EM 算法

learning note For reading translation

我真的不懂忧郁

Delft University of Technology

TU Delft

EM 算法

learning note For reading translation

by

我真的不懂忧郁

Student Name	Student Number
First Surname	1234567

Instructor:	I. Surname
Teaching Assistant:	I. Surname
Project Duration:	Month, Year - Month, Year
Faculty:	Faculty of Aerospace Engineering, Delft

Preface

A preface...

我真的不懂忧郁
Delft, June 2024

Summary

A summary...

目录

Preface **i**

Summary **ii**

Nomenclature **iv**

1 EM 算法 **1**

1.1 背景 . . . 1

1.2 EM 算法导出 . . . 2

1.3 EM 算法 . . . 3

2 从 KL 散度的视角看 EM 算法 **4**

2.1 主要过程 . . . 4

2.2 KL 散度正定性证明 . . . 5

3 算法收敛性 **6**

3.1 直观推导 . . . 6

3.2 收敛性定理 . . . 6

4 广义 EM 算法 **8**

4.1 问题 . . . 8

4.2 坐标上升法 . . . 8

References **9**

A Source Code Example **10**

B Task Division Example **11**

Nomenclature

If a nomenclature is required, a simple template can be found below for convenience. Feel free to use, adapt or completely remove.

Abbreviations

Abbreviation	Definition
ISA	International Standard Atmosphere
...	

Symbols

Symbol	Definition	Unit
V	Velocity	[m/s]
...		
ρ	Density	[kg/m^3]
...		

Chapter 1

EM 算法

1.1. 背景

EM 算法主要是为了解决生成模型参数 θ 不均匀问题，也就是含有隐变量模型的 learning 问题 $\hat{\theta} = arg\max_{\theta} P(X|\theta)$。

example 1.1.1: 假设有三枚硬币，分别记作 A, B, C，这些硬币正面出现的概率分别是 π、p、q，进行如下试验，先投掷硬币 A，根据其结果选出硬币 B 或硬币 C，正面选 B，反面选 C，然后掷选出的硬币，投掷硬币的结果，正面记作 *1*，反面记作 *0*，独立重复 $n = 10$ 次试验，结果如下

$$1, 1, 0, 1, 0, 0, 1, 0, 1, 1 \tag{1.1}$$

假设只能观测到投掷硬币的结果，不能观测到过程，问如何估计三硬币正面出现的概率，即三硬币模型的参数。

Solution.

概率模型写作

$$\begin{aligned} P(y|\theta) &= \sum_z P(y, z|\theta) = \sum_z P(z|\theta)P(y|z, \theta) \\ &= \pi p^y(1-p)^{1-y} + (1-\pi)q^y(1-q)^{1-y} \end{aligned} \tag{1.2}$$

这里 y 是观测变量，z 是**隐变量**，表示 A 投掷的结果，$\theta = (\pi, p, q)$ 表示模型参数。

将观测数据表示为 $Y = (Y_1, Y_2, \cdots, Y_n)^T$，未观测数据表示为 $Z = (Z_1, Z_2, \cdots, Z_n)^T$，则观测数据的似然函数为

$$\begin{aligned} P(Y|\theta) &= \sum_Z P(Z|\theta)P(y|Z, \theta) \\ &= \prod_{j=1}^{n} \pi p^{y_j}(1-p)^{1-y_j} + (1-\pi)q^{y_j}(1-q)^{1-y_j} \end{aligned} \tag{1.3}$$

考虑求模型参数 $\theta = (\pi, p, q)$ 的极大似然估计

$$\hat{\theta} = \arg\max_{\theta}\ log\ P(Y|\theta) \tag{1.4}$$

□

Question 1: 为什么这个问题没法儿求解。

这个问题是没有解析解的，只有通过迭代的算法求解，EM 算法就是求解这个问题的一种迭代算法。

1.2. EM 算法导出

EM 算法的基本问题时近似实现对观测数据的极大似然估计，即面对一个含有隐变量的模型，目标是最大观测数据 Y 关于 θ 的对数似然函数。

$$\begin{aligned} L(\theta) &= \log\ P(Y|\theta) = \log \sum_{Z} P(Y,Z|\theta) \\ &= log\left(\sum_{Z} P(Y|Z,\theta)P(Z|\theta)\right) \end{aligned} \tag{1.5}$$

EM 算法是通过迭代逐步近似极大化 $L(\theta)$ 的，假设第 i 次迭代后 θ 的估计值是 $\theta^{(i)}$，我们希望新的估计值能使得 $L(\theta)$ 增加，即 $L(\theta) > L(\theta^{(i)})$，并逐步达到极大值，所以考虑两者的差，注意到 log 函数是凹函数，利用凹函数的 *Jensen* 不等式。[1]

$$\begin{aligned} L(\theta) - L(\theta^{(i)}) &= \log\left(\sum_{Z} P(Z|Y,\theta^{(i)}) \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})}\right) - \log P(Y|\theta^{(i)}) \\ &\geqslant \sum_{Z} P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - \log P(Y|\theta^{(i)}) \\ &= \sum_{Z} P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})} \end{aligned} \tag{1.6}$$

令

$$B(\theta,\theta^{(i)}) = L(\theta^{(i)}) + \sum_{Z} P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})} \tag{1.7}$$

则

$$L(\theta) \geqslant B(\theta,\theta^{(i)}) \tag{1.8}$$

且容易知道

$$L(\theta^{(i)}) = B(\theta^{(i)},\theta^{(i)}) \tag{1.9}$$

即 $B(\theta,\theta^{(i)})$ 是 $L(\theta)$ 的一个下界，通过使得下界 $B(\theta,\theta^{(i)})$ 增大使得似然函数增大，这就是 EM 算法的思路。选择参数 $\theta^{(i+1)}$ 使得

$$\theta^{(i+1)} = \arg\max_{\theta}\ B(\theta,\theta^{(i)}) \tag{1.10}$$

现在求这个参数

[1] 凹函数的 Jensen 不等式：$f(\sum_i\ k_i x_i) \geqslant \sum_i k_i f(x_i)$

$$
\begin{aligned}
\theta^{(i+1)} &= \arg\max_{\theta} \left(L(\theta^{(i)}) + \sum_{Z} P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} \right) \\
&= \arg\max_{\theta} \left(\sum_{Z} P(Z|Y,\theta^{(i)}) \log P(Y|Z,\theta)P(Z|\theta) \right) \\
&= \arg\max_{\theta} \left(\sum_{Z} P(Z|Y,\theta^{(i)}) \log P(Y,Z|\theta) \right) \\
&= \arg\max_{\theta} \; Q(\theta,\theta^{(i)})
\end{aligned} \tag{1.11}
$$

即通过求 Q 函数及其极大化，EM 算法是通过不断求解下界的极大化来逼近求解对数似然函数的算法。

definition 1.2.1: ***(Q 函数)*** *完全数据的对数似然函数* $\log\ P(Y,Z|\theta)$ *关于在给定观测数据* Y *和当前参数* $\theta^{(i)}$ *下对未观测数据* Z *对条件概率分布* $P(Z|Y,\theta^{(i)})$ *对期望称为* Q *函数，即*

$$
Q(\theta,\theta^{(i)}) = E_Z\left[\log\ P(Y,Z|\theta) \mid Y,\theta^{(i)}\right] \tag{1.12}
$$

1.3. EM 算法

算法 (EM 算法)

输入：观测变量数据 Y，隐变量数据 Z，联合分布 $P(Y,Z|\theta)$，条件分布 $P(Z|Y,\theta)$；

输出：模型参数 θ;

（1）选择参数的初值 $\theta^{(i)}$，开始迭代；

（2）**E-step**：记 $\theta^{(i)}$ 为第 i 次迭代参数的估计值，在第 $i+1$ 次迭代的 E 步，计算

$$
\begin{aligned}
Q(\theta,\theta^{(i)}) &= E_Z\left[\log\ P(Y,Z|\theta) \mid Y,\theta^{(i)}\right] \\
&= \sum_{Z} \ log\, P(Y,Z|\theta)P(Z|Y,\theta^{(i)})
\end{aligned} \tag{1.13}
$$

（3）**M-step**：求 $Q(\theta,\theta^{(i)})$ 极大化 θ，确定 $i+1$ 次迭代的参数估计值 $\theta^{(i+1)}$

$$
\theta^{(i+1)} = \arg\ \max_{\theta} Q(\theta,\theta^{(i)}) \tag{1.14}
$$

（4）重复 (2) 和 (3)，直到收敛；

值得注意的是，算法参数的初值是任意选择，但是 EM 算法对初值是敏感的。

Chapter 2

从 KL 散度的视角看 EM 算法

2.1. 主要过程

我们要最大化对数似然函数

$$L(\theta) = log\ P(X|\theta) \tag{2.1}$$

其中概率 $P(X|\theta)$ 可以写成

$$P(X|\theta) = \frac{P(X,Z|\theta)}{P(Z|X,\theta)} \tag{2.2}$$

其中 Z 是隐变量。引入隐变量的概率分布 $Q(Z)$

$$P(X|\theta) = \frac{P(X,Z|\theta)/Q(Z)}{P(Z|X,\theta)/Q(Z)} \tag{2.3}$$

所以

$$logP(X|\theta) = log\frac{P(X,Z|\theta)}{Q(Z)} - log\frac{P(Z|X,\theta)}{Q(Z)} \tag{2.4}$$

两边对 $Q(Z)$ 求数学期望

$$\int_Z Q(Z)logP(X|\theta)dZ = \int_Z Q(Z)log\frac{P(X,Z|\theta)}{Q(Z)}dZ + \int_Z Q(Z)log\frac{Q(Z)}{P(Z|X,\theta)}dZ \tag{2.5}$$

分别看式子两边，式子左边 $P(X|\theta)$ 和 $Q(Z)$ 无关，因此

$$左边 = logP(X|\theta)\int_Z Q(Z)dZ = logP(X|\theta) = L(\theta) \tag{2.6}$$

式子右边第二项刚好就是 X 和 θ 下 Z 的分布和 Z 的理想分步下的 KL 散度，即

$$KL(Q(Z)||P(Z|X,\theta)) = \int_Z Q(Z)log\frac{Q(Z)}{P(Z|X,\theta)}dZ \tag{2.7}$$

式子第一项就是证据下界 *ELBO*

$$ELBO = \int_Z Q(Z)log\frac{P(X,Z|\theta)}{Q(Z)}dZ \tag{2.8}$$

由于 KL 散度正定性，有下面的不等关系

$$L(\theta) = ELBO + KL(Q(Z)||P(Z|X,\theta)) \geqslant ELBO \tag{2.9}$$

等号成立的关键是 $KL(Q(Z)||P(Z|X,\theta)) = 0$，即 $Q(Z) = P(Z|X,\theta)$。

$$\hat{\theta} = arg\max_{\theta} \int_Z Q(Z)log\frac{P(X,Z|\theta)}{Q(Z)}dZ \tag{2.10}$$

因此 EM 算法就是最大化证据下界 $ELBO$，因为 Z 是隐变量无法被观测，所以这里 $Q(Z)$ 的分布我们还没确定，但我们可以假设其等于给定 X 和 θ 的后验

$$Q(Z) = P(Z|X,\theta^{(i)}) \tag{2.11}$$

其中 $\theta^{(i)}$ 是确定的，我们用来估计 $\hat{\theta} = \theta^{(i+1)}$ 那么

$$\theta^{(i+1)} = arg\max_{\theta} \int_Z P(Z|X,\theta^{(i)})log\frac{P(X,Z|\theta)}{P(Z|X,\theta^{(i)})}dZ \tag{2.12}$$

这样 EM 算法就是一个迭代的算法，利用前次计算的参数来推断下一次。

$$\begin{aligned}
\theta^{(i+1)} &= arg\max_{\theta} \int_Z P(Z|X,\theta^{(i)})log\frac{P(X,Z|\theta)}{P(Z|X,\theta^{(i)})}dZ \\
&= arg\max_{\theta} \int_Z P(Z|X,\theta^{(i)})logP(X,Z|\theta)dZ - \int_Z P(Z|X,\theta^{(i)})logP(Z|X,\theta^{(i)})dZ \\
&= arg\max_{\theta} \int_Z P(Z|X,\theta^{(i)})logP(X,Z|\theta)dZ
\end{aligned} \tag{2.13}$$

2.2. KL 散度正定性证明

我们要证明 KL 散度的正定性，即

$$KL(p(z)||q(z)) = \int_z p(z)log\frac{p(z)}{q(z)}dz \geqslant 0 \tag{2.14}$$

注意到 log 函数是一个凹函数[1]，所以

$$\begin{aligned}
\int_z p(z)log\frac{p(z)}{q(z)}dz &= -\int_z p(z)log\frac{q(z)}{p(z)}dz \\
&\leqslant -log\int_z q(z)dz = -log\,1 = 0
\end{aligned} \tag{2.15}$$

因此

$$KL(p(z)||q(z)) = \int_z p(z)log\frac{p(z)}{q(z)}dz \geqslant 0 \tag{2.16}$$

[1]若 f 是凹函数，则 $\sum_i k_i f(x_i) \leqslant f(\sum_i k_i x_i)$

Chapter 3

算法收敛性

3.1. 直观推导

我们前面已经分别从 Jensen 不等式和 KL 散度两个视角推出了 EM 算法的形式，总体来说就是希望 $\theta^{(i)} \to \theta^{(i+1)}$ 时，有

$$logP(X|\theta^{(i)}) \leqslant logP(X|\theta^{(i+1)}) \tag{3.1}$$

根据前面的推导，我们最优化的方法是最大化证据下界

$$ELBO = \underbrace{\int_Z P(Z|X,\theta^{(i)})logP(X,Z|\theta)dZ}_{\mathcal{L}(\theta,\theta^{(i)})} - \underbrace{\int_Z P(Z|X,\theta^{(i)})logP(Z|X,\theta)dZ}_{\mathcal{H}(\theta,\theta^{(i)})} \tag{3.2}$$

我们显然可以得到 $\mathcal{L}(\theta^{(i+1)},\theta^{(i)}) \geqslant \mathcal{L}(\theta^{(i+1)},\theta^{(i)}$，如果我们想知道每次迭代是否 $ELBO$ 都在增加，只要

$$\mathcal{H}(\theta^{i+1},\theta^{^{(i)}}) \leqslant \mathcal{H}(\theta^{i},\theta^{^{(i)}}) \tag{3.3}$$

两者做差很容易证明

$$\mathcal{H}(\theta^{i+1},\theta^{^{(i)}}) - \mathcal{H}(\theta^{i},\theta^{^{(i)}}) \leqslant 0 \tag{3.4}$$

3.2. 收敛性定理

theorem 3.2.1: 设 $P(Y|\theta)$ 为观测数据的似然函数，$\theta^{(i)}(i=1,2,\cdots)$ 为 *EM* 算法得到参数估计序列，$P(Y|\theta^{(i)})(i=1,2,\cdots)$ 为对应的似然函数序列，则 $P(Y|\theta^{(i)})$ 是单调递增的，即

$$P(Y|\theta^{(i+1)}) \geqslant P(Y|\theta^{(i)}) \tag{3.5}$$

proof.

□

theorem 3.2.2: 设 $L(\theta) = \log\ P(Y|\theta)$ 为观测数据的对数似然函数，$\theta^{(i)}(i = 1, 2, \cdots)$ 为 EM 算法对参数估计序列，$L(\theta^{(i)})(i = 1, 2, \cdots)$ 对应的对数似然函数序列

(1) 如果 $P(Y|\theta)$ 有上界，则 $L(\theta^{(i)}) = \log\ P(Y|\theta^{(i)})$ 收敛到某一个值 L^**;*

(2) 在函数 $Q(\theta, \theta^*)$ 与 $L(\theta)$ 满足一定条件下，由 EM 算法得到的参数估计序列 $\theta^{(i)}$ 的收敛值 θ^* 是 $L(\theta)$ 的稳定点。

proof.

Chapter 4

广义 EM 算法

4.1. 问题

4.2. 坐标上升法

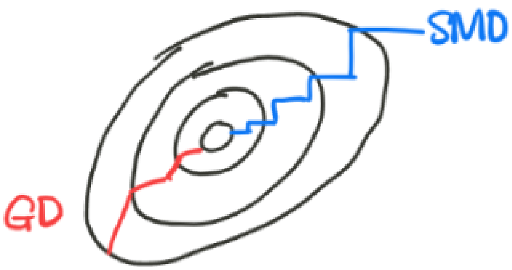

图 4.1: 坐标上升和梯度上升优化思路对比

References

[1] I. Surname, I. Surname, and I. Surname. “The Title of the Article”. In: *The Title of the Journal* 1.2 (2000), pp. 123–456.

Chapter A

Source Code Example

Adding source code to your report/thesis is supported with the package listings. *An example can be found below. Files can be added using* \lstinputlisting[language=<language>]{<filename>}.

```python
"""
ISA Calculator: import the function, specify the height and it will return a
list in the following format: [Temperature,Density,Pressure,Speed of Sound].
Note that there is no check to see if the maximum altitude is reached.
"""

import math
g0 = 9.80665
R = 287.0
layer1 = [0, 288.15, 101325.0]
alt = [0,11000,20000,32000,47000,51000,71000,86000]
a = [-.0065,0,.0010,.0028,0,-.0028,-.0020]

def atmosphere(h):
    for i in range(0,len(alt)-1):
        if h >= alt[i]:
            layer0 = layer1[:]
            layer1[0] = min(h,alt[i+1])
            if a[i] != 0:
                layer1[1] = layer0[1] + a[i]*(layer1[0]-layer0[0])
                layer1[2] = layer0[2] * (layer1[1]/layer0[1])**(-g0/(a[i]*R))
            else:
                layer1[2] = layer0[2]*math.exp((-g0/(R*layer1[1]))*(layer1[0]-layer0[0]))
    return [layer1[1],layer1[2]/(R*layer1[1]),layer1[2],math.sqrt(1.4*R*layer1[1])]
```

Chapter B

Task Division Example

If a task division is required, a simple template can be found below for convenience. Feel free to use, adapt or completely remove.

表 **B.1:** Distribution of the workload

	Task	Student Name(s)
	Summary	
Chapter 1	Introduction	
Chapter 2		
Chapter 3		
Chapter *		
Chapter *	Conclusion	
	Editors	
	CAD and Figures	
	Document Design and Layout	