采样:

   step1: 多项分布中采样

   多项分布 Z 的参数为 $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$, 此时对应 Z=k 高斯混合分布先从多项分布 Z 中依据 Z 的取值，sample 出 $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$，得到即将产生 x 的高斯分布参数

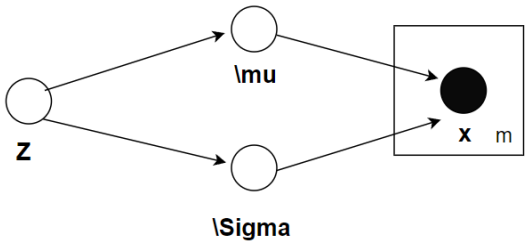
   step2: 高斯分布中采样

   从 Gaussian 分布 $P(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$ 中采样得到 $\boldsymbol{x}$

   综上，高斯混合分布模型的概率密度函数可以写成如下概率图形式

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} P(Z=k) \cdot P(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) = \sum_{Z,\boldsymbol{\mu},\boldsymbol{\Sigma}} P(Z) \cdot P(\boldsymbol{\mu}, |Z) \cdot P(\boldsymbol{x}|\boldsymbol{\mu},)$$

   对应的盘是记法也即下图



其中假设拿到的不完全数据集（不含隐变量）大小为 m，可以表示成 $\{\boldsymbol{x}_i\}_{i=1,2,...,m}$

2. 由 EM 算法，对参数的更新可以分成以下几步：

step1: 明确隐变量，求完全数据的对数似然

[1.1] 明确隐变量

高斯混合模型的隐变量也即：对每一个样本判断它是否来自子模型 i（第 i 个高斯分布，参数为 $\boldsymbol{\mu}_i, _i$）。所以此时的隐变量可以定义为

γ_j，是一个 K 维向量 one-hot 编码，指示第 j 个样本来自哪个子模型

$\gamma_{k,j}$，是一个标量，取值 $\{0,1\}$，指示第 j 个样本是否否自第 k 个子模型

所以综合所有隐变量，可以得到他们之间的关系

$$\Gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_m \end{pmatrix} = \begin{pmatrix} \gamma_{1,1} & \gamma_{2,1} & \cdots & \gamma_{K,1} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{1,m} & \gamma_{2,m} & \cdots & \gamma_{K,m} \end{pmatrix}$$

[1.2] 确定完全数据和对应参数集

定义隐变量后，完全数据和参数集可以表示为

$$(\boldsymbol{X}, \boldsymbol{\Gamma}) = \{(x_j, \gamma_{1,j}, ..., \gamma_{k,j})\}, \quad j \in [1, m]$$

$$\theta = (\theta_1, ..., \theta_K), \quad \theta_i = (\mu_i, \Sigma_i, \pi_i)$$

[1.3] 求解完全数据的对数似然

$$\begin{aligned} log(P(\boldsymbol{X}, \boldsymbol{\Gamma}|\theta)) &= log\left(P(\boldsymbol{X}|\boldsymbol{\Gamma}, \theta) \cdot P(\boldsymbol{\Gamma}|\theta)\right) \\ &= logP(\boldsymbol{X}|\boldsymbol{\Gamma}, \theta) + logP(\boldsymbol{\Gamma}|\theta) \\ &= log \prod_{j=1}^{m} P(\boldsymbol{x}_j|\boldsymbol{\Gamma}, \theta) + log \prod_{j=1}^{m} P(\gamma_j|\theta) \\ &= \sum_{j=1}^{m} logP(\boldsymbol{x}_j|\boldsymbol{\Gamma}, \theta) + \sum_{j=1}^{m} logP(\gamma_j|\theta) \end{aligned}$$

其中最后一行两项具体取值如下，

$$
\begin{aligned}
P(x_j|\Gamma,\theta) &= \prod_{k=1}^{K}(P(x_j|\theta_k))^{\gamma_{k,j}} \\
&= \prod_{k=1}^{K}\left(\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\cdot exp(-\frac{1}{2}(x_j-\mu_k)^T\Sigma^{-1}(x_j-\mu_k))\right)^{\gamma_{k,j}} \\
P(\gamma_j|\theta) &= \prod_{k=1}^{K}(\pi_j)^{\gamma_{k,j}}
\end{aligned}
$$

step2:E 步，确定 Q 函数

额外定义记号 $\hat{\gamma_{k,j}}$，指示迭代第 t 轮隐变量 $\gamma_{k,j}=1$ 时的概率

$$
\begin{aligned}
\hat{\gamma_{k,j}} &= p(\gamma_{k,j}=1|\boldsymbol{x}_j,\theta^i) \\
&= \frac{p(\boldsymbol{x}_j|\gamma_{k,j}=1,\theta^i)\cdot p(\gamma_{k,j}=1|\theta^i)}{p(\boldsymbol{x}_j|\theta^i)} \\
&= \frac{\pi_k^i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu_k}^i,\boldsymbol{\Sigma}^i)}{p(\boldsymbol{x}_j|\theta^i)} \\
&= \frac{\pi_k^i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu_k}^i,\boldsymbol{\Sigma}^i)}{\sum_{k=1}^{K}\pi_k^i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu_k}^i,\boldsymbol{\Sigma}^i)}
\end{aligned}
$$

$$
\begin{aligned}
Q(\theta,\theta^i) &= \mathbb{E}_{\Gamma|X,\theta^i}\left[logP(X,\Gamma|\theta)\right] \\
&= \mathbb{E}_{\Gamma|X,\theta^i}\left[\sum_{j=1}^{m}log(x_j|\Gamma,\theta)+\sum_{j=1}^{m}logP(\gamma_j|\theta)\right] \\
&= \mathbb{E}_{\Gamma|X,\theta^i}\left[\sum_{j=1}^{m}\sum_{k=1}^{K}\gamma_{k,j}\cdot\left(-\frac{1}{2}log|\Sigma|-\frac{1}{2}\cdot(x_j-\mu_k)^T\cdot\Sigma^{-1}\cdot(x_j-\mu_k)+log\pi_k-\frac{n}{2}log(2n)\right)\right] \\
&= \sum_{j=1}^{m}\sum_{k=1}^{K}\hat{\gamma_{k,j}}\cdot\left(-\frac{1}{2}log|\Sigma|-\frac{1}{2}\cdot(x_j-\mu_k)^T\cdot\Sigma^{-1}\cdot(x_j-\mu_k)+log\pi_k-\frac{n}{2}log(2n)\right) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{K}\hat{\gamma_{k,j}}\cdot\left(-\frac{1}{2}log|\Sigma|-\frac{1}{2}\cdot(x_j-\mu_k)^T\cdot\Sigma^{-1}\cdot(x_j-\mu_k)+log\pi_k\right)+const
\end{aligned}
$$

其中由 $\gamma_{k,j}$ 转换到 $\hat{\gamma_{k,j}}$ 时，是由于求期望往式子中多乘了 $p(\gamma_{k,j}=1|x_j,\theta^i)$ 项。由于 $\gamma_{k,j}$ 取值只能为 0、1，原式中本项存在只能是 $\gamma_{k,j}=1$ 时，此时期望公式中正好为 $p(\gamma_{k,j}=1|x_j,\theta^i)$ 项, 故可以改写为 $\hat{\gamma_{k,j}}$

此外，由于后续 M 步，是要对 θ 来求最大值，和 n 无关。故舍弃无关项，得到 Q 如上式所示。

step3：M 步，求解最大值

$$
Q^{(i+1)}=\arg\max_{\theta}Q(\theta,\theta^i)
$$

[3.1] 更新 π_k

由于 π_k 是有约束的，故无法直接求偏导求解。需使用拉格朗日函数以及对应 KKT 条件求解

此时的优化问题为,

$$\min -Q(\theta,\theta^i)$$
$$s.t.\sum_{k=1}^{K}\pi_k=1$$

对应的拉格朗日函数为,

$$L(\pi_1,...,\pi_K,\lambda)=-Q(\theta,\theta^i)+\lambda(\sum_{k=1}^{K}-1)$$

由于 L 是关于 $\pi_1,...,\pi_K$ 的凸函数，且满足 Slater 条件，故有强对偶性
所以可以使用充分必要条件的 KKT 条件 $\frac{\partial L}{\partial \pi^*}=0$

$$\because \frac{\partial L}{\partial \pi_k}=-\frac{\sum_{j=1}^{m}\hat{\gamma_{k,j}}}{\pi_k}+\lambda=0$$
$$\therefore \pi_k=\frac{\sum_{j=1}^{m}\hat{\gamma_{k,j}}}{\lambda}$$
$$\therefore \sum_{k=1}^{K}\pi_k=\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{j=1}^{m}=\frac{m}{\lambda}=1$$
$$\therefore \lambda=m$$
$$\therefore \pi_k^{(i+1)}=\frac{\sum_{j=1}^{m}\hat{\gamma_{k,j}}}{m}$$

[3.2] 更新 $\boldsymbol{\mu}_k$, 无约束优化由

$$\frac{\partial Q(\theta,\theta^i)}{\partial \mu_k}=\sum_{j=1}^{m}\hat{\gamma_{k,j}}\Sigma^{-1}(x_j-\mu_k)=0$$

等号右侧为零矩阵，等号两侧同时乘 Σ

$$\mu_k^{(i+1)}=\frac{\sum_{j=1}^{m}\hat{\gamma_{k,j}}\cdot \boldsymbol{x}_j}{\sum_{j=1}^{m}\hat{\gamma_{k,j}}}$$

[3.3] 更新 $\boldsymbol{\Sigma}$, 无约束优化【此时要所有的 Σ 一起更新，不能以某一个 Σ_k 的更新效果代替 Σ 的更新!】
同理用 $Q(\theta,\theta^i)$ 对 Σ 求偏导，令结果等于 0. 求得

$$\Sigma^{(i+1)}=\frac{\sum_{j=1}^{m}\sum_{k=1}^{K}\hat{\gamma_{k,j}}(x_j-\mu_k)(x_j-\mu_k)^T}{\sum_{j=1}^{m}\sum_{k=1}^{K}\hat{\gamma_{k,j}}}$$

由于对 $\sum_{k=1}^{K}\hat{\gamma_{k,j}}=1$（对某一个样本，属于所有子模型的概率加起来和为 1），所以更新式可以写成

$$\Sigma^{(i+1)}=\frac{\sum_{j=1}^{m}\sum_{k=1}^{K}\hat{\gamma_{k,j}}(x_j-\mu_k)(x_j-\mu_k)^T}{m}$$

3. **[3.1] 证明是混合概率分布**

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \frac{p(\boldsymbol{x}_a, \boldsymbol{x}_b)}{p(\boldsymbol{x}_b)} = \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}_b)} = \frac{\sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k)}{p(\boldsymbol{x}_b)}$$

由边缘概率分布的定义可得

$$p(\boldsymbol{x}_b) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_a = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k) d\boldsymbol{x}_a$$

假设 $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ 是一个概率分布

$$1 = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{\sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k)}{p(\boldsymbol{x}_b)} d\boldsymbol{x}_a = \frac{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k) d\boldsymbol{x}_a}{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k) d\boldsymbol{x}_a} = 1$$

所以 $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ 确实是一个混合概率分布

[3.2] 确定混合稀疏以及分量概率密度表达式

由上推导，各分量混合系数为 π_k，概率密度表达式为 $\frac{p(\boldsymbol{x}_a, \boldsymbol{x}_b|k)}{p(\boldsymbol{x}_b)}$

二. (40 points) **变分推断**

本题研究概率图模型里面的变分推断技术。现定义联合分布如下：

$$p(\mathbf{t}, \boldsymbol{w}, \alpha) = p(\mathbf{t} \mid \boldsymbol{w}) p(\boldsymbol{w} \mid \alpha) p(\alpha) \tag{3}$$

其中各具体分布为：

$$p(\mathbf{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}_n, \beta^{-1}\right) \tag{4}$$

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}\left(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1} \boldsymbol{I}\right) \tag{5}$$

$$p(\alpha) = \text{Gamma}\left(\alpha \mid a_0, b_0\right) = \frac{{b_0}^{a_0} \alpha^{a_0 - 1} e^{-b_0 \alpha}}{\Gamma(a_0)} \tag{6}$$

这里，$\mathcal{N}$ 代表高斯分布，$\text{Gamma}\left(\alpha \mid a_0, b_0\right)$ 表示变量为 α，参数为 a_0, b_0 的 Gamma 分布。

1. (10 points) 请使用盘式记法表示联合分布 $p(\mathbf{t}, \boldsymbol{w}, \alpha)$。
2. (20 points) 现在需要寻找对后验概率分布 $p(\boldsymbol{w}, \alpha \mid \mathbf{t})$ 的一个近似。使用变分框架进行分解，得到变分后验概率分布的分解表达式为 $q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w}) q(\alpha)$。首先计算 $q^*(\alpha)$，考虑 α 上的概率分布，利用教材公式（14.39），只保留与 α 有函数依赖关系的项，试证明

$$\ln q^*(\alpha) = (a_0 - 1) \ln \alpha - b_0 \alpha + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{w}\right] + \text{ 常数} \tag{7}$$

   这里 M 表示与 $\boldsymbol{w}$ 和 α 无关的常数。

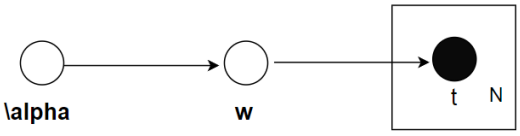
3. (10 points) 由上一题得到的结果，观察发现这是 Gamma 分布的对数，因此 $q^*(\alpha)$ 仍然服从如下 Gamma 分布

$$q^*(\alpha) = \text{Gamma}\left(\alpha \mid a_N, b_N\right) \tag{8}$$

   (a) 计算 a_N, b_N 的具体值（提示：由上一题计算 Gamma 分布对数 $\ln p(\alpha)$ 的结果观察 α 和 $\ln \alpha$ 的系数便可解决）
   (b) 比较 $q^*(\alpha)$ 与 $p(\alpha)$ 的相似度，总结变分推断的效果。（简单作答即可）

解：

1. $p(\boldsymbol{t}, \boldsymbol{w}, \alpha)$ 盘式记法为



2. 证明：

使用 $q(\boldsymbol{w}, \alpha)$ 来近似 $p(\boldsymbol{w}, \alpha|\boldsymbol{t})$，利用书中 14.39，且分解 $q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w}) \cdot q(\alpha)$ 有如下分析

$$\begin{aligned} ln(q^*(\alpha)) &= \mathbb{E}_{\boldsymbol{w}}\left[ln(p(\boldsymbol{w}, \alpha, \boldsymbol{t}))\right] + const \\ &= \int ln(p(\boldsymbol{w}, \alpha, \boldsymbol{t})) \cdot q(\boldsymbol{w}) d\boldsymbol{w} + const \\ &= \int \left(ln(p(\boldsymbol{t}|\boldsymbol{w})) + ln(p(\boldsymbol{w}|\alpha)) + ln(p(\alpha))\right) \cdot q(\boldsymbol{w}) d\boldsymbol{w} + const \end{aligned}$$

[2.1] 第一项

$$\begin{aligned} ln(p(\boldsymbol{t}|\boldsymbol{w})) &= ln \prod_{n=1}^{N} N(t_n|\boldsymbol{w}^T \cdot \phi_n, \beta^{-1}) \\ &= \sum_{n=1}^{N} ln(N(t_n|\boldsymbol{w}^T \cdot \phi_n, \beta^{-1})) \\ &= \sum_{n=1}^{N} \left(-\frac{1}{2} ln 2\pi - \frac{1}{2} ln|\beta^{-1}| - \frac{1}{2}(t_n - \boldsymbol{w}^T\phi_n)^T \beta (t_n - \boldsymbol{w}^T\phi_n)\right) \end{aligned}$$

此时 t_n 和 $\boldsymbol{w}^T\phi_n$ 都是标量，高斯概率密度函数中 n=1

[2.2] 第二项

$$\begin{aligned} ln(p(\boldsymbol{w}|\alpha)) &= lnN(\boldsymbol{w}|0, \alpha^{-1} \cdot I) \\ &= -\frac{M}{2} ln 2\pi - \frac{1}{2} ln|\alpha^{-1} \cdot I| - \frac{1}{2} \boldsymbol{w}^T \cdot I \cdot \alpha \cdot \boldsymbol{w} \\ &= -\frac{M}{2} ln 2\pi + \frac{M}{2} ln\alpha - \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} \end{aligned}$$

此时设定 $\boldsymbol{w}$ 和 $\boldsymbol{0}$ 都是 M 维向量，则此时高斯概率密度函数中 n=M

[2.3] 第三项

$$ln(p(\alpha)) = a_0 ln b_0 + (a_0 - 1) ln\alpha - b_0\alpha - ln(\Gamma(a_0))$$

综上，三项中都只保留和。且除了答案中涉及到的四项之外，其他项和 $\int q(\boldsymbol{w}) d\boldsymbol{w}$ 积分之后均为常数。所以由答案所示

$$lnq^*(\alpha) = (a_0 - 1) ln\alpha - b_0\alpha + \frac{M}{2} ln\alpha - \frac{\alpha}{2} \cdot \mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}^T \boldsymbol{w}] + const$$

3. **[3.1] 求解系数**

对第二问求出来的 $lnq^*(\alpha)$ 取对数

$$\begin{aligned} q^*(\alpha) &= exp(ln(\alpha^{(a_0-1)}) - b_0\alpha + ln(\alpha^{\frac{M}{2}}) - \frac{\alpha}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] + const) \\ &= \alpha^{(a_0-1)}\alpha^{\frac{M}{2}}\frac{1}{e^{b_0\alpha}e^{\frac{\alpha\mathbb{E}[\boldsymbol{w}^\mathrm{T}\boldsymbol{w}]}{2}}}e^{const} \\ &= \frac{\alpha^{\frac{M}{2}+a_0-1}.e^{-\alpha(b_0+\frac{\mathbb{E}[\boldsymbol{w}^\mathrm{T}\boldsymbol{w}]}{2})}}{e^{const1}}e^{const2} \end{aligned}$$

其中，令 $e^{const} = e^{const2-const1}$

对比系数，可以得到 a_Nb_N 的具体值为 $\begin{cases} a_N = \frac{M}{2} + a_0 \\ b_N = b_0 + \frac{\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]}{2} \end{cases}$

[3.2] 比较相似度，总结变分推断效果

首先，从 Gamma 分布的两参数来看，M 是 $\boldsymbol{w}, I$ 的维数，$\boldsymbol{w}$ 服从均值为 0，方差为 $\alpha^{-1}I$ 的正态分布，所以 $\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$ 的值不大。

所以在 M 值和 $\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$ 的值都不大的情况下，和原 Gamma 分布的相差不大，变分推断效果不错

其次，可以从 KL 散度的角度刻画两分布的相似程度

$$KL(p||q^*) = \int p(\alpha)ln\frac{p(\alpha)}{q^*(\alpha)}d\alpha \tag{9}$$

$$= \int p(\alpha)(lnp(\alpha) - lnq^*(\alpha))d\alpha \tag{10}$$

$$= \int \frac{b_0^{\alpha_0}.\alpha^{a_0-1}.e^{-b_0\alpha}}{\Gamma(a_0)}.\left(a_0lnb_0 - ln\Gamma(a_0) - \frac{M}{2}ln\alpha + \frac{\alpha}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] - const\right)d\alpha \tag{11}$$

三. (30 points) **强化学习 I**

1. (15 points) 值迭代用下面的贝尔曼最优方程来迭代地计算最优值函数：

$$V_n(s) \leftarrow \max_a\left(R(s,a) + \gamma\sum_{s'}T\left(s' \mid s,a\right)V_{n-1}\left(s'\right)\right) \tag{12}$$

其中 R 和 T 分别是奖赏函数和转移函数。令 V_k 为第 k 轮迭代得到的状态值函数，V^* 为最优状态值函数。请证明：当 $\|V_k - V_{k-1}\|_\infty < \delta, \delta = \frac{\epsilon(1-\gamma)}{\gamma}$ 时，$\|V^* - V_k\|_\infty < \epsilon$，即值迭代可以收敛得到最优值函数。

2. (8 points) 强化学习的目标是学习一个策略以最大化环境中的期望累积折扣奖赏 $J(\pi)$：

$$J(\pi) = \mathbb{E}_{\tau=(s_0,a_0,s_1,a_1,\ldots)\sim P_\pi(\cdot)}\left[\sum_{t=0}^{\infty}\gamma^t r\left(s_t,a_t\right)\right] \tag{13}$$

在智能体与环境的交互过程中，令采取策略 π 使得智能体在 t 时刻的状态为 s 的概率为 $P_t^\pi(s)$。对策略 π，我们定义状态动作对 (s,a) 被访问到的折扣概率为该策略的“占用度量”：

$$\rho^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_t^\pi(s)\pi(a \mid s) \tag{14}$$

请证明，强化学习的目标式可以写成：

$$J(\pi) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s,a\sim\rho^{\pi}(s,a)}\left[r(s,a)\right] \tag{15}$$

3. (7 points) 模仿学习要求在给定一些专家策略示例数据时学习一个策略以模仿专家的策略。请证明，任意策略的次优性（与专家策略的性能差距）有关于两个策略的占用度量的总变差的上界，即有下式：

$$\left|J\left(\pi^{E}\right) - J(\pi)\right| \leq \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{TV}\left(\rho^{\pi}, \rho^{E}\right) \tag{16}$$

其中奖赏函数满足 $\forall s, a, |R(s,a)| \leq R_{\max}$，总变差 $\mathbb{D}_{TV}(p,q)$ 可以度量 p, q 两个分布的差异：

$$\mathbb{D}_{TV}(p,q) = \frac{1}{2}\int_{x} |p(x) - q(x)|\,\mathrm{d}x \tag{17}$$

解：

1. 首先对原结论进行转化，如果可以证明以下结论，则原结论成立（是原结论的必要不充分条件）

$$\|V^* - V_k\|_{\infty} < \|V_k - V_{k-1}\|_{\infty} \cdot \frac{\gamma}{1-\gamma} < \frac{\epsilon(1-\gamma)}{\gamma} \cdot \frac{\gamma}{1-\gamma} = \epsilon$$

[**是原结论的必要条件 1**] 故以下尝试证明

$$\|V^* - V_k\|_{\infty} < \|V_k - V_{k-1}\|_{\infty} \cdot \frac{\gamma}{1-\gamma}$$

发现可以对结论式使用三角不等式 (假设从 V_{k+1} 开始，迭代更新 T 次值函数可以得到最优值函数 V^*)【以下均默认为无穷范数】

$$\begin{aligned}\|V^* - V_k\| &= \|(V^* - V_T) + (V_T - V_{T-1}) + ... + (V_{k+1} - V_k)\| \\ &\leq \|V^* - V_T\| + \|V_T - V_{T-1}\| + ... + \|V_{k+1} - V_k\|\end{aligned}$$

且由结论转化得到的关系很像等比数列求和的结果

[**是原结论的必要条件 2**] 故以下尝试证明

$$\|V_{k+2} - V_{k+1}\| \leq \gamma \|V_{k+1} - V_k\|$$

因为 V_{k+2} V_{k+1} V_k 都从相同的初始状态 s 出发，由值函数及策略单调递增的方式更新

令 a_1 和 a_0 分别为 V_{k+2} 及 V_{k+1} 从初始状态 s 选择的最优动作

$$a_1 = \arg\max_a R(s,a) + \gamma \sum_{s'} T(s'|s,a) V_{k+1}$$

$$a_0 = \arg\max_a R(s,a) + \gamma \sum_{s'} T(s'|s,a) V_k$$

由于 a_1 对于 V_{k+2} 来说最优、a_0 对于 V_{k+1} 来说最优。所以以下缩放时让 V_{k+1} 选择次优动作

a_1，对应值函数值会被缩小

$$\begin{aligned}
\|V_{k+2} - V_{k+1}\| &= \left\| R(s,a_1) + \gamma \sum_{s'} T(s'|s,a_1) V_{k+1}(s') - \left(R(s,a_0) + \gamma \sum_{s'} T(s'|s,a_0) V_k(s') \right) \right\| \\
&\leq \left\| R(s,a_1) + \gamma \sum_{s'} T(s'|s,a_1) V_{k+1}(s') - \left(R(s,a_1) + \gamma \sum_{s'} T(s'|s,a_1) V_k(s') \right) \right\| \\
&= \left\| \gamma \sum_{s'} T(s'|s,a) \left(V_{k+1}(s') - V_k(s') \right) \right\| \\
&\leq \gamma \sum_{s'} T(s'|s,a_1) \left\| V_{k+1}(s') - V_k(s') \right\| \\
&\leq \gamma \sum_{s'} T(s'|s,a_1) \left\| V_{k+1} - V_k \right\| \\
&= \gamma \left\| V_{k+1} - V_k \right\|
\end{aligned}$$

综上，结合原结论的两必要条件，可以证明原结论成立

$$\begin{aligned}
\|V^* - V_k\| &= \|(V^* - V_T) + (V_T - V_{T-1}) + ... + (V_{k+1} - V_k)\| \\
&\leq \|V^* - V_T\| + \|V_T - V_{T-1}\| + ... + \|V_{k+1} - V_k\| \\
&\leq (\gamma^{T+1} + \gamma^T + \cdots + \gamma) \|V_k - V_{k-1}\| \\
&= \frac{1 - \gamma^{T+1}}{1 - \gamma} \cdot \gamma \cdot \|V_k - V_{k-1}\| \\
&\leq \frac{\gamma}{1 - \gamma} \|V_k - V_{k-1}\| \\
&\leq \frac{\gamma}{1 - \gamma} \frac{\epsilon(1 - \gamma)}{\gamma} \\
&= \epsilon
\end{aligned}$$

2. 由于求期望的线性性质以及决策过程的马尔科夫性（新状态和动作只需要从上一个状态中生成，也即 $(s_t, a_t) \sim \rho^\pi(s,a)$）

$$\begin{aligned}
J(\pi) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tau=(s_0,a_0,...,)\sim p_\pi(\,\cdot\,)} \left[r(s_t, a_t) \right] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(s_t,a_t)\sim\rho^\pi(s_t,a_t)} \left[r(s_t, a_t) \right] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(s,a)\sim\rho^\pi(s,a)} \left[r(s, a) \right] \\
&= \frac{1 - \gamma^{\infty}}{1 - \gamma} \mathbb{E}_{(s,a)\sim\rho^\pi(s,a)} \left[r(s, a) \right] \\
&= \frac{1}{1 - \gamma} \mathbb{E}_{(s,a)\sim\rho^\pi(s,a)} \left[r(s, a) \right]
\end{aligned}$$

3.

$$
\begin{aligned}
\left|J(\pi^T) - J(\pi)\right| &= \frac{1}{1-\gamma}\left|\mathbb{E}_{(s,a)\sim\rho^{\pi^E}}[r(s,a)] - \mathbb{E}_{(s,a)\sim\rho^{\pi}}[r(s,a)]\right| \\
&= \frac{1}{1-\gamma}\left|\int_s\int_a r(s,a)\left(\rho^{\pi^E}(s,a) - \rho^{\pi}(s,a)\right)dsda\right| \\
&\leq \frac{1}{1-\gamma}\int_s\int_a |r(s,a)|\cdot\left|\rho^{\pi^E}(s,a) - \rho^{\pi}(s,a)\right|dsda \\
&\leq \frac{R_{max}}{1-\gamma}\int_s\int_a \left|\rho^{\pi^E}(s,a) - \rho^{\pi}(s,a)\right|dsda \\
&= \frac{2R_{max}}{1-\gamma}\mathbb{D}_{TV}(\rho^{\pi^E}, \rho^{\pi})
\end{aligned}
$$