1 EM算法简单回顾

已知观察变量x 求参数

$$\theta = \arg\max_\theta \log p(x|\theta) \tag{1}$$

假设隐变量为z，则EM算法通过不断提高$\log(x|\theta)$的下界来对其优化。
下面先求下界函数：

$$\begin{aligned}
&\log p(x|\theta) \\
&= \log \sum_z p(z)[p(x,z|\theta)/p(z)] \\
&\geq \sum_z p(z) \log[p(x,z|\theta)/p(z)]
\end{aligned} \tag{2}$$

EM将下界函数分成两个部分：θ 和$p(z)$
在E step中，固定θ，求最优的$p(z)$ 在M step中，固定$p(z)$, 求最优的θ
先来看E step,(2)等号成立当且仅当$p(x,z|\theta)/p(z)$ 为常数
即

$$\begin{aligned}
&p(x,z|\theta)/p(z) = c \\
&\Rightarrow \sum_z p(x,z|\theta)/p(z) \times p(z) = c \\
&\Rightarrow p(x|\theta) = c \\
&\Rightarrow p(x,z|\theta)/p(z) = p(x|\theta) \\
&\Rightarrow p(z) = p(z|x,\theta)
\end{aligned} \tag{3}$$

也就是E step:
求

$$p(z)^{(k)} = p(z|x,\theta^{(k-1)}) \tag{4}$$

此时下界函数正好取到原函数值
接着看M step，此时$p(z)$已经固定，最大化(2)等于最大化$\sum_z p(z)\log p(x,z|\theta)$，即

$$\theta^{(k)} = \arg\max_\theta \sum_z p(z)^{(k-1)} p(x,z|\theta) \tag{5}$$

(最大似然估计)

2 高斯混合模型中的EM

这部分请参考[1]

已知观察变量$x = \{x_1, x_2, ..., x_n\}$,x_i 为第i个样本观察值，且互相独立，这些样本由m个高斯模型$y = \{y_1, y_2, ..., y_m\}$产生。每个模型有三个参数：

1.先验概率$a_j = p(y_j), a_j > 0 \sum a_j = 1$

2.模型的期望$u_j = E(y_j)$

3.方差$S_j = (y_j)$.

记$\theta = \{a_1, u_1, S_1, ..., a_m, u_m, S_m\}$ 为整个模型的参数。

整个模型描述为：

$$\begin{aligned}
&\log p(x|\theta) \\
&= \sum_i \log \sum_j p(x_i|y_j, \theta) \times p(y_j|\theta) \qquad (\because\ x\ yields\ i.i.d) \\
&= \sum_i \log \sum_j a_j \times \frac{1}{(2\pi)^{1/2}|S_j|^{d/2}} \exp^{(x_i-u_j)^T S_j^{-1}(x_i-u_j)}
\end{aligned} \tag{6}$$

现在求参数$\theta = \arg\max_\theta \log p(x|\theta)$

这里隐变量为$p(z) = p(y|x)$，也就是我们不知道样本x_i是由模型y_j产生的概率。

$$\begin{aligned}
&\log p(x|\theta) \\
&= \sum_i \log p(x_i|\theta) \\
&= \sum_i \log \sum_j p(y_j|x_i)[p(x_i, y_j|\theta)/p(y_j|x_i)] \\
&\geq \sum_i \sum_j p(y_j|x_i) \log[p(x_i, y_j|\theta)/p(y_j|x_i)](low\ bound\ function)
\end{aligned} \tag{7}$$

等式成立当且仅当$p(y_j|x_i) = p(y_j|x_i, \theta)$ ，推导同前

即E step:

$p(y_j|x_i)^{(k)} = p(y_j|x_i, \theta^{(k-1)})$

M step:

$$\theta^{(k)} = \arg\max_\theta \sum_{ij} p(y_j|x_i)^{(k-1)} \log p(x_i, y_j|\theta) \tag{8}$$

以求最优的$a_j^{(k)}$为例：

$$\frac{\partial \sum_{ij} p(y_j|x_i)^{(k-1)} \log p(x_i, y_j|\theta) + \lambda(\sum_j a_j - 1)}{\partial a_j} = 0$$

$$\Rightarrow \sum_i p(y_j|x_i)^{(k-1)} \times \frac{1}{a_j \times p(x_i|y_j, \theta)} \times p(x_i|y_j, \theta) + \lambda = 0$$

$$\Rightarrow a_j = -\sum_i p(y_j|x_i)^{(k-1)}/\lambda$$

$$\Rightarrow -\sum_{ij} p(y_j|x_i)^{(k-1)}/\lambda = 1 \qquad (\because \sum_j a_j = 1)$$

$$\Rightarrow \lambda = -n$$

$$\Rightarrow a_j = \sum_i p(y_j|x_i)^{(k-1)}/n \tag{9}$$

3 基于EM算法的无监督中文分词算法

这部分请参考文献[3]，伪代码见[2] 所谓的监督和无监督的中文分词的一个区别在于前者是需要给定的字典，后者不需要。所谓的字典就是一个关于词的分布的参数θ。一个字典含有M个词，每个词的出现概率为θ_i，则$\theta_i \geq 0, \sum_i^M \theta_i = 1$; 给定一个句子$x = x_1 x_2 ... x_n$，共有$2^{(n-1)}$种分割方法，每一种分割方法有一个概率，假设词的分布相互独立，我们有：$p(x_1 x_2 ... x_n = w_{i1} w_{i2} .. w_{im}) = p(w_{i1}) p(w_{i2}) .. p(w_{im}) = \theta_{i1} \theta_{i2} ... \theta_{im}$ 这样的$2^{(n-1)}$种分割方法就有$2^{(n-1)}$个概率，我们记

$$p(z, x|\theta) = \{p(z_1, x|\theta), p(z_2, x|\theta), ..., p(z_{2^{n-1}}, x|\theta)\} \tag{10}$$

为分割的概率分布。

最大似然分词法：在监督分词中，θ和x已知，我们要求这样的一个分割

$$z^* = \arg\max_z p(x, z|\theta) \tag{11}$$

这一个z^*可以通过动态规划算法得到，不再多述。

在无监督分词中，θ未知，这样z^*也无法得到，因此我们首先要估计出θ，根据EM算法，我们有

$$\theta^* = \arg\max_\theta \log p(x|\theta) \tag{12}$$

在E step中我们固定θ，得到

$$p(z)^{(k)} = p(z|x, \theta^{(k-1)}) \tag{13}$$

也就是已知词典$\theta^{(k-1)}$，求出$2^{(n-1)}$种分割方法的概率，k为叠代的次数

在M step中我们固定$p(z)$,求

$$\theta^{(k)} = \arg\max_\theta \sum_z p(z)^{(k-1)} \log p(x, z|\theta) \tag{14}$$

注意到

$$\begin{aligned}
&\frac{\partial \sum_z p(z)^{(k-1)} \log p(x, z|\theta) + \lambda(\sum_i \theta_i - 1)}{\partial \theta_i} \\
&= \frac{\partial \sum_z p(z)^{(k-1)} \times \sum_i n(w_i, z) \log \theta_i}{\partial \theta_i} + \lambda \\
&= \sum_z p(z)^{(k-1)} \times n(w_i, z)/\theta_i + \lambda \\
&= 0
\end{aligned} \tag{15}$$

$n(w_i, z)$表示在分割z中w_i出现的次数

因此$\theta_i^{(k)} = -\sum_z p(z)^{(k-1)} \times n(w_i, z)/\lambda$

换句话说,$\theta_i^{(k)}$正比于在$k-1$次叠代中所有包含w_i的分割z的概率乘上包含个数的总和,而这个总和的概率,可以通过前向后向算法得到.因此我们无须在E step中求出具体的z分布,得到$\theta^{(k-1)}$后直接用前向后向算法就可以求得$\theta^{(k)}$

4 分词实验

我们用93-95年的人民日报作为语料，共131M。整个EM算法用C++实现。程序运行9小时56分钟，消耗内存817M，叠代37次收敛。生成24.6M个词条。接着，我们用80k的人工分好的语料作为测试，最后实验的精度为77%

下面是一段分词结果：

中国/已确定了/未来五年/高技术/研究重点/，/并/着手制订/下世纪的/高科技/研究计划/。/新华社/南京/十二月/四日电/中国/最大的/氨/伦/丝/生产基地/－/－/钟山/氨/伦/有限公司/，/日前在/连云港/开发区建/成并投产/。/目前/，/王翔/又开始了/九龙/街/的建设/。/与此同时/，/为改善/九江/的/投资环境/，/王翔/决定投资/数亿元/在/九江市区/开发建设/与世界/名山/庐山/相呼应的/“/大千世界/游乐园/”/。/据上海市/计委/专家分析/测算/，/要在二/０ ０ ０年/实现/人均国内/生产总值/五千美元/的目标/，/今后三年/上海/国内生产/总值平均/年增幅/要达到/百分之十/至百分之/十一/。/美国传统/基金会/副总裁/霍姆/斯/博士和/道/琼斯公司/副总裁/克/罗/维/茨/今天

早晨/率领由/传统/基金会及/《/华尔街日/报/》/组成的/高层/代表团/，/向/董建华/递交了/报告结果/。/据悉/，/一九七三年/中国/天津市与/日本神户/结成/第一/对中外/友好城市/。/据天津市/外经贸委/介绍/，/这/三百多家/企业/涉及/进出口/贸易/、/生产加工/、/医疗卫生/、/旅游/、/餐饮/、/工程承包/等行业/，/分布在/六十多个/国家/和地区/。/餐饮业/是天津市/在/海外投资/的/重点之一/。/天津/在新加坡/开设的/“/药膳/”/受到当地/政/要/的关注/。/

分析：EM算法喜欢把一些高频的词组作为词，比如“下世纪的”，“已确定了”，这对实验结果影响很大。在未来的工作中，我们希望用少量的分好的语料来改善EM分词的结果。

5 参考文献

1.Jeff A. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models

2.http://www.datalab.uci.edu/people/xge/mlword/node4.html

3.Xianping Ge, Wanda Pratt, Padhraic Smyth, Discovering Chinese Words from Unsegmented Text, SIGIR 1999.