浅谈 EM 算法的应用

喻瀚森

四川大学锦城学院 四川 成都 611731

【摘 要】本文通过对 EM 算法的简单原理介绍，并且选取了一个数据集，通过 C 语言和 python 语言编写了程序进行 EM 算法的实际应用，得出了相对应的实验结果。

【关键词】算法；EM；实验

1 引言

在生活中，经常会遇到数据缺失不完整的情况，在数据挖掘的过程中也经常涉及到对于缺失数据的处理。数据缺失的处理方法非常多，但是在大数据环境下补充缺失的数据是非常繁琐的，而 EM 算法是一个能够很稳定找到可靠的缺失值的算法，虽然它有一些缺点，但是它的优势在于算法很简单，能够处理的问题非常的广泛，因此也被人们广泛的使用。[1]

2 EM 算法相关理论

EM 算法的名称是叫最大期望算法，在数据挖掘中对于缺失值的寻找有着重大的意义，这个算法是一种迭代算法，它的主要应用场景是需要在一个模型之中，这个模型是一个概率参数模型中，并且这个概率参数模型之中含有未知的变量，也称为隐变量。在这样一个模型中，存在着一个最大似然估计，这个估计又被称为极大后验概率估计，在软件工程领域中，EM 算法被广泛应用于机器学习之中，并且十分火热。

2.1 算法过程

em 算法是一种优化参数的策略，通过迭代方式来计算优化参数，它的优化计算主要可以分为两步，分别称为 e 步和 m 步，也可以被称为期望步和极大步，因此这个算法也被我们称为 em 算法，是这两个优化步骤的一个简称组合。[2] 这个模型算法的基本思想是：首先根据事先已经提供给出的观测数据集，也可以称为目标观测数据，通过对这个数据集的分析和处理方法来准确地估计得出这个模型参数的平均值；接下来根据上一个参数模型所估计得出的参数值作为一个基本点，继续进行对下一个缺失的数据值进行估计，然后再次继续结合之前上一步的估计得出的这个缺失数据，并根据之前已经观测得出的数据，反复地进行迭代，并对参数值进行估计，在迭代的过程中，这个估计值会收敛，当发生收敛后，迭代结束，程序结束。[3]

2.2 Jensen 不等式

对于 Jensen 不等式的定义如下：

（1）假设函数 f 的一个定义域区间是作为一个实数的，如果任意选取一个实数 x ，都可以存在 f(x) 的二阶导数都是大于 0, 那么 f 就是一个凸函数。

（2）如果存在一个 f 是凸函数，并且当存在 x 不是随机变量而且 x 不是随机常量时，那么 e[f(x)] 就远大于 f(e[x])。当且仅当 x(x) 为一个常量时，e[f(x)] 就相当于 f(e[x])。其中，e(x) 所代表的含义就是 x 对于某个数学预测期望。

（3））当 jensen 不等式被广泛的应用于凹函数的情况下时候，不等号的逆转方向为一个反向，即 E[f(X)] 小于 f(E[X])。当且仅当 x 为常量的时候，E[f(X)] 等于 f(E[X])。其中，e(x) 表示的意思就是 x 对于一个数学期望。

（4）Jensen 不等式的主要用途是用于证明 EM 算法的收敛性。[4]

3 实验设计

实验分别选择了 c++ 语言和 python 两种编程语言环境，本文通过采用一个简单的实例对 em 算法的过程进行了解和学习，em 算法主要是 dempster,laind,rubin 在 1977 年提出的一种估计方法，这个估计方法主要目的就是通过对参数极大似然的估计，这个估计方法我们就能从一个不知名的情况下得到。数的取值可以放置到一个不完整的数据集里面，来对其中各个参数取值的估计量进行 mle 估计，这也是一种很简单和十分实用的机器学习算法。这种技术已经在数据挖掘中得到了广泛的研究和应用，目前我们主要的目标就是为了收集和处理那些缺损的数据，截尾数据。[5]

3.1 数据集选取与初始化参数

本次实验数据如表 3-1 所示，表中的数据是一个交易数据集，T100-T500 分别表示 5 个顾客购买商品 I1-I5 的情况，1 表示购买，0 表示没有。

表 1：交易数据集

	L1	L2	L3	L4	L5
T100	0	0	1	1	1
T200	1	1	0	0	0
T300	0	0	0	1	1
T400	0	1	1	1	1
T500	1	1	0	0	0

3.2 C++ 语言环境下实验

3.2.1 初始化参数

参数初始化：

```
doublearr[2][5]={{0.3,0.4,0.3,0.7,0.4},{0.6,0.5,0.3,0.4,0.4}};// 男女购买各种商品的概率
doublex[5][5]={{0,0,1,1,1},{1,1,0,0,0},{0,0,0,1,1},{0,1,1,1,1},{1,1,0,0,0}};// 数据集
```

3.2.2 E 步

计算期望（E），利用对隐藏变量的现有估计值，计算其最大似然估计值；

```
voide_step(doublearr[2][5],doublex[5][5],doublea1[5],doublea2[5]){
// 数组 a1 是男生购买的概率，数组 a2 是女生购买的概率，M 步相同
for(inth=0;h<5;h++){a1[h]=a2[h]=1;}
for(inti=0;i<5;i++){
for(intj=0;j<5;j++){a1[i]*=pow(arr[0][j],x[i][j])*pow(1-arr[0][j],1-x[i][j]);a2[i]*=pow(arr[1][j],x[i][j])*pow(1-arr[1][j],1-x[i][j]);}}
for(into=0;o<5;o++){a1[o]=(a1[o]*k1)/(a1[o]*k1+a2[o]*k2);a2[o]=1-a1[o];}// 计算概率
}
```

在每一次进入 E 步的时候，先初始化，然后通过双重循环计算 θ 的值，并更新所有 θ 的值。

3.2.3M 步

最大化（M），是在最大化在 E 步的基础上，利用前面所述的理论求得的最大似然值，通过这个值来计算参数的值。

```
voidm_step(doublearr[2][5],doublex[5][5],doublea1[5],doublea2[5]){
doublesum1=0.0,sum2=0.0;
for(inti=0;i<5;i++){sum1+=a1[i];sum2+=a2[i];cout<<a1[i]<<" "<<a2[i]<<endl;}// 更新男女的概率
k1=sum1/5;k2=sum2/5;// 更新商品的概率
for(intj=0;j<5;j++){doubles1=0.0,s2=0.0;
for(intk=0;k<5;k++){
s1+=a1[k]*x[k][j];s2+=a2[k]*x[k][j];}
arr[0][j]=s1/sum1;
arr[1][j]=s2/sum2;}
}
```

在每一次进入 M 步的时候，都会更新 π 值。

3.2.4 迭代

通过 M 步上找到的一个参数估计值，用这个值反复的进入 E 步计算，这个过程不断反复循环迭代交替进行，直到迭代次数过多或者聚合完成

```
while(flag){e_step(arr,x,a1,a2);m_step(arr,x,a1,a2);print(a1,a2,arr);
cout<<endl<<" k1="<<k1<<" k2="<<k2<<endl;
cout<<"--------------"<<endl;F--;
if(F<=0||K1==k1&&K2==k2){flag=false;}
```

以迭代次数过多为停止条件的原因是避免数据集选取不当而造成程序进入死循环。同时，程序选择迭代结束的条件还可以优化，因为目前选取的条件过于精确。事实上，通过观察迭代过程，迭代结果的精度可以在 0.01，而不是百分之百，所以，这是程序在效率上的优化空间。

3.3 Python 环境下实验

符号解释：（1）πk 表示第 k 个分量出现的概率 (k=2, 可以理解为男性分量和女性分量）这里男女生的概率总和应该为 1。代码中使用角标 _1 代表男性，_2 代表女性。

（2）x(i)j=1（或 0）表示顾客 i 购买了商品 j（或是没有），例如 x(1)1=1 代表的是用户 1 购买了商品 1。代码中用两个字典表示 P1-P5 五个用户的购物情况。分为纵向和横向两种表示。

（3）θkj 表示在第 k 个分量重中观察到第 j 个变量取值为 1 的概率。这里可以理解为男性和女性对五件物品喜好程度。

（4）p(k|i) 表示第 i 条数据属于第 k 个分量的概率。这里可以理解为第 i 个顾客是男性和是女性的概率。

在 python 中，对 float 的计算能表示的最高位数有限，在概率运算中，最后都会停止于 1.0 或 0.0。所就可根据这一点设立停止条件。在 EM 算法中的 p（k|i）表示表示第 i 条数据属于第 k 个分量的概率。即分别有五条数据属于男生和女生，所以当迭代次数足够多时，p 的值则只剩 1.0 或 0.0。可统计数组中 1.0 或 0.0 的个数相加为 10，则停止迭代。

在分析过变量后可以考虑编写 E,M 两步的实现，E 步实现计算第四个变量即五个角色性别的概率。初始设定男生的概率为 0.4，女生的概率为 0.6。

```
defexpectation():
foriinrange(5):
forjinrange(5):
p[0][i]*=theta[0][j]**x[i][j]*(1-theta[0][j])**(1-x[i][j])
p[1][i]*=theta[1][j]**x[i][j]*(1-theta[1][j])**(1-x[i][j])
foriinrange(5):
p[0][i]=(p[0][i]*PI1)/(p[0][i]*PI1+p[1][i]*PI2)
p[1][i]=1-p[0][i]
print('p:')
print(p[0])
print(p[1])
```

以上为 E 步的代码实现。

M 步实现对男女性概率和男女性对物品的喜好程度的更新。

```
defmaximization():
sum1=0.0
sum2=0.0
foriinrange(5):
sum1+=p[0][i]
```

```
sum2+=p[1][i]
PI1=sum1/5
PI2=sum2/5
foriinrange(5):
s1=0.0
s2=0.0
forjinrange(5):
s1+=p[0][j]*x[j][i]
s2+=p[1][j]*x[j][i]
theta[0][i]=s1/sum1
theta[1][i]=s2/sum2
print(‘Π1:’,PI1,’ Π2:’,PI2)
print(‘ :’)
print(theta[0])
print(theta[1])
```

这里设计当新老 πk 的小数点后 5 位不在变化时判断迭代停止。以上为 M 步的代码实现过程。

4 实验结果

通过两种语言完成了对 EM 算法的实例后，运行程序，根据图 1 中输出结果显示，可以看到在第三次迭代后，k1 与 k2 已经趋于稳定后续不再变化。

```
0 0 1 1 1
1 1 0 0 0
0 0 0 1 1
0 1 1 1 1
1 1 0 0 0
P(1|1) = 0.846449,    P(2|1) = 0.153551
P(1|2) = 0.230769,    P(2|2) = 0.769231
P(1|3) = 0.846449,    P(2|3) = 0.153551
P(1|4) = 0.786096,    P(2|4) = 0.213904
P(1|5) = 0.230769,    P(2|5) = 0.769231
arr1:0.156957 0.424289 0.555187 0.843043 0.843043
arr2:0.747019 0.850883 0.178422 0.252981 0.252981
k1=0.588107  k2=0.411893
------------
P(1|1) = 0.998427,    P(2|1) = 0.00157286
P(1|2) = 0.0035628,    P(2|2) = 0.996437
P(1|3) = 0.991027,    P(2|3) = 0.00897268
P(1|4) = 0.98795,    P(2|4) = 0.0120502
P(1|5) = 0.0035628,    P(2|5) = 0.996437
arr1:0.00238751 0.333411 0.665558 0.997612 0.997612
arr2:0.988789 0.994768 0.00675927 0.0112112 0.0112112
k1=0.596906  k2=0.403094
------------
P(1|1) = 1,    P(2|1) = 7.64039e-011
P(1|2) = 2.35262e-009,    P(2|2) = 1
P(1|3) = 1,    P(2|3) = 2.23427e-008
P(1|4) = 1,    P(2|4) = 2.90417e-008
P(1|5) = 2.35262e-009,    P(2|5) = 1
arr1:1.56841e-009 0.333333 0.666667 1 1
arr2:1 1 1.45591e-008 2.57304e-008 2.57304e-008
k1=0.6  k2=0.4
------------
P(1|1) = 1,    P(2|1) = 0
P(1|2) = 6.43027e-028,    P(2|2) = 1
P(1|3) = 1,    P(2|3) = 0
P(1|4) = 1,    P(2|4) = 0
P(1|5) = 6.43027e-028,    P(2|5) = 1
arr1:4.28684e-028 0.333333 0.666667 1 1
arr2:1 1 0 0 0
k1=0.6  k2=0.4
------------
```

图 1：EM 算法执行输出图

【参考文献】

[1] 张宏东．EM 算法及其应用 [D]. 山东大学 , 2014.

[2] 林东方．基于 EM 算法的不完全测量数据的处理方法研究 [D]. 中南大学 , 2012.

[3] 杨晴．EM 算法在混合模型参数估计中的应用 [D]. 宁夏大学 , 2014.

[4] 李顺静．基于 EM 算法的缺失数据的统计分析及应用 [D]. 重庆工商大学 .

[5] 张宝龙．有限混合分布模型参数估计的 EM 算法及模拟 [D]. 宁夏大学 , 2015. 学 , 2015.