人工智能导论第三次作业

张立博 2021012487

2023 年 6 月 24 日

1 第一题 贝叶斯网络

1.1

$$
\begin{gathered}
P(-e,-s,-m,-b) \\
= P(-e)P(-s|-e,-m)P(-m)P(-b|-m) \\
= 0.6 \times 0.9 \times 0.9 \times 0.9 \\
= 0.4374
\end{gathered}
$$

1.2

$$
\begin{gathered}
P(+b) \\
= P(+b|+m)P(+m) + P(+b|-m)P(-m) \\
= 1.0 \times 0.1 + 0.1 \times 0.9 \\
= 0.19
\end{gathered}
$$

1.3

$$
\begin{gathered}
P(+m|+b) \\
= \frac{P(+b|+m)P(+m)}{P(+b)} \\
= \frac{1.0 \times 0.1}{0.19} \\
= 0.5263
\end{gathered}
$$

1.4

$$
\begin{gathered}
P(+m|+s,+b,+e) \\
=\frac{P(+m,+s,+b,+e)}{P(+s,+b,+e)} \\
=\frac{P(+m)P(+s|+e,+m)P(+b|+m)P(+e)}{P(+m,+s,+b,+e)+P(-m,+s,+b,+e)} \\
=\frac{P(+m)P(+s|+e,+m)P(+b|+m)P(+e)}{P(+m)P(+s|+e,+m)P(+b|+m)P(+e)+P(-m)P(+s|+e,-m)P(+b|-m)P(+e)} \\
=\frac{0.1\times 1.0\times 1.0\times 0.4}{0.1\times 1.0\times 1.0\times 0.4+0.9\times 0.8\times 0.1\times 0.4} \\
=0.5814
\end{gathered}
$$

1.5

在 s 未知的情况下，m 和 e 是独立的，因此：

$$
\begin{gathered}
P(+e|+m) \\
=P(+e) \\
=0.4
\end{gathered}
$$

2 第二题 变量消除法

2.1

$$\tau_2(b,d,e,f,g)=\sum_c P(c|b)P(d|c)P(e|c,d)P(f|c,e)P(g|c,f)$$

剩余因子如下：

$$\tau_1(b),\tau_2(b,d,e,f,g)$$

2.2

$$\tau_3(b,d,f,g)=\sum_e \tau_2(b,d,e,f,g)$$

剩余因子如下：

$$\tau_1(b), \tau_3(b,d,f,g)$$

2.3

$$\tau_4(b,d,f) = \sum_g \tau_2(b,d,f,g)$$

剩余因子如下：

$$\tau_1(b), \tau_4(b,d,f)$$

2.4

$$\begin{aligned} &P(B=b, D=d|F=f) \\ &= \frac{\tau_1(b)\tau_4(b,d,f)}{\sum_b \tau_1(b)(\sum_d \tau_4(b,d,f))} \end{aligned}$$

2.5

变量消除顺序为：A, G, E, C

消除 A 后，剩余因子为：

$$\tau_1(b), P(c|b), P(d|c), P(e|c,d), P(f|c,e), P(g|c,f)$$

产生的新因子 $\tau_1(b)$ 大小为 1

消除 G 后，剩余因子为：

$$\tau_1(b), P(c|b), P(d|c), P(e|c,d), P(f|c,e), \tau_2(c,f)$$

产生的新因子 $\tau_2(c,f)$ 大小为 1

消除 E 后，剩余因子为：

$$\tau_1(b), P(c|b), P(d|c), \tau_2(c,f), \tau_3(c,d,f)$$

产生的新因子 $\tau_3(c,d,f)$ 大小为 2

消除 C 后，剩余因子为：

$$\tau_1(b), \tau_4(b,d,f)$$

产生的新因子 $\tau_4(b,d,f)$ 大小为 2

3 第三题 带方差的高斯线性回归

由已知

$$
\begin{aligned}
logp(D_n;w,\sigma) &= \sum_{i=1}^{N} logp(y_n|x_n,w,\sigma) \\
&= \sum_{i=1}^{N} log\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(y_n-w^Tx_n)^2}{2\sigma^2}) \\
&= -\frac{N}{2}log2\pi - \frac{N}{2}log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_n-w^Tx_n)^2
\end{aligned}
$$

分别对 w 和 σ 求导，得到

$$
\frac{\partial logp(D_n;w,\sigma)}{\partial w} = \sum_{i=1}^{N}\frac{y_n-w^Tx_n}{\sigma^2}x_n
$$

$$
\frac{\partial logp(D_n;w,\sigma)}{\partial \sigma} = \sum_{i=1}^{N}\frac{1}{\sigma} - \frac{(y_n-w^Tx_n)^2}{\sigma^3}
$$

令导数为 0，得到

$$
\sum_{i=1}^{N}\frac{y_n-w^Tx_n}{\sigma^2}x_n = 0
$$

$$
\sum_{i=1}^{N}\frac{1}{\sigma} - \frac{(y_n-w^Tx_n)^2}{\sigma^3} = 0
$$

解得

$$
w_{ML} = (X^TX)^{-1}X^TY
$$

$$
\sigma_{ML}^2 = \frac{1}{n}(Y-Xw_{ML})^T(Y-Xw_{ML})
$$

故

$$
\hat{w} = (X^TX)^{-1}X^TY
$$

$$
\hat{\sigma} = \sqrt{\frac{1}{n}(Y-X\hat{w})^T(Y-X\hat{w})}
$$

其中 x_n 为 $d\times 1$ 的向量，y_n 为标量，w 为 $d\times 1$ 的向量。
X 为 $n\times d$ 的矩阵，Y 为 $n\times 1$ 的向量，X 的第 i 行为 x_i^T，Y 的第 i 行为 y_i。

4 第四题 采样

4.1

$$\alpha_{ij} = min\{1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\}$$

下一个状态为 $(1,3)$ 的概率为

$$\begin{aligned}
\alpha_{(2,2),(1,3)} &= min\{1, \frac{p(1,3)Q}{p(2,2)Q}\} \\
&= min\{1, \frac{p(1,3)}{p(2,2)}\} \\
&= min\{1, \frac{1 + ln(1\times 3 + 2\times 1 + 3\times 3)}{2 + ln(2\times 2 + 2\times 2 + 3\times 2)}\} \\
&= min\{1, \frac{1 + ln(14)}{2 + ln(14)}\} \\
&= 0.784
\end{aligned}$$

下一个状态为 $(3,4)$ 的概率为

$$\begin{aligned}
\alpha_{(2,2),(3,4)} &= min\{1, \frac{p(3,4)Q}{p(2,2)Q}\} \\
&= min\{1, \frac{p(3,4)}{p(2,2)}\} \\
&= min\{1, \frac{3 + ln(3\times 4 + 2\times 3 + 3\times 4)}{2 + ln(2\times 2 + 2\times 2 + 3\times 2)}\} \\
&= min\{1, \frac{3 + ln(30)}{2 + ln(14)}\} \\
&= 1
\end{aligned}$$

4.2

若初始点位于 $(2,2)$，则第一次采样过程需要采样 x_1，提议分布 $q(x_1)$ 的形式为

$$\begin{aligned}
q(x_1) &= p(x_1|x_2) \\
&= \frac{p(x_1,x_2)}{p(x_2)} \\
&= \frac{x_1 + ln(x_1x_2 + 2x_1 + 3x_2)}{\sum_{x_1=1}^{100}(x_1 + ln(x_1x_2 + 2x_1 + 3x_2))} \\
&= \frac{x_1 + ln(x_1\times 2 + 2x_1 + 3\times 2)}{\sum_{x_1=1}^{100}(x_1 + ln(x_1\times 2 + 2x_1 + 3\times 2))}
\end{aligned}$$

第二次采样过程中需要采样的变量为 x_2，若采样到 $x_1 = 42$，则提议分布的形式为

$$\begin{aligned} q(x_2) &= p(x_2|x_1) \\ &= \frac{p(x_1, x_2)}{p(x_1)} \\ &= \frac{42 + ln(42x_2 + 2 \times 42 + 3x_2)}{\sum_{x_2=1}^{100}(42 + ln(42x_2 + 2 \times 42 + 3x_2))} \end{aligned}$$

5 第五题 Baum-Welch 算法

5.1

由已知:

$$q(Z) = P(Z|X, \theta^{old})$$

所以：

$$\begin{aligned} J(\theta) := L(q(Z), \theta) &= \sum_Z q(Z) log \frac{P(X, Z|\theta)}{q(Z)} \\ &= \sum_Z q(Z) log P(X, Z|\theta) - \sum_Z q(Z) log q(Z) \\ &= \sum_Z P(Z|X, \theta^{old}) log P(X, Z|\theta) - \sum_Z P(Z|X, \theta^{old}) log P(Z|X, \theta^{old}) \end{aligned}$$

又因为：

$$\hat{Z} = \arg\max_Z P(Z|X, \theta^{old}) \propto \arg\max_Z P(X, Z|\theta^{old})$$

所以：

$$\arg\max_\theta J(\theta) = \arg\max_\theta \sum_Z P(X, Z|\theta^{old}) log P(X, Z|\theta)$$

5.2

根据 HMM 的定义，可以将 $P(X, Z|\theta)$ 表示为：

$$P(X, Z|\theta) = P(x_1, x_2, ..., x_T, z_1, z_2, ..., z_T|\theta)$$

根据 HMM 的独立性假设

$$\begin{aligned}P(X,Z|\theta) &= P(x_1,x_2,...,x_T|z_1,z_2,...,z_T,\theta)P(z_1,z_2,...,z_T|\theta)\\ &= P(x_1,x_2,...,x_T|z_1,z_2,...,z_T,\theta)P(z_1|\theta)P(z_2|z_1,\theta)...P(z_T|z_{T-1},\theta)\\ &= P(x_1|z_1,\theta)P(x_2|z_2,\theta)...P(x_T|z_T,\theta)P(z_1|\theta)P(z_2|z_1,\theta)...P(z_T|z_{T-1},\theta)\end{aligned}$$

其中

$$P(x_t|z_t,\theta) = b_{z_t}(x_t)$$
$$P(z_t|z_{t-1},\theta) = a_{z_{t-1}z_t}$$
$$P(z_1|\theta) = \pi_{z_1}$$

所以

$$\begin{aligned}P(X,Z|\theta) &= \pi(z_1)B(z_1,x_1)A(z_1,z_2)B(z_2,x_2)...A(z_{T-1},z_T)B(z_T,x_T)\\ &= \pi(z_1)\prod_{t=1}^{T-1}A(z_t,z_{t+1})\prod_{t=1}^{T}B(z_t,x_t)\end{aligned}$$

5.3

令 $Q(\theta,\theta^{old}) = \sum_Z P(X,Z|\theta^{old})logP(X,Z|\theta)$，将 (2) 中结果代入得到

$$\begin{aligned}Q(\theta,\theta^{old}) &= \sum_Z P(X,Z|\theta^{old})logP(X,Z|\theta)\\ &= \sum_Z P(X,Z|\theta^{old})log\{\pi(z_1)\prod_{t=1}^{T-1}A(z_t,z_{t+1})\prod_{t=1}^{T}B(z_t,x_t)\}\\ &= \sum_Z P(X,Z|\theta^{old})log\pi(z_1) + \sum_Z P(X,Z|\theta^{old})\sum_{t=1}^{T-1}logA(z_t,z_{t+1}) + \sum_Z P(X,Z|\theta^{old})\sum_{t=1}^{T}logB(z_t,x_t)\end{aligned}$$

上式中 Q 被拆分成了三式之和，其中每一项仅与 θ 中的一个参数有关

5.4

证：

$$
\begin{aligned}
&\sum_{Z} P(X,Z|\theta^{old}) \log \pi(z_1) \\
&= \sum_{z_1=1}^{N} \sum_{z_2,...,z_T} P(X, z_1, z_2, ..., z_T|\theta^{old}) \log \pi(z_1) \\
&= \sum_{z_1=1}^{N} P(X, z_1|\theta^{old}) \sum_{z_2,...,z_T} P(z_2, ..., z_T|X, z_1, \theta^{old}) \log \pi(z_1) \\
&= \sum_{z_1=1}^{N} P(X, z_1|\theta^{old}) \log \pi(z_1) \sum_{z_2,...,z_T} P(z_2, ..., z_T|X, z_1, \theta^{old})
\end{aligned}
$$

由于内部求和符号中的条件概率 $P(z_2, ..., z_T|X, z_1, \theta^{old})$ 是对给定的 z_1 条件下后续的隐藏状态的概率，根据概率的归一性，有：

$$
\sum_{z_2,...,z_T} P(z_2, ..., z_T|X, z_1, \theta^{old}) = 1
$$

所以

$$
\begin{aligned}
&\sum_{Z} P(X,Z|\theta^{old}) log \pi(z_1) \\
&= \sum_{z_1=1}^{N} P(X, z_1|\theta^{old}) \log \pi(z_1) \\
&= \sum_{i=1}^{N} P(X, z_1 = i|\theta^{old}) \log \pi_i
\end{aligned}
$$

使用拉格朗日乘子法，求 $\sum_Z P(X, Z|\theta^{old}) \log \pi_{z_1}$ 在 $\sum_{i=1}^{N} \pi_i = 1$ 下取到极大值时 π_i 的值

首先定义拉格朗日函数：

$$
L(\pi, \lambda) = \sum_{i=1}^{N} P(X, z_1 = i|\theta^{old}) \log \pi_i + \lambda(\sum_{i=1}^{N} \pi_i - 1)
$$

题目要求 $\sum_Z P(X, Z|\theta^{old}) \log \pi_{z_1}$ 在 $\sum_{i=1}^{N} \pi_i = 1$ 下取到极大值，即找到使得拉格朗日函数取到极大值的 π_i 的值

对拉格朗日函数求偏导并令其为 0。
首先对 π_i 求偏导：

$$\frac{\partial L(\pi,\lambda)}{\partial \pi_i} = \frac{P(X, z_1 = i|\theta^{old})}{\pi_i} + \lambda = 0$$

然后对 λ 求偏导：

$$\frac{\partial L(\pi,\lambda)}{\partial \lambda} = \sum_{i=1}^{N} \pi_i - 1 = 0$$

联立求解得到：

$$\begin{aligned}
\pi_i &= -\frac{P(X, z_1 = i|\theta^{old})}{\lambda} \\
\sum_{i=1}^{N} \pi_i &= -\frac{\sum_{i=1}^{N} P(X, z_1 = i|\theta^{old})}{\lambda} = 1 \\
\lambda &= -\sum_{i=1}^{N} P(X, z_1 = i|\theta^{old}) \\
\pi_i &= \frac{P(X, z_1 = i|\theta^{old})}{\sum_{i=1}^{N} P(X, z_1 = i|\theta^{old})}
\end{aligned}$$

5.5

求解 $Q(\theta, \theta^{old})$ 取极大值时 a_{ij} 和 $b_j(k)$ 的取值
求解过程与上一小题类似，首先证明：

$$\sum_{Z} P(X, Z|\theta^{old}) \sum_{t=1}^{T-1} log A(z_t, z_{t+1}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(X, z_t = i, z_{t+1} = j|\theta^{old}) log a_{ij} \tag{1}$$

证：

$$\sum_Z P(X,Z|\theta^{old})\sum_{t=1}^{T-1} logA(z_t,z_{t+1})$$

$$=\sum_{z_1=1}^{N}\sum_{z_2,...,z_T} P(X,z_1,z_2,...,z_T|\theta^{old})\sum_{t=1}^{T-1} logA(z_t,z_{t+1})$$

$$=\sum_{z_1=1}^{N} P(X,z_1|\theta^{old})\sum_{z_2,...,z_T} P(z_2,...,z_T|X,z_1,\theta^{old})\sum_{t=1}^{T-1} logA(z_t,z_{t+1})$$

$$=\sum_{z_1=1}^{N} P(X,z_1|\theta^{old})\sum_{z_2,...,z_T} P(z_2,...,z_T|X,z_1,\theta^{old})\sum_{t=1}^{T-1} loga_{z_t z_{t+1}}$$

$$=\sum_{z_1=1}^{N} P(X,z_1|\theta^{old})\sum_{z_2,...,z_T} P(z_2,...,z_T|X,z_1,\theta^{old})\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} I(z_t=i,z_{t+1}=j)loga_{ij}$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1}\sum_{z_1=1}^{N} P(X,z_1|\theta^{old})P(z_2,...,z_T|X,z_1,\theta^{old})I(z_t=i,z_{t+1}=j)loga_{ij}$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(X,z_t=i,z_{t+1}=j|\theta^{old})loga_{ij}$$

使用拉格朗日乘子法，求 $\sum_Z P(X,Z|\theta^{old})\sum_{t=1}^{T-1} logA(z_t,z_{t+1})$ 在 $\sum_{j=1}^{N} a_{ij}=1$ 下取到极大值时 a_{ij} 的值
首先定义拉格朗日函数：

$$L(A,\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(X,z_t=i,z_{t+1}=j|\theta^{old})loga_{ij}+\lambda(\sum_{j=1}^{N} a_{ij}-1)$$

题目要求 $\sum_Z P(X,Z|\theta^{old})\sum_{t=1}^{T-1} logA(z_t,z_{t+1})$ 在 $\sum_{j=1}^{N} a_{ij}=1$ 下取到极大值，即找到使得拉格朗日函数取到极大值的 a_{ij} 的值
对拉格朗日函数求偏导并令其为 0。
首先对 a_{ij} 求偏导：

$$\frac{\partial L(A,\lambda)}{\partial a_{ij}}=\frac{\sum_{t=1}^{T-1} P(X,z_t=i,z_{t+1}=j|\theta^{old})}{a_{ij}}+\lambda=0$$

然后对 λ 求偏导：

$$\frac{\partial L(A,\lambda)}{\partial \lambda}=\sum_{j=1}^{N} a_{ij}-1=0$$

联立求解得到：

$$a_{ij} = -\frac{\sum_{t=1}^{T-1} P(X, z_t = i, z_{t+1} = j|\theta^{old})}{\lambda}$$

$$\sum_{j=1}^{N} a_{ij} = -\frac{\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(X, z_t = i, z_{t+1} = j|\theta^{old})}{\lambda} = 1$$

$$\lambda = -\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(X, z_t = i, z_{t+1} = j|\theta^{old})$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} P(X, z_t = i, z_{t+1} = j|\theta^{old})}{\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(X, z_t = i, z_{t+1} = j|\theta^{old})}$$

接着证明：

$$\sum_{Z} P(X, Z|\theta^{old}) \sum_{t=1}^{T} logB(z_t, x_t) = \sum_{j=1}^{N}\sum_{k=1}^{M}\sum_{t=1}^{T} P(X, z_t = j, x_t = k|\theta^{old}) logb_j(k) \tag{2}$$

证：

$$\sum_{Z} P(X, Z|\theta^{old}) \sum_{t=1}^{T} logB(z_t, x_t)$$

$$= \sum_{z_1=1}^{N}\sum_{z_2,...,z_T} P(X, z_1, z_2, ..., z_T|\theta^{old}) \sum_{t=1}^{T} logB(z_t, x_t)$$

$$= \sum_{z_1=1}^{N} P(X, z_1|\theta^{old}) \sum_{z_2,...,z_T} P(z_2, ..., z_T|X, z_1, \theta^{old}) \sum_{t=1}^{T} logB(z_t, x_t)$$

$$= \sum_{z_1=1}^{N} P(X, z_1|\theta^{old}) \sum_{z_2,...,z_T} P(z_2, ..., z_T|X, z_1, \theta^{old}) \sum_{t=1}^{T} logb_{z_t}(x_t)$$

$$= \sum_{z_1=1}^{N} P(X, z_1|\theta^{old}) \sum_{z_2,...,z_T} P(z_2, ..., z_T|X, z_1, \theta^{old}) \sum_{j=1}^{N}\sum_{k=1}^{M}\sum_{t=1}^{T} I(z_t = j, x_t = k) logb_j(k)$$

$$= \sum_{j=1}^{N}\sum_{k=1}^{M}\sum_{t=1}^{T}\sum_{z_1=1}^{N} P(X, z_1|\theta^{old}) P(z_2, ..., z_T|X, z_1, \theta^{old}) I(z_t = j, x_t = k) logb_j(k)$$

$$= \sum_{j=1}^{N}\sum_{k=1}^{M}\sum_{t=1}^{T} P(X, z_t = j, x_t = k|\theta^{old}) logb_j(k)$$

使用拉格朗日乘子法，求 $\sum_Z P(X, Z|\theta^{old}) \sum_{t=1}^{T} logB(z_t, x_t)$ 在 $\sum_{k=1}^{M} b_j(k) = 1$ 下取到极大值时 $b_j(k)$ 的值

首先定义拉格朗日函数：

$$L(B,\lambda)=\sum_{j=1}^{N}\sum_{k=1}^{M}\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})logb_j(k)+\lambda(\sum_{k=1}^{M}b_j(k)-1)$$

题目要求 $\sum_Z P(X,Z|\theta^{old})\sum_{t=1}^{T}logB(z_t,x_t)$ 在 $\sum_{k=1}^{M}b_j(k)=1$ 下取到极大值，即找到使得拉格朗日函数取到极大值的 $b_j(k)$ 的值
对拉格朗日函数求偏导并令其为 0。
首先对 $b_j(k)$ 求偏导：

$$\frac{\partial L(B,\lambda)}{\partial b_j(k)}=\frac{\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})}{b_j(k)}+\lambda=0$$

然后对 λ 求偏导：

$$\frac{\partial L(B,\lambda)}{\partial\lambda}=\sum_{k=1}^{M}b_j(k)-1=0$$

联立求解得到：

$$\begin{aligned}b_j(k)&=-\frac{\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})}{\lambda}\\\sum_{k=1}^{M}b_j(k)&=-\frac{\sum_{k=1}^{M}\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})}{\lambda}=1\\\lambda&=-\sum_{k=1}^{M}\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})\\b_j(k)&=\frac{\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})}{\sum_{k=1}^{M}\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})}\end{aligned}$$

5.6

由已知

$$\begin{aligned}\gamma_t^{old}(i)&=P(z_t=i|X,\theta^{old})\\\xi_t^{old}(i,j)&=P(z_t=i,z_{t+1}=j|X,\theta^{old})\end{aligned}$$

使用课上讲过的前向-后向算法，对于给定参数 θ^{old} 的 HMM，可以求出：

$$\alpha_t^{old}(i) = P(z_t = i | x_1, x_2, ..., x_t, \theta^{old})$$
$$\beta_t^{old}(i) = P(x_{t+1}, x_{t+2}, ..., x_T | z_t = i, \theta^{old})$$

下面使用 $\alpha_t^{old}(i), \beta_t^{old}(i), a_{ij}^{old}, b_j^{old}(k)$ 表示出 $\gamma_t^{old}(i), \xi_t^{old}(i, j)$
首先表示出 $\gamma_t^{old}(i)$：

$$\begin{aligned}
\gamma_t^{old}(i) &= P(z_t = i | X, \theta^{old}) \\
&= \frac{P(z_t = i, X | \theta^{old})}{P(X | \theta^{old})} \\
&= \frac{P(z_t = i, x_1, x_2, ..., x_T | \theta^{old})}{P(X | \theta^{old})} \\
&= \frac{P(z_t = i, x_1, x_2, ..., x_t, x_{t+1}, ..., x_T | \theta^{old})}{P(X | \theta^{old})} \\
&= \frac{P(z_t = i, x_1, x_2, ..., x_t | \theta^{old}) P(x_{t+1}, ..., x_T | z_t = i, \theta^{old})}{P(X | \theta^{old})} \\
&= \frac{P(z_t = i, x_1, x_2, ..., x_t | \theta^{old}) P(x_{t+1}, ..., x_T | z_t = i, \theta^{old})}{\sum_{j=1}^{N} P(z_t = j, x_1, x_2, ..., x_t | \theta^{old}) P(x_{t+1}, ..., x_T | z_t = j, \theta^{old})} \\
&= \frac{\alpha_t^{old}(i) \beta_t^{old}(i)}{\sum_{j=1}^{N} \alpha_t^{old}(j) \beta_t^{old}(j)}
\end{aligned}$$

然后表示出 $\xi_t^{old}(i,j)$：

$$
\begin{aligned}
\xi_t^{old}(i,j) &= P(z_t=i, z_{t+1}=j|X,\theta^{old}) \\
&= \frac{P(z_t=i, z_{t+1}=j, X|\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i, z_{t+1}=j, x_1, x_2, ..., x_T|\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i, z_{t+1}=j, x_1, x_2, ..., x_t, x_{t+1}, ..., x_T|\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i, z_{t+1}=j, x_1, x_2, ..., x_t|\theta^{old})P(x_{t+1}, ..., x_T|z_t=i, z_{t+1}=j, \theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i, x_1, x_2, ..., x_t|\theta^{old})P(z_{t+1}=j|x_1, x_2, ..., x_t, z_t=i, \theta^{old})P(x_{t+1}, ..., x_T|z_t=i, z_{t+1}=j, \theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i, x_1, x_2, ..., x_t|\theta^{old})P(z_{t+1}=j|x_1, x_2, ..., x_t, z_t=i, \theta^{old})P(x_{t+1}, ..., x_T|z_{t+1}=j, \theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i, x_1, x_2, ..., x_t|\theta^{old})P(z_{t+1}=j|x_1, x_2, ..., x_t, z_t=i, \theta^{old})P(x_{t+1}, ..., x_T|z_{t+1}=j, \theta^{old})}{\sum_{i=1}^N \sum_{j=1}^N P(z_t=i, z_{t+1}=j, x_1, x_2, ..., x_t|\theta^{old})P(x_{t+1}, ..., x_T|z_t=i, z_{t+1}=j, \theta^{old})} \\
&= \frac{\alpha_t^{old}(i)a_{ij}^{old}b_j^{old}(x_{t+1})\beta_{t+1}^{old}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t^{old}(i)a_{ij}^{old}b_j^{old}(x_{t+1})\beta_{t+1}^{old}(j)}
\end{aligned}
$$

5.7

回顾 5.4 和 5.5 的结果，使用 $\gamma_t^{old}(i), \xi_t^{old}(i,j)$ 表示出 $\pi_i, a_{ij}, b_j(k)$

首先表示出 π_i：

$$
\begin{aligned}
\pi_i &= \frac{P(z_1=i|X,\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_1=i, X|\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_1=i, x_1, x_2, ..., x_T|\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_1=i, x_1, x_2, ..., x_T|\theta^{old})}{\sum_{j=1}^N P(z_1=j, x_1, x_2, ..., x_T|\theta^{old})} \\
&= \frac{\sum_{t=1}^T P(X, z_1=i, x_1, x_2, ..., x_T|\theta^{old})}{\sum_{j=1}^N \sum_{t=1}^T P(X, z_1=j, x_1, x_2, ..., x_T|\theta^{old})} \\
&= \frac{\sum_{t=1}^T \gamma_t^{old}(i)}{\sum_{j=1}^N \sum_{t=1}^T \gamma_t^{old}(j)}
\end{aligned}
$$

然后表示出 a_{ij}：

$$
\begin{aligned}
a_{ij} &= \frac{P(z_t=i,z_{t+1}=j|X,\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i,z_{t+1}=j,x_1,x_2,...,x_T|\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(z_t=i,z_{t+1}=j,x_1,x_2,...,x_T|\theta^{old})}{\sum_{i=1}^{N}\sum_{j=1}^{N}P(z_t=i,z_{t+1}=j,x_1,x_2,...,x_T|\theta^{old})} \\
&= \frac{\sum_{t=1}^{T-1}P(X,z_t=i,z_{t+1}=j|\theta^{old})}{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1}P(X,z_t=i,z_{t+1}=j|\theta^{old})} \\
&= \frac{\sum_{t=1}^{T-1}\xi_t^{old}(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1}\xi_t^{old}(i,j)}
\end{aligned}
$$

最后表示出 $b_j(k)$：

$$
\begin{aligned}
b_j(k) &= \frac{P(x_t=k|z_t=j,X,\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(x_t=k,z_t=j,X|\theta^{old})}{P(X|\theta^{old})} \\
&= \frac{P(x_t=k,z_t=j,X|\theta^{old})}{\sum_{j=1}^{N}P(x_t=k,z_t=j,X|\theta^{old})} \\
&= \frac{\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})}{\sum_{j=1}^{N}\sum_{t=1}^{T}P(X,z_t=j,x_t=k|\theta^{old})} \\
&= \frac{\sum_{t=1}^{T}\gamma_t^{old}(j)I(x_t=k)}{\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_t^{old}(j)}
\end{aligned}
$$

结合上述推理过程，Baum-Welch 算法的伪代码如下：

重复直到收敛：

E 步：

使用前向-后向算法计算 $\alpha_t^{old}(i),\beta_t^{old}(i)$

使用 $\alpha_t^{old}(i),\beta_t^{old}(i),a_{ij}^{old},b_j^{old}(k)$ 表示出 $\gamma_t^{old}(i),\xi_t^{old}(i,j)$

M 步：

使用 $\gamma_t^{old}(i),\xi_t^{old}(i,j)$ 表示出 $\pi_i,a_{ij},b_j(k)$

使用 $\pi_i,a_{ij},b_j(k)$ 更新模型参数 θ

使用更新后的模型参数 θ 计算 $Q(\theta,\theta^{old})$

计算 $Q(\theta,\theta^{old})$ 的相对变化 δ

更新 θ^{old} 为 θ

输出： 模型参数 θ

6 第六题 LDA

6.1

Variational EM LDA 的伪代码如下：

```
预处理数据集，得到docs和vocab
初始化lda部分模型参数：K(主题数),V(词汇表大小),alpha(超参数)
对给定的文档集合进行模型拟合：
    初始化监测收敛性的变量prev_bound(inf),delta(-inf),D(文档数)
    根据数据初始化模型参数：
        使用随机选择的文档子集的词频计数初始化每个主题
        通过最大似然估计计算主题-词分布的对数矩阵log_betas(V,K)
        初始化主题-文档分布矩阵gammas(D,K)
        # 重复优化参数直到收敛
        for i in range(max_epochs):
            更新参数并计算bound的值
            计算bound的相对变化delta
            打印进度
            更新prev_bound为当前bound
            检查是否收敛：若是则停止迭代
            如果vocab不为空且选择展示主题：输出每个主题的高频词
```

6.2

完整代码见附件 **main.py**

6.3

设置不同的 K，得到的结果如下：

```
0 孩子 情况 网友 时间 工作 介绍 事发 发生
1 医院 公司 警方 发现 孩子 妻子 昨日 医生
2 孩子 调查 发现 现场 工作人员 男子 万元 人员
3 发现 孩子 警方 发生 男子 学生 女士 老人
4 民警 男子 孩子 手机 嫌疑人 警方 报警 派出所
```

图 1: K=5

```
0 男子 警方 民警 发现 现场 人员 发生 报警
1 发现 告诉 现场 昨日 找到 相关 医院 万元
2 男子 法院 张某 死亡 医生 发现 情况 案件
3 孩子 妻子 儿子 时间 家人 告诉 医院 离开
4 万元 车辆 发生 发现 公司 法院 赔偿 被告
5 父亲 手机 发生 母亲 生活 孩子 妈妈 事情
6 公司 发现 工作 现场 调查 同学 中国 医院
7 法院 犯罪 被告人 医院 村民 老人 民警 发现
8 民警 孩子 下午 交警 工作 女子 手机 警察
9 孩子 发现 老师 女士 警方 家长 女儿 情况
```

图 2: K=10

```
0 孩子 男子 妈妈 医生 网友 公司 女士 法院
1 发生 公司 司机 发现 现场 男子 告诉 事故
2 儿子 医院 李某 发现 告诉 李先生 死亡 人员
3 孩子 民警 男子 两人 法院 陈某 公司 发生
4 男子 现场 民警 司机 公安局 孩子 介绍 情况
5 民警 嫌疑人 男子 工作 警方 医院 案件 发生
6 发现 学生 孩子 情况 学校 广州 调查 男子
7 警方 发现 调查 网友 万元 孩子 事情 情况
8 学生 发现 学校 现场 工作 生活 老师 时间
9 老人 孩子 下午 女士 事发 调查 现场 时许
10 民警 警方 发现 孩子 家中 情况 调查 下午
11 孩子 介绍 提供 男子 公司 标题 人员 女士
12 孩子 老师 学生 学校 处罚 昨日 告诉 家长
13 女子 万元 银行 发现 孩子 告诉 电话 男子
14 发现 老人 民警 医院 报警 女儿 大爷 警方
15 情况 交警 电话 视频 警察 相关 告诉 手机
16 房屋 发生 赔偿 法院 医院 万元 死亡 孩子
17 孩子 警方 男子 介绍 父亲 回家 公司 标题
18 警方 孩子 车辆 报道 现场 男子 公司 标题
19 民警 警方 发现 派出所 调查 医院 介绍 手机
```

图 3: K=15

6.4

观察结果发现，分类效果最好的是 K=10 的情况，效果比较明显的主题如：

1. 主题 0：男子 警方 民警 发现 现场 人员 发生 报警
2. 主题 5：父亲 手机 发生 母亲 生活 孩子 妈妈 事情

可能的原因如下：

1. K=5 时，每个主题的词汇量较大，负责解释更多的数据变化，这可能导致模型欠拟合，无法捕捉数据中的复杂结构，分类效果较差
2. K=15 时，每个主题的词汇量较小，可能导致模型过拟合，将噪声也作为主题进行建模，影响分类效果，分类效果较差
3. K=10 时，每个主题的词汇量适中，适当地平衡了模型的复杂性和表达能力方面，

分类效果较好