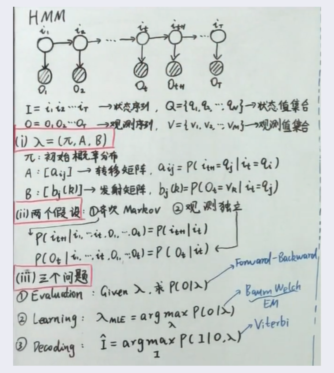
HMM

HMM模型的三个基本问题



- **评估观察序列概率**：即给定模型λ=(A,B,pi)和观测序列O={o1,o2,...oT}，计算在模型λ下观测序列O出现的概率P(O|λ)。这个问题的求解需要用到前向后向算法
- **模型参数学习问题**：即给定观测序列O={o1,o2,...oT}，估计模型λ=(A,B,pi)的参数，使该模型下观测序列的条件概率P(O|λ)最大。这个问题的求解需要用到基于EM算法的鲍姆-韦尔奇算法
- **预测问题，也称为解码问题**：即给定模型λ=(A,B, Π)和观测序列O={o1,o2,...oT}，求给定观测序列条件下，最可能出现的对应的状态序列，这个问题的求解需要用到基于动态规划的维特比算法

HMM模型的定义

- 对于HMM模型，首先我们假设Q是所有可能的隐藏状态的集合，V是所有可能的观测状态的集合，其中，N是可能的隐藏状态数，M是所有的可能的观察状态数。

$$Q=\{q_1,q_2,\dots,q_N\},\ V=\{v_1,v_2,\dots v_M\}$$

- 对于一个长度为T的序列，I对应的状态序列，O是对应的观察序列，其中，任意一个隐藏状态it∈Q,任意一个观察状态ot∈V

$$I=\{i_1,i_2,\dots,i_T\},\ O=\{o_1,o_2,\dots o_T\}$$

- **两个基本假设**
  - 齐次马尔科夫链假设。即任意时刻的隐藏状态只依赖于它前一个隐藏状态，如果在时刻t的隐藏状态是it=qi,在时刻t+1的隐藏状态是it+1=qj，则从时刻t到时刻t+1的HMM状态转移概率aij可以表示为

$$a_{ij}=P(i_{t+1}=q_j|i_t=q_i)$$

    - 这样aij可以组成马尔科夫链的状态转移矩阵A:

$$A=\Big[a_{ij}\Big]_{N\times N}$$

  - 观测独立性假设。即任意时刻的观察状态只仅仅依赖于当前时刻的隐藏状态，这也是一个为了简化模型的假设。

$$b_j(k)=P(o_t=v_k|i_t=q_j)$$

    - 这样bj(k)可以组成观测状态生成的概率矩阵B

$$B=\Big[b_j(k)\Big]_{N\times M}$$

- 除此之外，我们需要一组在时刻t=1的隐藏状态概率分布pi:

$$\Pi=\Big[\pi(i)\Big]_N \text{其中 } \pi(i)=P(i_1=q_i)$$

- 一个HMM模型，可以由隐藏状态初始概率分布pi，状态转移概率矩阵A和观测状态概率矩阵B决定。pi，A决定状态序列，B决定观测序列。因此，HMM模型可以由一个三元组λ表示如下：

$$\lambda=(A,B,\Pi)$$

评估观察序列概率

即给定模型λ=(A,B,pi)和观测序列O={o1,o2,...oT}，计算在模型λ下观测序列O出现的概率P(O|λ)，这个问题的求解需要用到前向后向算法

问题描述

这个问题是这样的。我们已知HMM模型的参数λ=(A,B,pi)。其中A是隐藏状态转移概率的矩阵，B是观测状态生成概率的矩阵， pi是隐藏状态的初始概率分布。同时我们也已经得到了观测序列O={o1,o2,...oT},现在我们要求观测序列O在模型λ下出现的条件概率P(O|λ)。

直接计算

- 任意一个隐藏序列I={i1,i2,...,iT}出现的概率

$$P(I|\lambda)=\pi_{i_1}a_{i_1i_2}a_{i_2i_3}\dots a_{i_{T-1}\ i_T}$$

- 对于固定的状态序列I={i1,i2,...,iT}，我们要求的观察序列O={o1,o2,...oT}出现的概率是

$$P(O|I,\lambda)=b_{i_1}(o_1)b_{i_2}(o_2)\dots b_{i_T}(o_T)$$

  - 则O和I联合出现的概率是：

$$P(O,I|\lambda)=P(I|\lambda)P(O|I,\lambda)=\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\dots a_{i_{T-1}\ i_T}b_{i_T}(o_T)$$

    - 然后求边缘概率分布，即可得到观测序列O在模型λ下出现的条件概率P(O|λ)：

$$P(O|\lambda)=\sum_I P(O,I|\lambda)=\sum_{i_1,i_2,\dots i_T}\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\dots a_{i_{T-1}\ i_T}b_{i_T}(o_T)$$

- 如果隐藏状态数N非常多的那就麻烦了，此时预测状态有N^T种组合，算法的时间复杂度是O(TN^T)阶的

向前算法

输入：HMM模型$\lambda=(A,B,\Pi)$，观测序列$O=(o_1,o_2,\dots o_T)$

输出：观测序列概率$P(O|\lambda)$

1) 计算时刻1的各个隐藏状态前向概率：

$$\alpha_1(i)=\pi_i b_i(o_1),\ i=1,2,\dots N$$

2) 递推时刻$2,3,\dots T$时刻的前向概率：

$$\alpha_{t+1}(i)=\Big[\sum_{j=1}^N\alpha_t(j)a_{ji}\Big]b_i(o_{t+1}),\ i=1,2,\dots N$$

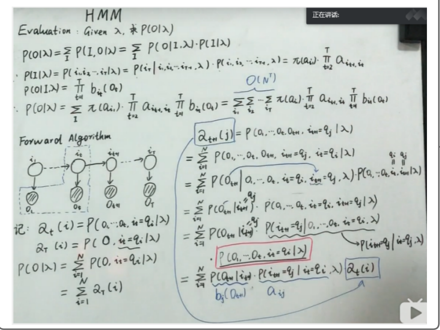
3) 计算最终结果：

$$P(O|\lambda)=\sum_{i=1}^N\alpha_T(i)$$

- 前向算法本质上属于动态规划的算法，也就是我们要通过找到局部状态递推的公式，这样一步步的从子问题的最优解拓展到整个问题的最优解。
- 定义时刻t时隐藏状态为qi，观测状态的序列为o1,o2,...ot的概率为前向概率。记为：

$$\alpha_t(i)=P(o_1,o_2,\dots o_t,i_t=q_i|\lambda)$$

  - 推导过程



    - 这样我们得到了前向概率的递推关系式如下

$$\alpha_{t+1}(i)=\Big[\sum_{j=1}^N\alpha_t(j)a_{ji}\Big]b_i(o_{t+1})$$

- 从递推公式可以看出，我们的算法时间复杂度是O(TN^2)，比暴力解法的时间复杂度O(TN^T)少了几个数量级。

向后算法

输入：HMM模型$\lambda=(A,B,\Pi)$，观测序列$O=(o_1,o_2,\dots o_T)$

输出：观测序列概率$P(O|\lambda)$

1) 初始化时刻T的各个隐藏状态后向概率：

$$\beta_T(i)=1,\ i=1,2,\dots N$$

2) 递推时刻$T-1,T-2,\dots 1$时刻的后向概率：

$$\beta_t(i)=\sum_{j=1}^N a_{ij}b_j(o_{t+1})\beta_{t+1}(j),\ i=1,2,\dots N$$

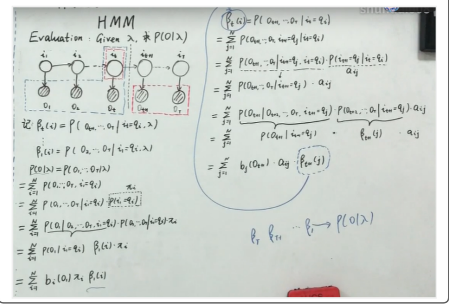
3) 计算最终结果：

$$P(O|\lambda)=\sum_{i=1}^N\pi_i b_i(o_1)\beta_1(i)$$

- 定义时刻t时隐藏状态为qi，从时刻t+1到最后时刻T的观测状态的序列为ot+1,ot+2,...oT的概率为后向概率。记为

$$\beta_t(i)=P(o_{t+1},o_{t+2},\dots o_T|i_t=q_i,\lambda)$$

  - 推导过程



    - 这样我们得到了后向概率的递推关系式如下：

$$\beta_t(i)=\sum_{j=1}^N a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$

- 此时我们的算法时间复杂度仍然是O(TN^2)。

HMM常用概率的计算

- 利用前向概率和后向概率，我们可以计算出HMM中单个状态和两个状态的概率公式。

$$\gamma_t(i)=P(i_t=q_i|O,\lambda)=\frac{P(i_t=q_i,O|\lambda)}{P(O|\lambda)}$$

- 利用前向概率和后向概率的定义可知：

$$P(i_t=q_i,O|\lambda)=\alpha_t(i)\beta_t(i)$$

  - 于是我们得到

$$\gamma_t(i)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^N\alpha_t(j)\beta_t(j)}$$

- 给定模型λ和观测序列O,在时刻t处于状态qi，且时刻t+1处于状态qj的概率记为:

$$\xi_t(i,j)=P(i_t=q_i,i_{t+1}=q_j|O,\lambda)=\frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{P(O|\lambda)}$$

  - 从而最终我们得到

$$\xi_t(i,j)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{r=1}^N\sum_{s=1}^N\alpha_t(r)a_{rs}b_s(o_{t+1})\beta_{t+1}(s)}$$

- 在观测序列O下状态i出现的期望值$\sum_{t=1}^T\gamma_t(i)$
  - 在观测序列O下由状态i转移的期望值$\sum_{t=1}^{T-1}\gamma_t(i)$
    - 在观测序列O下由状态i转移到状态j的期望值$\sum_{t=1}^{T-1}\xi_t(i,j)$

模型参数学习问题

即给定观测序列O={o1,o2,...oT}，估计模型λ=(A,B,pi)的参数，使该模型下观测序列的条件概率P(O|λ)最大。这个问题的求解需要用到基于EM算法的鲍姆-韦尔奇算法

问题重述

- 第一种情况较为简单，就是我们已知D个长度为T的观测序列和对应的隐藏状态序列，即{(O1,I1),(O2,I2),...(OD,ID)}是已知的，此时我们可以很容易的用最大似然来求解模型参数。
- 假设样本从隐藏状态qi转移到qj的频率计数是Aij,那么状态转移矩阵求得为 $A=\Big[a_{ij}\Big]$, 其中$a_{ij}=\frac{A_{ij}}{\sum_{s=1}^N A_{is}}$
  - 假设样本隐藏状态为qj且观测状态为vk的频率计数是Bjk,那么观测状态概率矩阵为 $B=\Big[b_j(k)\Big]$, 其中$b_j(k)=\frac{B_{jk}}{\sum_{s=1}^M B_{js}}$
    - 假设所有样本中初始隐藏状态为qi的频率计数为C(i),那么初始概率分布为 $\Pi=\pi(i)=\frac{C(i)}{\sum_{s=1}^N C(s)}$
- 可见第一种情况下求解模型还是很简单的。但是在很多时候，我们无法得到HMM样本观察序列对应的隐藏序列，只有D个长度为T的观测序列，即{(O1),(O2),...(OD)}是已知的，它的解法最常用的是鲍姆-韦尔奇算法，其实就是基于EM算法的求解

鲍姆-韦尔奇算法原理

- 鲍姆-韦尔奇算法原理既然使用的就是EM算法的原理，那么我们需要在E步求出联合分布P(O,I|λ)基于条件概率P(I|O,λ~)的期望，其中λ~为当前的模型参数，然后再M步最大化这个期望，得到更新的模型参数λ。接着不停的进行EM迭代，直到模型参数的值收敛为止。
- 首先来看看E步，当前模型参数为λ~，联合分布P(O,I|λ)基于条件概率P(I|O,λ~)的期望表达式为

$$L(\lambda,\overline{\lambda})=\sum_I P(I|O,\overline{\lambda})logP(O,I|\lambda)$$

- 在M步，我们极大化上式，然后得到更新后的模型参数如下

$$\overline{\lambda}=arg\ \max_{\lambda}\sum_I P(I|O,\overline{\lambda})logP(O,I|\lambda)$$

- 通过不断的E步和M步的迭代，直到λ~收敛

鲍姆-韦尔奇算法的推导

输入：D个观测序列样本$\{(O_1),(O_2),\dots(O_D)\}$

输出：HMM模型参数

1)随机初始化所有的$\pi_i,a_{ij},b_j(k)$

2) 对于每个样本$d=1,2,\dots D$，用前向后向算法计算$\gamma_t^{(d)}(i)$，$\xi_t^{(d)}(i,j),t=1,2\dots T$

3) 更新模型参数：

$$\pi_i=\frac{\sum_{d=1}^D\gamma_1^{(d)}(i)}{D}$$

$$a_{ij}=\frac{\sum_{d=1}^D\sum_{t=1}^{T-1}\xi_t^{(d)}(i,j)}{\sum_{d=1}^D\sum_{t=1}^{T-1}\gamma_t^{(d)}(i)}$$

$$b_j(k)=\frac{\sum_{d=1}^D\sum_{t=1,o_t^{(d)}=v_k}^T\gamma_t^{(d)}(j)}{\sum_{d=1}^D\sum_{t=1}^T\gamma_t^{(d)}(j)}$$

4) 如果$\pi_i,a_{ij},b_j(k)$的值已经收敛，则算法结束，否则回到第2）步继续迭代。

问题重述

- 训练数据 $\{(O_1,I_1),(O_2,I_2),\dots(O_D,I_D)\}$
  - 观测序列 $O_d=\{o_1^{(d)},o_2^{(d)},\dots o_T^{(d)}\}$
    - 隐藏状态序列 $I_d=\{i_1^{(d)},i_2^{(d)},\dots i_T^{(d)}\}$
- 我们需要先计算联合分布P(O,I|λ)的表达式如下：

$$P(O,I|\lambda)=\prod_{d=1}^D\pi_{i_1^{(d)}}b_{i_1^{(d)}}(o_1^{(d)})a_{i_1^{(d)}i_2^{(d)}}b_{i_2^{(d)}}(o_2^{(d)})\dots a_{i_{T-1}^{(d)}i_T^{(d)}}b_{i_T^{(d)}}(o_T^{(d)})$$

  - 我们的E步得到的期望表达式为

$$L(\lambda,\overline{\lambda})=\sum_I P(I|O,\overline{\lambda})logP(O,I|\lambda)$$

    - 在M步我们要极大化上式。由于P(I|O,λ~)=P(I,O|λ~)/P(O|λ~),而P(O|λ~)是常数，因此我们要极大化的式子等价于：

$$\overline{\lambda}=arg\ \max_{\lambda}\sum_I P(O,I|\overline{\lambda})logP(O,I|\lambda)$$

Pi迭代推导

- 将上面P(O,I|λ)的表达式带入我们的极大化式子，得到的表达式如下：

$$\overline{\lambda}=arg\ \max_{\lambda}\sum_{d=1}^D\sum_I P(O,I|\overline{\lambda})(log\pi_{i_1}+\sum_{t=1}^{T-1}log\ a_{i_t,i_{t+1}}+\sum_{t=1}^T logb_{i_t}(o_t))$$

  - 我们的隐藏模型参数λ=(A,B,pi),因此下面我们只需要对上式分别对A,B,pi求导即可得到我们更新的模型参数λ~
    - 首先我们看看对模型参数pi的求导。由于pi只在上式中括号里的第一部分出现，因此我们对于pi的极大化式子为:

$$\overline{\pi_i}=arg\ \max_{\pi_{i_1}}\sum_{d=1}^D\sum_I P(O,I|\overline{\lambda})log\pi_{i_1}=arg\ \max_{\pi_i}\sum_{d=1}^D\sum_{i=1}^N P(O,i_1^{(d)}=i|\overline{\lambda})log\pi_i$$

- 因此根据拉格朗日子乘法，我们得到pi要极大化的拉格朗日函数为:

$$arg\ \max_{\pi_i}\sum_{d=1}^D\sum_{i=1}^N P(O,i_1^{(d)}=i|\overline{\lambda})log\pi_i+\gamma(\sum_{i=1}^N\pi_i-1)$$

  - 其中，γ为拉格朗日系数。上式对pi求偏导数并令结果为0， 我们得到

$$\sum_{d=1}^D P(O,i_1^{(d)}=i|\overline{\lambda})+\gamma\pi_i=0$$

- 令i分别等于从1到N，从上式可以得到N个式子，对这N个式子求和可得： $\sum_{d=1}^D P(O|\overline{\lambda})+\gamma=0$
  - 从上两式消去γ,得到pi的表达式为

$$\pi_i=\frac{\sum_{d=1}^D P(O,i_1^{(d)}=i|\overline{\lambda})}{\sum_{d=1}^D P(O|\overline{\lambda})}=\frac{\sum_{d=1}^D P(O,i_1^{(d)}=i|\overline{\lambda})}{DP(O|\overline{\lambda})}=\frac{\sum_{d=1}^D P(i_1^{(d)}=i|O,\overline{\lambda})}{D}=\frac{\sum_{d=1}^D P(i_1^{(d)}=i|O^{(d)},\overline{\lambda})}{D}$$

- 第二节中前向概率的定义可得

$$P(i_1^{(d)}=i|O^{(d)},\overline{\lambda})=\gamma_1^{(d)}(i)$$

  - 因此最终我们在M步pi的迭代公式为 $\pi_i=\frac{\sum_{d=1}^D\gamma_1^{(d)}(i)}{D}$

迭代推导

- 现在我们来看看A的迭代公式求法。方法和pi的类似。由于A只在最大化函数式中括号里的第二部分出现，而这部分式子可以整理为：

$$\sum_{d=1}^D\sum_I\sum_{t=1}^{T-1}P(O,I|\overline{\lambda})log\ a_{i_t,i_{t+1}}=\sum_{d=1}^D\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^{T-1}P(O,i_t^{(d)}=i,i_{t+1}^{(d)}=j|\overline{\lambda})log\ a_{ij}$$

  - 我们可以用拉格朗日子乘法并对aij求导，并令结果为0，可以得到aij的迭代表达式为：

$$a_{ij}=\frac{\sum_{d=1}^D\sum_{t=1}^{T-1}P(O^{(d)},i_t^{(d)}=i,i_{t+1}^{(d)}=j|\overline{\lambda})}{\sum_{d=1}^D\sum_{t=1}^{T-1}P(O^{(d)},i_t^{(d)}=i|\overline{\lambda})}$$

    - 前向后向算法评估观察序列概率里第二节中前向概率的定义和第五节ξt(i,j)的定义可得们在M步aij的迭代公式为：

$$a_{ij}=\frac{\sum_{d=1}^D\sum_{t=1}^{T-1}\xi_t^{(d)}(i,j)}{\sum_{d=1}^D\sum_{t=1}^{T-1}\gamma_t^{(d)}(i)}$$

B迭代推导

- 现在我们来看看B的迭代公式求法。方法和pi的类似。由于B只在最大化函数式中括号里的第三部分出现，而这部分式子可以整理为：

$$\sum_{d=1}^D\sum_I\sum_{t=1}^T P(O,I|\overline{\lambda})log\ b_{i_t}(o_t)=\sum_{d=1}^D\sum_{j=1}^N\sum_{t=1}^T P(O,i_t^{(d)}=j|\overline{\lambda})log\ b_j(o_t)$$

  - 由于bj(ot)还满足∑k=1Mbj(ot=vk)=1。和求解pi类似，我们可以用拉格朗日子乘法并对bj(k)求导，并令结果为0，得到bj(k)的迭代表达式为：

$$b_j(k)=\frac{\sum_{d=1}^D\sum_{t=1}^T P(O,i_t^{(d)}=j|\overline{\lambda})I(o_t^{(d)}=v_k)}{\sum_{d=1}^D\sum_{t=1}^T P(O,i_t^{(d)}=j|\overline{\lambda})}$$

    - 其中I(o(d)t=vk)当且仅当o(d)t=vk时为1，否则为0.利用隐马尔科夫模型HMM（二）前向后向算法评估观察序列概率里第二节中前向概率的定义可得bj(ot)的最终表达式为：

$$b_j(k)=\frac{\sum_{d=1}^D\sum_{t=1,o_t^{(d)}=v_k}^T\gamma_t^{(d)}(j)}{\sum_{d=1}^D\sum_{t=1}^T\gamma_t^{(d)}(j)}$$

预测问题，也称为解码问题

即给定模型λ=(A,B, Π)和观测序列O={o1,o2,...oT}，求给定观测序列条件下，最可能出现的对应的状态序列，这个问题的求解需要用到基于动态规划的维特比算法

输入：HMM模型$\lambda=(A,B,\Pi)$，观测序列$O=(o_1,o_2,\dots o_T)$

输出：最有可能的隐藏状态序列$I^*=\{i_1^*,i_2^*,\dots i_T^*\}$

1）初始化局部状态：

$$\delta_1(i)=\pi_i b_i(o_1),\ i=1,2\dots N$$

$$\Psi_1(i)=0,\ i=1,2\dots N$$

2) 进行动态规划递推时刻$t=2,3,\dots T$时刻的局部状态：

$$\delta_t(i)=\max_{1\le j\le N}[\delta_{t-1}(j)a_{ji}]b_i(0_t),\ i=1,2\dots N$$

$$\Psi_t(i)=arg\ \max_{1\le j\le N}[\delta_{t-1}(j)a_{ji}],\ i=1,2\dots N$$

3) 计算时刻T最大的$\delta_T(i)$,即为最可能隐藏状态序列出现的概率。计算时刻T最大的$\Psi_t(i)$,即为时刻T最可能的隐藏状态。

$$P*=\max_{1\le j\le N}\delta_T(i)$$

$$i_T^*=arg\ \max_{1\le j\le N}[\delta_T(i)]$$

4) 利用局部状态$\Psi(i)$开始回溯。对于$t=T-1,T-2,\dots,1$:

$$i_t^*=\Psi_{t+1}(i_{t+1}^*)$$

- 维特比算法是一个通用的解码算法，是基于动态规划的求序列最短路径的方法。在文本挖掘的分词原理中我们已经讲到了维特比算法的一些细节。既然是动态规划算法，那么就需要找到合适的局部状态，以及局部状态的递推公式。在HMM中，维特比算法定义了两个局部状态用于递推。
- 第一个局部状态是在时刻t隐藏状态为i所有可能的状态转移路径i1,i2,...it中的概率最大值。记为δt(i)

$$\delta_t(i)=\max_{i_1,i_2,\dots i_{t-1}}P(i_t=i,i_1,i_2,\dots i_{t-1},o_t,o_{t-1},\dots o_1|\lambda),\ i=1,2,\dots N$$

  - 递推得到

$$\delta_{t+1}(i)=\max_{i_1,i_2,\dots i_t}P(i_{t+1}=i,i_1,i_2,\dots i_t,o_{t+1},o_t,\dots o_1|\lambda)=\max_{1\le j\le N}[\delta_t(j)a_{ji}]b_i(o_{t+1})$$

- 第二个局部状态由第一个局部状态递推得到。我们定义在时刻t隐藏状态为i的所有单个状态转移路径(i1,i2,...,it□1,i)中概率最大的转移路径中第t□1个节点的隐藏状态为Ψt(i),其递推表达式可以表示为

$$\Psi_t(i)=arg\ \max_{1\le j\le N}[\delta_{t-1}(j)a_{ji}]$$