AI2651 智能语音识别：Baum-Welch 算法

521030910387 薛翔元

GMM-HMM 模型的参数如下表所示

名称	记号	参数
转移概率	$a_{ij} = P\left(q_t = j \mid q_{t-1} = i\right)$	a_{ij}
输出概率	$b_j\left(o_t\right) = p\left(o_t \mid q_t = j\right)$	$c_{jm}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}$

为了估计模型参数，我们需要最大化似然函数

$$\hat{\theta} = \arg\max_{\theta} \prod_{r=1}^{R} p\left(\boldsymbol{O}^{(r)} \mid \theta\right)$$

其中 R 是语音片段的数量，$\boldsymbol{O}^{(r)}$ 是第 r 个语音片段的观测序列。于是可以得到对数似然函数

$$\mathcal{L}(\theta) = \sum_{r=1}^{R} \log p\left(\boldsymbol{O}^{(r)} \mid \theta\right) = \sum_{r=1}^{R} \log \left[\sum_{\boldsymbol{q}} p\left(\boldsymbol{O}^{(r)}, \boldsymbol{q} \mid \theta\right)\right]$$

其中 $\boldsymbol{q}$ 是状态序列。根据 Jensen 不等式可知

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{r=1}^{R} \log \left[\sum_{\boldsymbol{q}} P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right) \frac{p\left(\boldsymbol{O}^{(r)}, \boldsymbol{q} \mid \theta\right)}{P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right)}\right] \\
&\geq \sum_{r=1}^{R} \sum_{\boldsymbol{q}} P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right) \log \frac{p\left(\boldsymbol{O}^{(r)}, \boldsymbol{q} \mid \theta\right)}{P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right)} \\
&= \sum_{r=1}^{R} \sum_{\boldsymbol{q}} P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right) \log p\left(\boldsymbol{O}^{(r)}, \boldsymbol{q} \mid \theta\right) + H\left[P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right)\right]
\end{aligned}$$

其中 $H\left[P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right)\right]$ 是给定最优参数的熵，是一个常数。因此，我们将辅助函数定义为

$$Q(\theta, \hat{\theta}) = \sum_{r=1}^{R} \sum_{\boldsymbol{q}} P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right) \log p\left(\boldsymbol{O}^{(r)}, \boldsymbol{q} \mid \theta\right)$$

上述辅助函数给出了对数似然函数的一个下界。注意到

$$\sum_{\boldsymbol{q}} P\left(\boldsymbol{q} \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right) = \sum_{j=1}^{N} P\left(q_t = j \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} P\left(q_{t-1} = i, q_t = j \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right)$$

为了书写简便，我们将软分配的占用率记为

$$\begin{aligned}
\gamma_j(t) &= P\left(q_t = j \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right) \\
\gamma_{(i,j)}(t) &= P\left(q_{t-1} = i, q_t = j \mid \boldsymbol{O}^{(r)}, \hat{\theta}\right)
\end{aligned}$$

因此，辅助函数可以重写为

$$\begin{aligned}
Q(\theta,\hat{\theta}) &= \sum_{r=1}^{R}\sum_{\boldsymbol{q}} P\left(\boldsymbol{q}|\boldsymbol{O}^{(r)},\hat{\theta}\right)\left[\sum_{t=1}^{T}\log P(q_t|q_{t-1},\theta)+\sum_{t=1}^{T}\log p(\boldsymbol{o}_t|q_t,\theta)\right] \\
&= \sum_{r=1}^{R}\sum_{\boldsymbol{q}}\left[\sum_{t=1}^{T} P\left(\boldsymbol{q}|\boldsymbol{O}^{(r)},\hat{\theta}\right)\log P(q_t|q_{t-1},\theta)+\sum_{t=1}^{T} P\left(\boldsymbol{q}|\boldsymbol{O}^{(r)},\hat{\theta}\right)\log p(\boldsymbol{o}_t|q_t,\theta)\right] \\
&= \sum_{r=1}^{R}\left[\sum_{t=1}^{T}\sum_{\boldsymbol{q}} P\left(\boldsymbol{q}|\boldsymbol{O}^{(r)},\hat{\theta}\right)\log P(q_t|q_{t-1},\theta)+\sum_{t=1}^{T}\sum_{\boldsymbol{q}} P\left(\boldsymbol{q}|\boldsymbol{O}^{(r)},\hat{\theta}\right)\log p(\boldsymbol{o}_t|q_t,\theta)\right] \\
&= \sum_{r=1}^{R}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_{(i,j)}(t)\log a_{ij}+\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_j(t)\log b_j(\boldsymbol{o}_t)\right] \\
&= \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_{(i,j)}(t)\log a_{ij}+\sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_j(t)\log b_j(\boldsymbol{o}_t)
\end{aligned}$$

于是我们将辅助函数分为两部分，分别进行优化

$$\begin{aligned}
Q_A(\theta,\hat{\theta}) &= \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_{(i,j)}(t)\log a_{ij} \\
Q_B(\theta,\hat{\theta}) &= \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_j(t)\log b_j(\boldsymbol{o}_t)
\end{aligned}$$

对于 $Q_A(\theta,\hat{\theta})$，我们有如下优化问题

$$\begin{aligned}
\max_{a_{ij}} \quad & Q_A(\theta,\hat{\theta}) \\
\text{s.t.} \quad & \sum_{j=1}^{N} a_{ij} = 1
\end{aligned}$$

其拉格朗日函数为

$$\mathcal{L}_A(a_{ij},\lambda)=\sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_{(i,j)}(t)\log a_{ij}+\lambda\left(1-\sum_{j=1}^{N}a_{ij}\right)$$

根据拉格朗日条件有

$$\frac{\partial\mathcal{L}_A}{\partial a_{ij}}=\sum_{r=1}^{R}\sum_{t=1}^{T}\frac{\gamma_{(i,j)}(t)}{a_{ij}}-\lambda=0$$

解得

$$a_{ij}=\frac{1}{\lambda}\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{(i,j)}(t)$$

因此，优化问题的最优解为

$$\lambda^* = \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_{(i,j)}(t)$$

$$a_{ij}^* = \frac{\sum\limits_{r=1}^{R}\sum\limits_{t=1}^{T}\gamma_{(i,j)}(t)}{\sum\limits_{r=1}^{R}\sum\limits_{t=1}^{T}\sum\limits_{j=1}^{N}\gamma_{(i,j)}(t)}$$

对于 $Q_B(\theta,\hat{\theta})$，我们假定 $b_j(\boldsymbol{o}_t) = \sum\limits_{m=1}^{M} c_{jm}\mathcal{N}(\boldsymbol{o}_t|\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_{jm})$，即

$$b_j(\boldsymbol{o}_t) = \sum_{m=1}^{M}\frac{c_{jm}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_{jm}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})^T\boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})\right]$$

其中 D 是观测向量的维度，M 是高斯分量的个数。$\boldsymbol{\mu}_{jm}$ 和 $\boldsymbol{\Sigma}_{jm}$ 分别是第 j 个状态的第 m 个高斯分量的均值向量和协方差矩阵。由 Jensen 不等式可知

$$\begin{aligned}
Q_B(\theta,\hat{\theta}) &= \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_j(t)\log\sum_{m=1}^{M}\frac{c_{jm}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_{jm}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})^T\boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})\right]\\
&\geq \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_j(t)\sum_{m=1}^{M}\left\{\log c_{jm}-\frac{1}{2}\left[\log|\boldsymbol{\Sigma}_{jm}|+(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})^T\boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})\right]+k\right\}\\
&= k+\sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{m=1}^{M}\gamma_{jm}(t)\left\{\log c_{jm}-\frac{1}{2}\left[\log|\boldsymbol{\Sigma}_{jm}|+(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})^T\boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})\right]\right\}
\end{aligned}$$

其中 $\gamma_{jm}(t) = p\left(q_t=j,g_t=m|\boldsymbol{O}^{(r)},\hat{\theta}\right)$，而 k 是一个常数。因此，我们定义新的辅助函数为

$$Q'_B(\theta,\hat{\theta}) = \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{m=1}^{M}\gamma_{jm}(t)\left\{\log c_{jm}+\frac{1}{2}\left[\log|\boldsymbol{\Sigma}_{jm}^{-1}|-(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})^T\boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})\right]\right\}$$

上述辅助函数给出了 $Q_B(\theta,\hat{\theta})$ 的一个下界。对于 $Q'_B(\theta,\hat{\theta})$，我们有如下优化问题

$$\begin{aligned}
\max_{c_{jm},\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_{jm}}\ & Q'_B(\theta,\hat{\theta})\\
\text{s.t.}\quad & \sum_{m=1}^{M}c_{jm}=1
\end{aligned}$$

其拉格朗日函数为

$$\mathcal{L}_B(c_{jm},\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_{jm},\lambda) = Q'_B(\theta,\hat{\theta})+\lambda\left(1-\sum_{m=1}^{M}c_{jm}\right)$$

根据拉格朗日条件有

$$\frac{\partial\mathcal{L}_B}{\partial c_{jm}} = \sum_{r=1}^{R}\sum_{t=1}^{T}\frac{\gamma_{jm}(t)}{c_{jm}}-\lambda=0$$

解得

$$c_{jm} = \frac{1}{\lambda}\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{jm}(t)$$

因此，λ 和 c_{jm} 的最优解为

$$\lambda^* = \sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{jm}(t)$$
$$c_{jm}^* = \frac{1}{\lambda^*}\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{jm}(t)$$

根据拉格朗日条件有

$$\frac{\partial \mathcal{L}_B}{\partial \boldsymbol{\mu}_{jm}} = \sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{jm}(t)\boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm}) = 0$$
$$\frac{\partial \mathcal{L}_B}{\partial \boldsymbol{\Sigma}_{jm}^{-1}} = \sum_{r=1}^{R}\sum_{t=1}^{T}\frac{1}{2}\gamma_{jm}(t)\left[\boldsymbol{\Sigma}_{jm} - (\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})^T\right] = 0$$

因此，$\boldsymbol{\mu}_{jm}$ 和 $\boldsymbol{\Sigma}_{jm}$ 的最优解为

$$\boldsymbol{\mu}_{jm}^* = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{jm}(t)\boldsymbol{o}_t}{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{jm}(t)}$$
$$\boldsymbol{\Sigma}_{jm}^* = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{jm}(t)(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm}^*)(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm}^*)^T}{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{jm}(t)}$$

将前向概率定义为

$$\begin{aligned}
\alpha_j(t) &= p\left(\boldsymbol{O}_1^t, q_t = j\right) \\
&= \sum_{i=1}^{N} p\left(\boldsymbol{O}_1^{t-1}, \boldsymbol{o}_t, q_{t-1} = i, q_t = j\right) \\
&= \sum_{i=1}^{N} p\left(\boldsymbol{o}_t, q_t = j | \boldsymbol{O}_1^{t-1}, q_{t-1} = i\right) p\left(\boldsymbol{O}_1^{t-1}, q_{t-1} = i\right) \\
&= \sum_{i=1}^{N} p\left(\boldsymbol{o}_t | q_t = j\right) P\left(q_t = j | q_{t-1} = i\right) \alpha_i(t-1) \\
&= \sum_{i=1}^{N} b_j(\boldsymbol{o}_t) a_{ij} \alpha_i(t-1)
\end{aligned}$$

将后向概率定义为

$$\begin{aligned}
\beta_j(t) &= p\left(\boldsymbol{O}_{t+1}^T|q_t=j\right) \\
&= \sum_{i=1}^N p\left(\boldsymbol{o}_{t+1}, \boldsymbol{O}_{t+2}^T, q_{t+1}=i|q_t=j\right) \\
&= \sum_{i=1}^N p\left(\boldsymbol{o}_{t+1}, \boldsymbol{O}_{t+2}^T|q_{t+1}=i, q_t=j\right) P\left(q_{t+1}=i|q_t=j\right) \\
&= \sum_{i=1}^N p\left(\boldsymbol{o}_{t+1}|q_{t+1}=i\right) P\left(q_{t+1}=i|q_t=j\right) p\left(\boldsymbol{O}_{t+2}^T|q_{t+1}=i\right) \\
&= \sum_{i=1}^N b_i(\boldsymbol{o}_{t+1}) a_{ji} \beta_i(t+1)
\end{aligned}$$

因此前向概率和后向概率可以递归计算。于是可以将软分配的占用率重写为

$$\begin{aligned}
\gamma_j(t) &= P\left(q_t=j|\boldsymbol{O}_1^T, \hat{\theta}\right) \\
&= \frac{p\left(\boldsymbol{O}_1^T, q_t=j|\hat{\theta}\right)}{p\left(\boldsymbol{O}_1^T|\hat{\theta}\right)} \\
&= \frac{p\left(\boldsymbol{O}_1^t, \boldsymbol{O}_{t+1}^T, q_t=j|\hat{\theta}\right)}{p\left(\boldsymbol{O}_1^T|\hat{\theta}\right)} \\
&= \frac{p\left(\boldsymbol{O}_1^t, q_t=j|\hat{\theta}\right) p\left(\boldsymbol{O}_{t+1}^T|q_t=j, \hat{\theta}\right)}{p\left(\boldsymbol{O}_1^T|\hat{\theta}\right)} \\
&= \frac{\alpha_j(t)\beta_j(t)}{\alpha_N(T+1)}
\end{aligned}$$

$$\begin{aligned}
\gamma_{(i,j)}(t) &= P\left(q_{t-1}=i, q_t=j|\boldsymbol{O}_1^T, \hat{\theta}\right) \\
&= \frac{p\left(\boldsymbol{O}_1^T, q_{t-1}=i, q_t=j|\hat{\theta}\right)}{p\left(\boldsymbol{O}_1^T|\hat{\theta}\right)} \\
&= \frac{p\left(\boldsymbol{O}_1^{t-1}, \boldsymbol{O}_t, q_{t-1}=i, q_t=j|\hat{\theta}\right)}{p\left(\boldsymbol{O}_1^T|\hat{\theta}\right)} \\
&= \frac{p\left(\boldsymbol{O}_1^{t-1}, q_{t-1}=i|\hat{\theta}\right) p\left(\boldsymbol{O}_{t+1}^T|q_t=j, \hat{\theta}\right) P\left(q_t=j|q_{t-1}=i\right) p\left(\boldsymbol{o}_t|q_t=j\right)}{p\left(\boldsymbol{O}_1^T|\hat{\theta}\right)} \\
&= \frac{\alpha_i(t-1)\beta_j(t) a_{ij} b_j(\boldsymbol{o}_t)}{\alpha_N(T+1)}
\end{aligned}$$

$$\begin{aligned}
\gamma_{jm}(t) &= P\left(q_t=j, g_t=m|\boldsymbol{O}_1^T, \hat{\theta}\right) \\
&= P\left(q_t=j|\boldsymbol{O}_1^T, \hat{\theta}\right) p\left(g_t=m|\boldsymbol{O}_1^T, \hat{\theta}\right) \\
&= \gamma_j(t)\gamma_m(t)
\end{aligned}$$

其中 $\gamma_m(t)$ 可以写成

$$\gamma_m(t) = \frac{b_{jm}(\boldsymbol{o}_t)}{b_j(\boldsymbol{o}_t)} = \frac{c_{jm}}{b_j(\boldsymbol{o}_t)} \cdot \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_{jm}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})^T \boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})\right]$$

至此，我们已经推导了 Baum-Welch 算法中的所有更新公式。下面，我们以伪代码的形式总结 Baum-Welch 算法。

算法 1 Baum-Welch 算法

输入： 观测序列 $\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \ldots, \boldsymbol{O}^{(R)}$

输出： 估计参数 $\hat{\theta} = (a_{ij}, c_{jm}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$

初始化 $\hat{\theta}^{(0)} = \left(a_{ij}^{(0)}, c_{jm}^{(0)}, \boldsymbol{\mu}_{jm}^{(0)}, \boldsymbol{\Sigma}_{jm}^{(0)}\right)$

for $k \leftarrow 1$ to K **do**

$\alpha_j^{(k)}(t) \leftarrow \sum_{i=1}^{N} b_j^{(k-1)}(\boldsymbol{o}_t) a_{ij}^{(k-1)} \alpha_i^{(k)}(t-1)$ ▷ 前向概率

$\beta_j^{(k)}(t) \leftarrow \sum_{i=1}^{N} b_j^{(k-1)}(\boldsymbol{o}_{t+1}) a_{ji}^{(k-1)} \beta_i^{(k)}(t+1)$ ▷ 后向概率

$\gamma_j^{(k)}(t) \leftarrow \dfrac{\alpha_j^{(k)}(t)\beta_j^{(k)}(t)}{\alpha_N^{(k)}(T+1)}$

$\gamma_{(i,j)}^{(k)}(t) \leftarrow \dfrac{\alpha_i^{(k)}(t-1)\beta_j^{(k)}(t) a_{ij}^{(k-1)} b_j^{(k-1)}(\boldsymbol{o}_t)}{\alpha_N^{(k)}(T+1)}$

$\gamma_{jm}^{(k)}(t) \leftarrow \gamma_j^{(k)}(t)\gamma_m^{(k)}(t)$ ▷ 软分配占用率

$a_{ij}^{(k)} \leftarrow \dfrac{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_{(i,j)}^{(k)}(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{j=1}^{N} \gamma_{(i,j)}^{(k)}(t)}$

$c_{jm}^{(k)} \leftarrow \dfrac{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_{jm}^{(k)}(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{m=1}^{M} \gamma_{jm}^{(k)}(t)}$

$\boldsymbol{\mu}_{jm}^{(k)} \leftarrow \dfrac{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_{jm}^{(k)}(t)\boldsymbol{o}_t}{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_{jm}^{(k)}(t)}$

$\boldsymbol{\Sigma}_{jm}^{(k)} \leftarrow \dfrac{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_{jm}^{(k)}(t)(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm}^{(k)})(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm}^{(k)})^T}{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_{jm}^{(k)}(t)}$ ▷ 模型参数

end for

return $\hat{\theta}^{(K)} = \left(a_{ij}^{(K)}, c_{jm}^{(K)}, \boldsymbol{\mu}_{jm}^{(K)}, \boldsymbol{\Sigma}_{jm}^{(K)}\right)$