如何进行用户数据挖掘？数据分析下的受众画像

2018-03-09 中国统计网

很多广告后台的受众画像数据，只告诉了我们看了广告的这部分人群是什么样的，而缺失了发生转化的这部分用户的画像数据。原因主要有两点：一是在大部分广告投放过程中，前后端数据是割裂的；二是媒体不愿意公开这么多的数据，甚至受众画像本身都有一定的问题。

如今信息流优化已经成为业内交流的热点，优化创意、定向等已是老生常谈，唯独受众画像的数据分析少有人提及，尚有可挖的地方。今天借此机会，和大家分享一种受众数据分析的思路。

需要强调的是，接下来的广告数据分析有一个最基本的前提：假设媒体提供的数据和甲方监测的数据都是真实准确的。本文以一个真实的案例和数据（今日头条，家装类）向大家介绍，如何用朴素贝叶斯的算法，对今日头条的受众画像进行数据挖掘和分析，从而实现精准定向下的转化率预测。

1.朴素贝叶斯的原理

这个定理解决了现实生活里经常遇到的问题：已知某条件概率，如何得到两个事件交换后的概率，也就是在已知P(A|B)的情况下如何求得P(B|A)。比如，我知道发生转化的用户中，女性的比例是36%，那么当一个女性用户看到我的广告时，她有多大的可能性发生转化。

这里先解释什么是条件概率：

P(A|B)表示事件B已经发生的前提下，事件A发生的概率，叫做事件B发生下事件A的条件概率。其基本求解公式为：

$$P(A|B) = \frac{P(AB)}{P(B)}$$

贝叶斯定理之所以有用，是因为我们在生活中往往遇到这种问题：可以不费力气地直接得出P(A|B)，P(B|A)则很难直接得出，但其实我们更关心P(B|A)，这时候，贝叶斯定理就为我们提供了从P(A|B)获得P(B|A)的道路。

下面省略证明过程，直接给出贝叶斯定理，相信对高中数学还有印象的朋友对这个公式应该不陌生：

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

2.朴素贝叶斯的数据挖掘原理

下面以一个简单的例子，介绍朴素贝叶斯的数据挖掘原理。虽然样本量不多，但足以说明原理和思路。

这里是一份受众画像数据，总共20笔数据，即代表20个UV。填写表单这一字段值为1的合计9笔，即发生转化的用户数为9。

ID	省级地域	地级市	性别	年龄	兴趣分类	填写表单
1	广东	清远	女	24-30岁	餐饮美食	1
2	广东	深圳	男	18-23岁	餐饮美食	1
3	广东	清远	男	31-40岁	法律服务	1
4	广东	佛山	女	24-30岁	房产	1
5	广东	深圳	男	18-23岁	房产	0
6	广东	深圳	男	50岁以上	服饰箱包	1
7	广东	佛山	男	41-50岁	服饰箱包	1
8	广东	东莞	女	31-40岁	家装百货	0
9	广东	深圳	女	18-23岁	家装百货	1
10	广东	清远	女	31-40岁	教育培训	0
11	广东	深圳	女	24-30岁	房产	0
12	广东	深圳	女	31-40岁	金融理财	0
13	广东	佛山	男	24-30岁	金融理财	1
14	广东	深圳	男	24-30岁	科技数码	1
15	广东	深圳	女	31-40岁	科技数码	0
16	广东	东莞	男	41-50岁	旅游出行	0
17	广东	佛山	男	31-40岁	旅游出行	0
18	广东	佛山	女	18-23岁	美容化妆	0
19	广东	深圳	男	24-30岁	美容化妆	0
20	广东	东莞	男	31-40岁	母婴儿童	0

表1

然后，我们把除了 ID（只是编号，对于挖掘没有价值）、省级地域（因为都是广东，对于挖掘没有价值）外的其他字段，做一个占比分布，如图所示：

地级市	分布	性别	分布	年龄	分布	兴趣分类	分布	填写表单	分布
清远	0.15	男	0.55	18-23岁	0.2	餐饮美食	0.1	1	0.45
深圳	0.45	女	0.45	24-30岁	0.3	法律服务	0.05	0	0.55
佛山	0.25			31-40岁	0.35	房产	0.15		
东莞	0.15			41-50岁	0.1	服饰箱包	0.1		
				50岁以上	0.05	家装百货	0.1		
						教育培训	0.05		
						金融理财	0.1		
						科技数码	0.1		
						旅游出行	0.1		
						美容化妆	0.1		
						母婴儿童	0.05		

表2

假设，我想知道 定向 X=（地级市="佛山"，性别="男"，年龄="18-23岁"，兴趣分类="房产"）的转化率，即我想求：**P（填写表单="1" | X)**。

直接是计算不出来的，回到上文提到的朴素贝叶斯，专门解决的就是这种问题，我只需知道**P（X | 填写表单="1")**，就可以通过公式得到P（填写表单="1" | X)。

具体的直接套公式得：

P（填写表单="1" | X) = P（X | 填写表单="1") * P（填写表单="1") / P（X)

同理可得，

P（填写表单="0" | X) = P（X | 填写表单="0") * P（填写表单="0") / P（X)

这里需要引出另外一个重要的公式，P(A,B)代表事件A与B同时发生的概率。

当事件A与B的发生是各自独立时，P(A,B) = P(A|B) * P(B) =P(A)P(B)。

因为，地级市、性别、年龄等这些字段（或定向）的发生可以理解为是各自独立的，所以P（X | 填写表单="1") = P（X ） * P（填写表单="1")，又P（X | 填写表单="1"）= P（地级市="佛山"，性别="男"，年龄="18-23岁"，兴趣分类="房产" | 填写表单="1"）=P（地级市="佛山" | 填写表单="1"） * P（性别="男" | 填写表单="1"） * P（年龄="18-23岁" | 填写表单="1"） * P（兴趣分类="房产" | 填写表单="1"），此时，看起来同样无法直接得到的P（X | 填写表单="1")，被拆分为看起来更简单的5个事件的概率的乘积。

带入具体值，计算得：

P（填写表单="1" | X) = P（X | 填写表单="1") * P（填写表单="1") / P（X）= （3/9 * 6/9 * 2/9 * 1/9）*0.45 / P（X） = 0.002469 / P（X）.....①

P（填写表单="0" | X) = P（X | 填写表单="0") * P（填写表单="0") / P（X）= （2/11 * 5/11 * 2/11 * 2/11）*0.55 / P（X） = 0.0015026 / P（X）.....②

接下来，遇到一个问题，P（X）是多少，不知道！不过不要紧，当定向 X的用户进来时，ta要么转化，要么不转化，所以

P（填写表单="1" | X) + P（填写表单="0" | X) =1.....③

联立①②③，最终求出：

P（填写表单="1" | X) = 62.2%

P（填写表单="0" | X) = 37.8%

所以，当定向为X时，朴素贝叶斯数据挖掘模型认为，该类用户的转化率在62.2%。

3.朴素贝叶斯的数据挖掘的优势

主流的数据挖掘算法，如神经网络、决策树等。多半依赖如表1所示的数据，每一个字段代表用户的不同维度，每一行代表一个独立用户的数据。但实际优化过程中，媒体方不可能提供如此详尽的受众画像数据， 但朴素贝叶斯不一样，对原始数据的要求略低，只需提供不同维度组合下的比例，而不必细化到每一个用户的情况。

4.朴素贝叶斯的数据挖掘案例解读

4.1 原生数据及预处理

我们从今日头条广告后台拿到的数据经过简单处理后，是下面这样的：

合计13339点击，转化量为37。

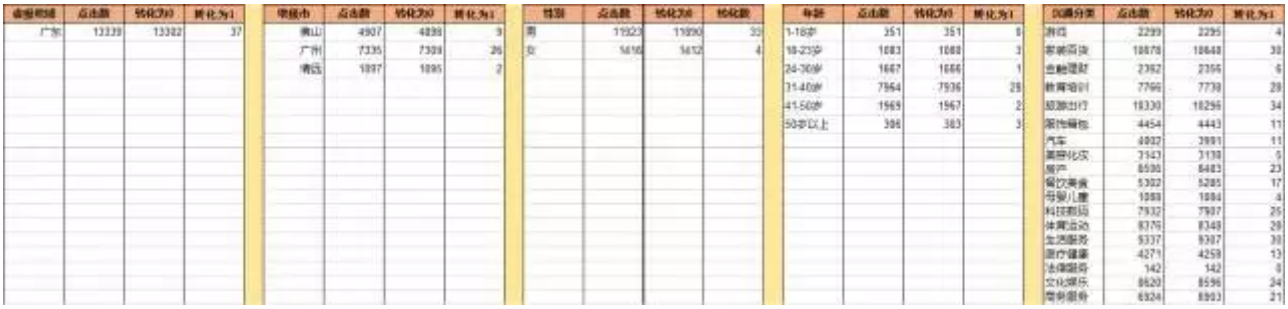


表3

4.2 计算字段重要性，确定输入字段

因为所有字段都是类别型字段（区别于数值型字段），这里介绍一个比较通用的算法，用于评估**所有可能的输入字段**对**输出字段**的**重要性**。

字段重要性	地级市	0			1			Attribute ValueImportance
		Frequency	Support	Confidence	Frequency	Support	Confidence	
0.1595	佛山	4898	0.3682	0.6022	9	0.2432	0.3978	0.0752
	广州	7309	0.5495	0.4388	26	0.7027	0.5612	0.0673
	清远	1095	0.0823	0.6036	2	0.0541	0.3964	0.0170

公式解读如下：

	A	B	C	D	E	F	G	H	I	J
1										
2		字段重要性	地级市	0			1			Attribute ValueImportance
3				Frequency	Support	Confidence	Frequency	Support	Confidence	
4		=J4+J5+J6	佛山	4898	=D4/(D4+D5+D6)	=E4/(E4+H4)	9	=G4/(G4+G5+G6)	=H4/(E4+H4)	=ABS(F4-I4)*(D4+G4)/13339
5			广州	7309	0.5495	0.4388	26	0.7027	0.5612	0.0673
6			清远	1095	0.0823	0.6036	2	0.0541	0.3964	0.0170
7										

注：ABS函数，用于求绝对值。

所有可能的输入字段对**输出字段**的**重要性**计算结果如下：

字段重要性	地级市	0			1			Attribute ValueImportance
		Frequency	Support	Confidence	Frequency	Support	Confidence	
0.1595	佛山	4898	0.3682	0.6022	9	0.2432	0.3978	0.0752
	广州	7309	0.5495	0.4388	26	0.7027	0.5612	0.0673
	清远	1095	0.0823	0.6036	2	0.0541	0.3964	0.0170

字段重要性	性别	0			1			Attribute ValueImportance
		Frequency	Support	Confidence	Frequency	Support	Confidence	
0.0020	男	11890	0.8939	0.5005	33	0.8919	0.4995	0.0010
	女	1412	0.1061	0.4954	4	0.1081	0.5046	0.0010

字段重要性	年龄	0			1			Attribute ValueImportance
		Frequency	Support	Confidence	Frequency	Support	Confidence	
0.2591	1-18岁	351	0.0264	1.0000	0	0.0000	0.0000	0.0263
	18-23岁	1080	0.0812	0.5003	3	0.0811	0.4997	0.0001
	24-30岁	1666	0.1252	0.8225	1	0.0270	0.1775	0.0806
	31-40岁	7936	0.5966	0.4408	28	0.7568	0.5592	0.0707
	41-50岁	1967	0.1479	0.7323	2	0.0541	0.2677	0.0686
	50岁以上	303	0.0228	0.2193	3	0.0811	0.7807	0.0129

字段重要性	兴趣分类	0			1			Attribute ValueImportance
		Frequency	Support	Confidence	Frequency	Support	Confidence	
0.4557	游戏	2295	0.0222	0.6358	4	0.0127	0.3642	0.0468
	家装百货	10648	0.1032	0.5192	30	0.0955	0.4808	0.0307
	金融理财	2356	0.0228	0.5443	6	0.0191	0.4557	0.0157
	教育培训	7738	0.0750	0.4567	28	0.0892	0.5433	0.0504
	旅游出行	10296	0.0998	0.4795	34	0.1083	0.5205	0.0318
	服饰箱包	4443	0.0430	0.5513	11	0.0350	0.4487	0.0343
	汽车	3991	0.0387	0.5247	11	0.0350	0.4753	0.0148
	美容化妆	3138	0.0304	0.6563	5	0.0159	0.3437	0.0736
	房产	6483	0.0628	0.4616	23	0.0732	0.5384	0.0374
	餐饮美食	5285	0.0512	0.4861	17	0.0541	0.5139	0.0111
	母婴儿童	1084	0.0105	0.4519	4	0.0127	0.5481	0.0079
	科技数码	7907	0.0766	0.4904	25	0.0796	0.5096	0.0115
	体育运动	8348	0.0809	0.4756	28	0.0892	0.5244	0.0306
	生活服务	9307	0.0902	0.4855	30	0.0955	0.5145	0.0203
	医疗健康	4258	0.0413	0.4991	13	0.0414	0.5009	0.0006
	法律服务	142	0.0014	1.0000	0	0.0000	0.0000	0.0106
	文化娱乐	8596	0.0833	0.5214	24	0.0764	0.4786	0.0277
	商务服务	6903	0.0669	0.5000	21	0.0669	0.5000	0.0000

一般经验来说，字段重要性小于0.1的字段可以不予纳入数据挖掘模型中。

所以，目前根据有限的数据，” 性别 “这一字段，对于判断用户是否转化的帮助不大，故在接下来的数据挖掘模型中，输入字段包括：地级市、年龄、兴趣分类。

开始做数据挖掘，具体原理这里不再赘述，直接给出结果。

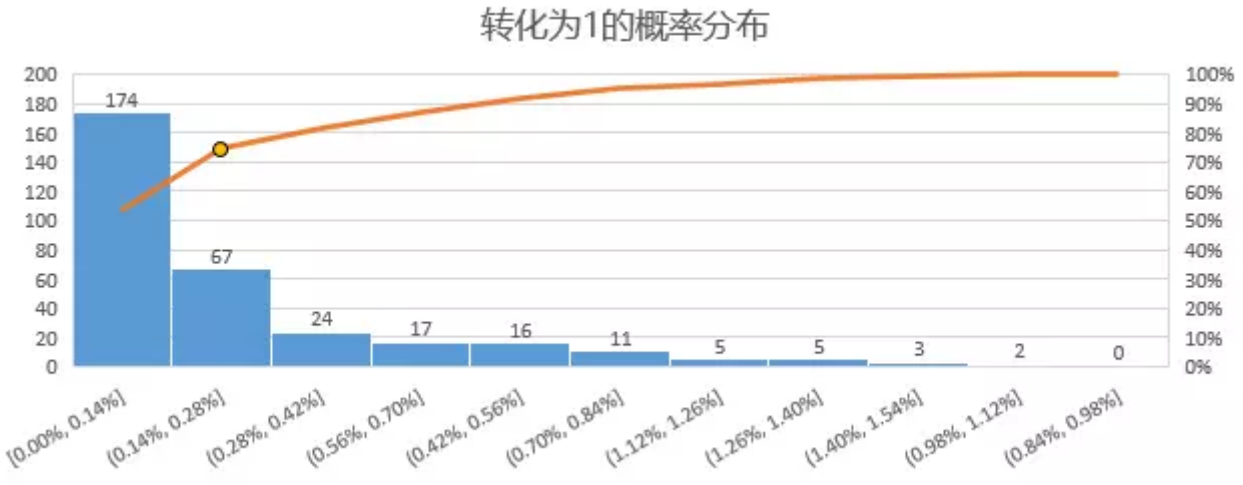
序号	转化为0	转化为1	转化&地级市&年龄	兴趣分类	转化为1&地级市&年龄&兴趣分类
209	98.49%	1.51%	12	11	转化&广州&(50岁以上)-母婴儿童
202	98.52%	1.48%	12	4	转化&广州&(50岁以上)-教育培训
207	98.54%	1.46%	12	9	转化&广州&(50岁以上)-房产
211	98.62%	1.38%	12	13	转化&广州&(50岁以上)-体育运动
203	98.64%	1.36%	12	5	转化&广州&(50岁以上)-旅游出行
212	98.68%	1.32%	12	14	转化&广州&(50岁以上)-生活服务
208	98.68%	1.32%	12	10	转化&广州&(50岁以上)-餐饮美食
210	98.70%	1.30%	12	12	转化&广州&(50岁以上)-科技数码
213	98.75%	1.25%	12	15	转化&广州&(50岁以上)-医疗健康
216	98.75%	1.25%	12	18	转化&广州&(50岁以上)-商务服务
200	98.84%	1.16%	12	2	转化&广州&(50岁以上)-家装百货
215	98.85%	1.15%	12	17	转化&广州&(50岁以上)-文化娱乐
205	98.87%	1.13%	12	7	转化&广州&(50岁以上)-汽车
201	98.95%	1.05%	12	3	转化&广州&(50岁以上)-金融理财
204	98.98%	1.02%	12	6	转化&广州&(50岁以上)-服饰箱包
101	99.21%	0.79%	6	11	转化&佛山&(50岁以上)-母婴儿童
317	99.22%	0.78%	18	11	转化&清远&(50岁以上)-母婴儿童
94	99.23%	0.77%	6	4	转化&佛山&(50岁以上)-教育培训
310	99.23%	0.77%	18	4	转化&清远&(50岁以上)-教育培训
99	99.24%	0.76%	6	9	转化&佛山&(50岁以上)-房产

。。。（中间太长，省略了）

80	99.96%	0.04%	5	8	转化&佛山&(41-50岁)-美容化妆
296	99.97%	0.03%	17	8	转化&清远&(41-50岁)-美容化妆
39	99.97%	0.03%	3	3	转化&佛山&(24-30岁)-金融理财
255	99.97%	0.03%	15	3	转化&清远&(24-30岁)-金融理财
42	99.97%	0.03%	3	6	转化&佛山&(24-30岁)-服饰箱包
258	99.97%	0.03%	15	6	转化&清远&(24-30岁)-服饰箱包
37	99.98%	0.02%	3	1	转化&佛山&(24-30岁)-游戏
253	99.98%	0.02%	15	1	转化&清远&(24-30岁)-游戏
44	99.98%	0.02%	3	8	转化&佛山&(24-30岁)-美容化妆
260	99.98%	0.02%	15	8	转化&清远&(24-30岁)-美容化妆
1	100.00%	0.00%	1	1	转化&佛山&(1-18岁)-游戏
2	100.00%	0.00%	1	2	转化&佛山&(1-18岁)-家装百货
3	100.00%	0.00%	1	3	转化&佛山&(1-18岁)-金融理财
4	100.00%	0.00%	1	4	转化&佛山&(1-18岁)-教育培训

我们看到，数据挖掘显示，转化为1的最大概率是1.51%，此时的定向条件是”广州&(50岁以上)-母婴儿童“。而样本数据的整体转化率是37/13339 = 0.28%。

下图是转化为1的概率分布，可以看到大于0.28%的数据约有25%。随着数据量的增加，模型也会不断改进，对精准定向组合的转化率预测效能也会越来越好，将有限的广告费花在最有可能转化的用户上。



下面考虑怎么将这一洞察，应用于广告投放，创造更高的ROI。比如制作针对性的创意、提高出价等等，这个方面各位都是老手了，我就不多说了。

最后强调一句，受众画像的数据挖掘需要满足一定的条件，即要能区分转化和未转化的用户。