选择工作的数学建模

HL 2019.8.7

## **问题描述：**

找工作，同时被多个公司录取，需要做出选择应该去哪家。

**已知：**

一个公司可以由多个维度的特征表示，例如（是否是国企，是否为外资，是否为私企）。

可以通过互联网，查看别人对某类公司多个维度的评价，例如（工资是否满意，环境是否好，加班是否少，管理是否好）。

**求：**

找出公司特征和工作感受的关系。当给出某个公司的特征后，能通过该关系估算出在该公司工作的感受。

## **建模：**

有若干工作选择，每个工作的特征A由n个属性(a1,a2,…an)组成，ai ∈ {0,1}

工作选择后，有感受结果R，每个R为m个属性(r1,r2,…rm)组成，ri为属性ri的分值，表示ri发生的似然度。

则可以用下面的公式表示估计结果发生可能性的算法：

也可简化为：

A为原因向量，表示做出某个行动的属性。在该问题中，行动就是选择某个公司，各个属性就是这个公司的特征。

W为权值矩阵。

R为结果向量，表示某个结果的属性。在改问题中，结果就是在某公司工作的感受。

**该算法的实质是：**

把 因 🡪 果的映射 f: A -> R，

简化为线性函数：

有样本集合S={(Ai,Ri)} | i=1…K, 根据样本估算出W，从而回归出函数f。

**回归法：**

多元线性回归，有样本集S={(A1,R1),(A2,R2),….(Ap,Rp)}, 估算出W

**步骤1.对结果变量每一维度进行回归，求出权值举证每一行的估计值b：**

对每一维的结果ri, 有：

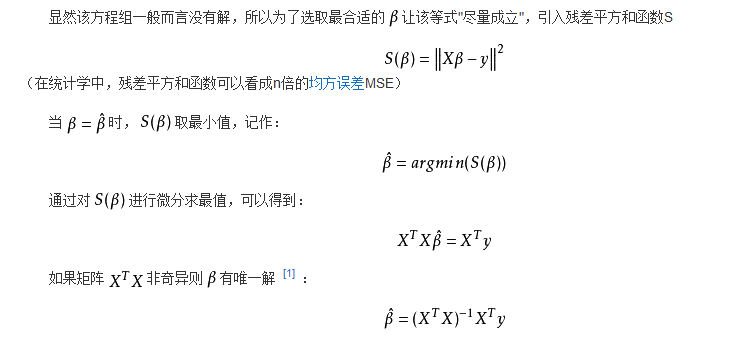
wi1\*a1+wi2\*a2+…+win\*an = ri,

如果省略下标i, 令y=ri, bj=wij, 则b1\*a1+…+bn\*an = y

对W回归转换为对每一行b=(b1,…bn)进行回归：

当有P个样本X=(x1,x2,..xn),Y=(y1,y2..yp) 则对使p个等式组成的超定方程组（b为未知变量）尽量成立：

向量化后：



**步骤2，组合权值矩阵：**

## **计算举例：**

公司有属性（是否是国企，是否为外企，是否为私企）

工作感受有属性（工资是否满意，加班是否少，管理是否好）

从看准网收集评论样本：

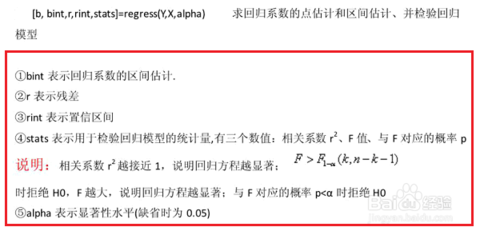
可以搜索2个意向公司的样本：

公司名C1（外企）, C2(国企)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 样本来源 | 样本序号 | a1国企 | a2外企 | r1工资满意 | r2加班少 | r3管理好 |
| C1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 2 | 0 | 1 | 1 | 1 | 1 |
| 3 | 0 | 1 | 1 | 1 | 1 |
| 4 | 0 | 1 | 1 | 1 | 1 |
| 5 | 0 | 1 | 0 | 1 | 0 |
| C2 | 6 | 1 | 0 | 1 | 1 | 0 |
| 7 | 1 | 0 | 1 | 1 | 1 |
| 8 | 1 | 0 | 0 | 1 | 1 |
| 9 | 1 | 0 | 1 | 0 | 0 |
| 10 | 1 | 0 | 0 | 0 | 0 |
| 11 | 1 | 0 | 1 | 0 | 1 |
| 12 | 1 | 0 | 1 | 1 | 0 |
| 13 | 1 | 0 | 0 | 0 | 0 |

对每一个结果维度r1,r2,r3, 用Matlab求回归：

[b,bint,r,rint,stats]=regress(Y,X)，其中b是方程的系数矩阵，bint是回归系数的区间，r残差,rint置信区间，stats用于检验回归模型是否正确，分别是R的平方，F值，概率P



样本：

X=

|  |  |
| --- | --- |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |

样本数:

p=13

对维度r1, Y1=[

|  |
| --- |
| 1 |
| 1 |
| 1 |
| 1 |
| 0 |
| 1 |
| 1 |
| 0 |
| 1 |
| 0 |
| 1 |
| 1 |
| 0 |

]

[b1,bint,r,rint,stats1]=regress(Y1,X)

结果：

b1= 0.6250

0.8000

stats1 =

0.0340 0.3875 0.5463 0.2432

即：r1 = 0.625a1+0.8\*a2

用同样的方法，求出r2:

Y2=[

|  |
| --- |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 0 |
| 0 |
| 0 |
| 1 |
| 0 |

]

[b2,bint,r,rint,stats2]=regress(Y2,X)

b2 =

0.5000

1.0000

stats2 = 0.2778 4.2308 0.0642 0.1818

Y3=[

|  |
| --- |
| 1 |
| 1 |
| 1 |
| 1 |
| 0 |
| 0 |
| 1 |
| 1 |
| 0 |
| 0 |
| 1 |
| 0 |
| 0 |

]

[b3,bint,r,rint,stats3]=regress(Y3,X)

b3 =

0.3750

0.8000

stats3 =

0.1720 2.2854 0.1588 0.2432

**b组合W:**

W=[b1';b2';b3']

W =

0.6250 0.8000

0.5000 1.0000

0.3750 0.8000

**统计量的解释：**

stats1 = 0.0340 0.3875 0.5463 0.2432

stats2 = 0.2778 4.2308 0.0642 0.1818

stats3 = 0.1720 2.2854 0.1588 0.2432

统计量的含义：

Stat=[r2, F, p, e]

r2 – 方差解释率，越接近1，数据拟合效果越好

F – F统计量，用于检查模型是否通过检验，查F分布表，若F>F分布表中对应值，则通过检验。

p - F统计量对应的概率，越接近0越好，当P<α时拒绝假设H0（正态分布），回归模型成立。 α为显著性水平，缺省值为0.05。

e – 期望误差

从统计量来看，F都大于显著性水平，回归模型不成立。

**估算：**

R=W.A

对C1公司，A1=[0,1]’,

R1=W\*A2=

0.8000

1.0000

0.8000

对C2公司A2=[1,0]’

R2=W\*A1=

0.6250

0.5000

0.3750

**结论：**

通过比较，在r1(工资满意度), r2(加班少), r3(管理好)三个维度上，C1都好于C2

但是由于样本的自变量过于集中，回归模型的效果不显著。需要考虑线性模型可能不合适，采用非线性模型。

## **简化方法：**

由以上过程，可以发现，我们的自变量定义a1,a2,a3对各个公司来说，是互斥的。某个公司C如果自变量ai=1，则其他自变量aj一定为0。

这种自变量的选取会导致回归的样本的X值过度集中，回归效果不好。

针对这种情况，可以用平均法，估计各个公司工作感受。

E(R)=（E(r1),E(r2),E(r3)）

例如对公司C1,

r=

|  |  |  |
| --- | --- | --- |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 0 | 1 | 0 |

mean(r) = [0.8000 1.0000 0.8000]

对公司c2,

r=

|  |  |  |
| --- | --- | --- |
| 1 | 1 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 0 | 0 | 0 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |
| 0 | 0 | 0 |

mean(r) = [0.6250 0.5000 0.3750]

**结论：**

通过比较，在r1(工资满意度), r2(加班少), r3(管理好)三个维度上，C1都好于C2

和回归法1得出的结论一致。

## **推广**

用线性回归的方法，可以推广到具有大量“选择-结果”样本，需要比较多种选择的决策问题。

例如：

小学 -- 升学率

医院 -- 治愈率