机器学习系列|从白富美相亲看特征预处理与选择

大数据文摘

> 作者授权转载 作者：龙心尘&寒小阳 摘自：
> http://blog.csdn.net/longxinchen_ml/article/details/50471682,
> http://blog.csdn.net/han_xiaoyang/article/details/50481967

大数据文摘愿意为读者打造高质量【机器学习讨论群】，措施如下

（1）群内定期组织分享

（2）确保群内分享者和学习者数量适合，有分享能力者不限名额，学习者数量少于分享者，按申请顺序排序。

点击文末“阅读原文”填表入群

上篇

1. 引言

再过一个月就是春节，相信有很多码农就要准备欢天喜地地回家过（xiang）年(qin)了。我们今天也打算讲一个相亲的故事。

讲机器学习为什么要讲相亲？被讨论群里的小伙伴催着相亲，哦不，催着讲特征工程紧啊。只是我们不太敢讲这么复杂高深的东西，毕竟工程实践的经验太复杂了，没有统一的好解释的理论，一般的教材讲这方面的内容不多。我们就打算以一个相亲的故事为例，串一些特征工程的内容。

2. 故事背景

事先声明：本故事纯属虚构，如有雷同，纯属巧合！

> 海归白富美韩梅梅刚回国，还没适应工作，母亲就催着相亲。以父母的关系，他们了解到的适龄单身男青年有100个。要从100个男生中找到1个理想的女婿，可谓百里挑一。韩梅梅母亲也担心女儿相亲多了会反感，打算草拟一个相亲名单，人数不多。怎么从中挑出优秀男青年就是一个首要的问题。

3. 用机器学习的框架去分析

我们用机器学习的框架分析，在父母眼中，这100个男生最终将会分成两类：“女婿”（1人）和“非女婿”（99人）。“女婿”和“非女婿”就叫做“标签”。

而选择相亲名单的标准——如“是否高富帅”、“是否海归”等等——就叫作“特征”。最好能有一个特征能够精确定位理想女婿。但这太过理想了。比较现实的方法是从这些“特征”中选择、拆分、组合出最合适的特征，逐渐逼近我们的标签，以形成一个精简的相亲名单。而这个过程，就可以理解成特征处理、特征工程的过程。

但是，现实中的特征有千千万，拆分重组之后特征又是几何级数地增加，可能永远也穷举不完。因此需要有统一客观的指标来衡量这些特征对标签的识别能力，以便进一步地深入分析。而评估这些“特征”对我们的“标签”的有效程度的过程就叫作“特征有效性分析”。

4. 剧情一：韩妈妈的“如意算盘”

> 为人父母嘛，总是希望女儿嫁得好。韩妈妈的第一反应的就是要找“高富帅”。先她先从这100个男生中挑了挑，符合高富帅这个标准的有5个

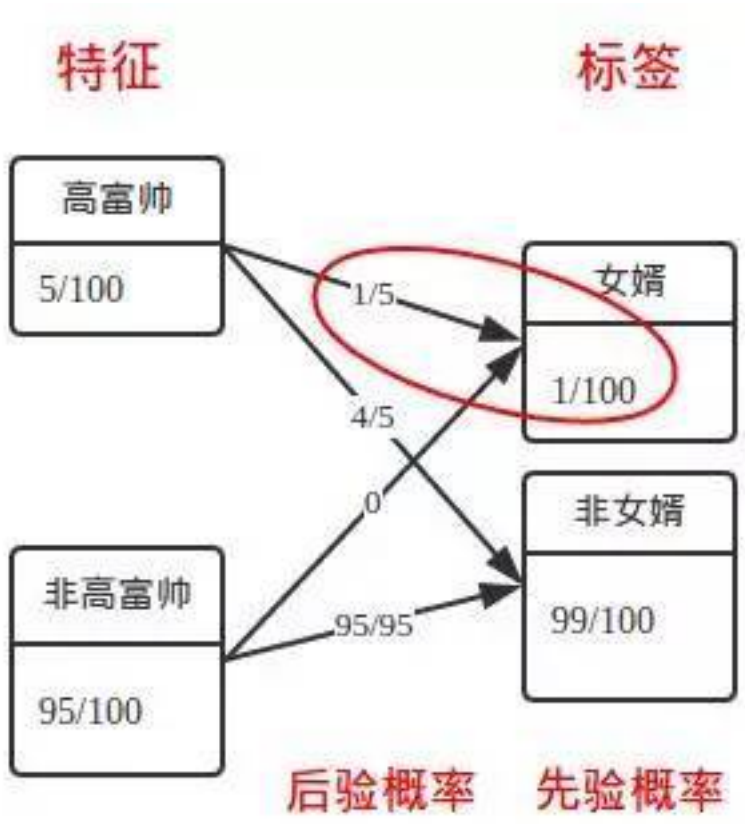
人。 韩妈妈的如意算盘是这样的：女婿就从这5个人中挑，概率就是20%，比之前的1%整整提高了20倍，嘿嘿嘿。。。

5. 特征有效性分析

其实，这就韩妈妈不知不觉就走了一个特征有效性分析的过程。我们用图表演示一下：

标签 / 人数 / 特征	女婿	非女婿
高富帅	1	4
不是高富帅	0	95

考虑到各方面的概率，用下图表示更加直观：



为了表述方便，我们以随机挑女婿而不考虑任何特征的概率叫做“先验概率”（1%）。而中间的箭头中的概率则表示在已经知道样本所属特征前提

下，属于女婿还是不属于女婿的概率，也可以叫作“标签相对于某个特征的后验概率”（20%）。而母亲的如意算盘就是考虑了上图中红圈部分的先验概率与后验概率（也可以叫条件概率）。这其实是一种很朴素的特征有效性分析的方法。而且她还做了个更加精确的数量化描述：

$$\frac{\text{后验概率}}{\text{先验概率}} = \frac{20\%}{1\%} = 20\text{（倍）}。$$

只是在工程上做除法可能运算会麻烦些，而两边同时取对数转换成减法则更方便：

$$log(\frac{\text{后验概率}}{\text{先验概率}}) = log(\text{后验概率}) - log(\text{先验概率})$$

概率表示着选女婿的可能性或者确定性。在本例中，后验概率的确定性比先验概率的确定性更高。可见，“确定性的增加”可以作为特征有效性分析的一个指标。

我们进一步分析，无论先验概率还是后验概率，其本身是0-1之间的一个数，取完对数之后是一个负数，这在现实中不太方便找到其对应的现象解释。但是概率的倒数一定大于1，取完对数之后就是一个正数，就好找现实解释了。我们可以把这个“概率倒数的对数”理解成不确定性的指标。于是上式就变成：

$$log(\frac{\text{后验概率}}{\text{先验概率}}) = log(\frac{1}{\text{先验概率}}) - log(\frac{1}{\text{后验概率}})$$

这里面的$log(\frac{\text{后验概率}}{\text{先验概率}})$我们叫做互信息。

因此，“不确定性的减少”可以作为特征有效性分析的一个指标。这个结论我们接下来将会反复用到。

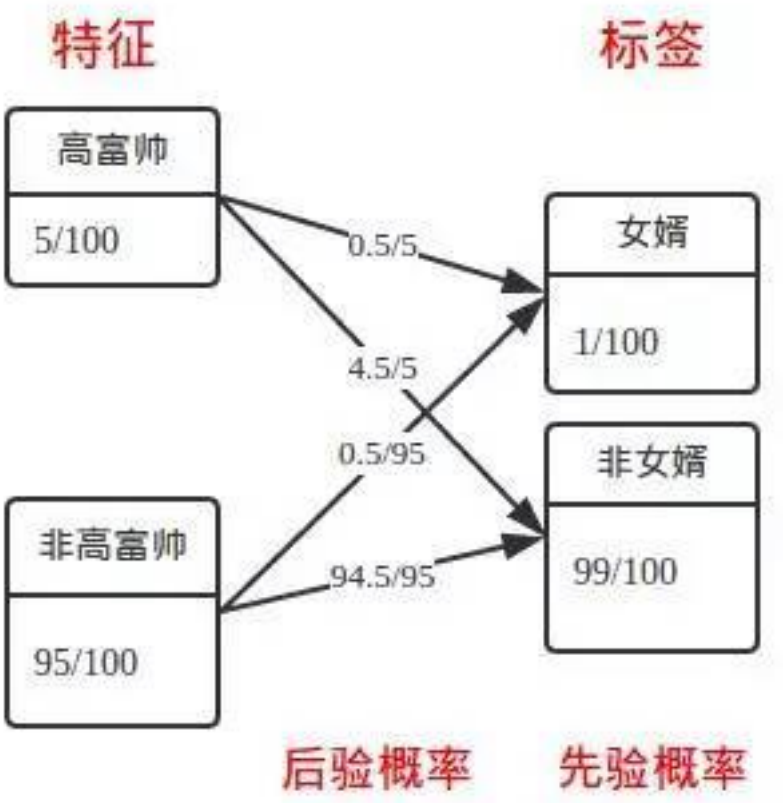
6. 剧情二：白富美巧劝慈母

> 韩妈妈半开玩笑地问韩梅梅：“我们家闺女只挑高富帅的怎么样？”女儿想了想，说：“如果人家看不上我们怎么办？”母亲笑着说：“我们的家境哪里差了？何况我们的女儿这么优秀，我们还看不上他们呢。”女儿说：“这就是说明我们双方不合适了。我们家条件虽然还不错，但是比下有余、比上不足，跟真正条件好的家庭比较起来我们根本不算事儿。如果一味挑高富帅，他们可能觉得我们只是看中他们的钱，反倒把我们家看低了。相反，要是真要有个真正对我好的男生，比什么都幸福，而他不一定必须是高富帅。毕竟跟我一起相处一辈子的是一个活生生的人，而不是他背后的东西嘛。”母亲很有感慨地说：“嗯，你能这样想我就放心了。梅梅真是长大了。那么，你打算怎么办？”女儿说道：“高富帅也得分人，踏实人品好的也可以接触一下，但是玩心太重不会照顾人的我就不喜欢。估计高富帅里面这两种人一半一半吧。很多男生并不是高富帅，其中没准也有合适的人呢。”

7. 特征有效性分析

现在特征的分布发生了新变化。按韩梅梅的分析，高富帅中可能有一半她就不会喜欢，而不是高富帅的男生中没准有合适的人。我们可以简单假设高富帅中与非高富帅中各有0.5个合适的人。则分析图表如下：

标签 / 人数 / 特征	女婿	非女婿
高富帅	0.5	4.5
不是高富帅	0.5	94.5



现在的情况是，“是不是女婿”的可能性同时分布在“是高富帅”和“不是高富帅”中，单独衡量“高富帅”本身的后验概率已经不够描述特征的整体效果了。我们可以有一个考虑特征整体情况的指标。

还是回到之前的那句话：

> “不确定性的减少”可以作为特征有效性分析的一个指标。

我们之前考虑了“是女婿”的不确定性是log(11%)，相应的“不是女婿”的不确定性是log(199%)，那么标签“是否女婿”作为整体的平均不确定性则可以理解为这两个状态的加权平均：

> $H(Y) = 1\% \times log(\frac{1}{1\%}) + 99\% \times log(\frac{1}{99\%}) = 0.08079$ 。(全文假定对数log的底数取为2)
>
> 这就是传说中的信息熵。我们用Y表示标签，用$H(Y)$表示“是否女婿”的信息熵，也就是其整体的平均不确定性。

那么考虑特征（“是否高帅富”）后的标签（“是否女婿”）的平均不确定性怎么衡量？我们用

X:{“是高富帅”,“不是高富帅”}来表示特征。其实，与上面的思路类似，我们在已知特征为“是高富帅”的前提下，“是否女婿”这个标签的整体平均不确定性可以用相对“是高富帅”的后验概率来求出：

$$H(Y|X=\text{“是高富帅”}) = (0.5/5)\times log(\frac{1}{(0.5/5)}) + (4.5/5)\times log(\frac{1}{(4.5/5)}) = 0.46900$$

在已知特征为“不是高富帅”的前提下，“是否女婿”这个标签的整体平均不确定性可以用相对“不是高富帅”的后验概率来求出：

$$H(Y|X=\text{“不是高富帅”}) = (0.5/95)\times log(\frac{1}{(0.5/95)}) + (94.5/95)\times log(\frac{1}{(94.5/95)}) = 0.04741。$$

因此，已知特征(无论具体是“是高富帅”还是“不是高富帅”)情况下的标签平均不确定性为前面两种情况的加权平均：

$$\begin{aligned}&H(Y|X)\\&= P(X=\text{“是高富帅”})\times H(Y|X=\text{“是高富帅”}) + P(X=\text{“不是高富帅”})\times H(Y|X=\text{“不是高富帅”})\\&= 5/100\times 0.46900 + 95/100\times 0.04741 = 0.06849\end{aligned}$$

这就是传说中的**条件熵**。

所以，考虑特征后，标签的“不确定性的减少”为：

$$I(Y,X) = H(Y) - H(Y|X) = 0.01230$$

这个$I(Y,X)$就叫做**平均互信息**。

我们用同样的方法去评价之前母亲设想的女婿只在高富帅中的理想情况（也就是女婿只在高富帅中产生的情况）的互信息

I（Y,X′）=0.04470

平均互信息从理想情况的0.04470下降到0.01230，也就是说原以为特征“是否高富帅”与标签“是否女婿”的相关性很高，后来发现相关性其实是比较低的。可见理想很丰满，现实很骨感。

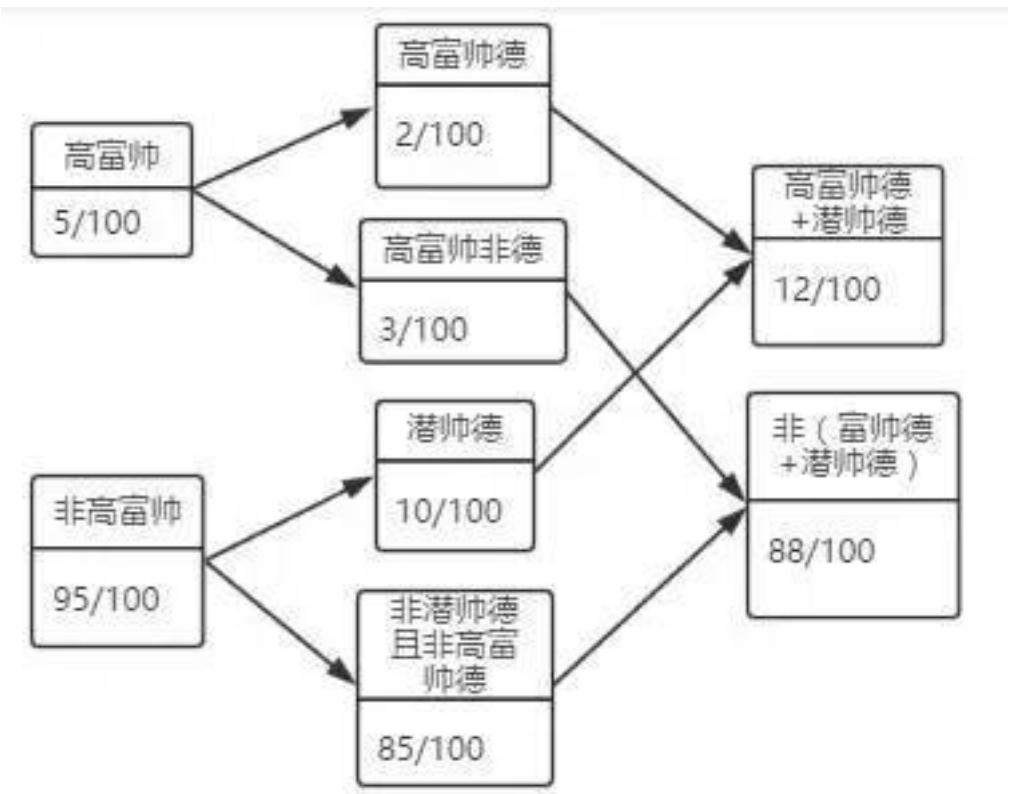
8. 剧情三：白富美重定名单

其实，韩梅梅没有说出来的话是她有一个青梅竹马的码农叫李雷。她出国之前的对他的印象还不错。如果按母亲的标准李雷肯定排除在相亲名单外了，而她想给他一个机会。 这时母亲说话了：“我们家女儿考虑得挺好，那相亲名单你来定吧。”女儿说：“不是高富帅的男生也该

> 好好区分一下，那些品行端正、气度不凡、踏实肯干的潜力股的男生我也比较欣赏，其他的就暂且不考虑了。”母亲说：“就是说可以从高富帅中挑出部分品德好的，还有从不是高帅富的男生中挑出部分潜力股，共同组成一个新的名单，我们的女婿就在这里面了?”女儿不好意思地说：“妈妈您真着急，八字还没一撇呢。”接着，韩梅梅母女俩从高富帅中挑了2个口碑不错的，又从不是高富帅的男生中条了10个很不错的。最终组成了12人的相亲名单。李雷的名字在其中。

9. 拆分重组成为新特征

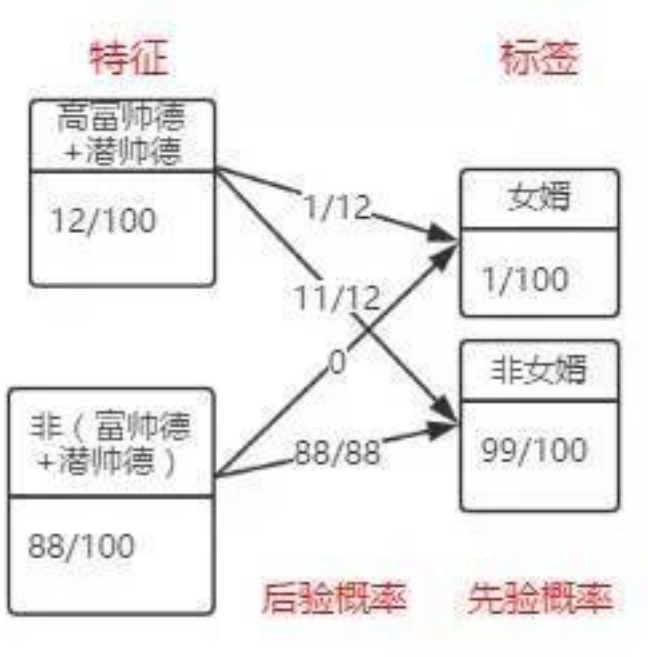
其实以上韩梅梅母女俩完成了一次特征的拆分与重组过程。具体图示如下:



这里用“潜帅德”表示韩梅梅对“品行端正、气度不凡、踏实肯干的潜力股”的特征的描述。

特征进行拆分与重组的过程在特征工程中经常出现。因为当你对特征与标签的相关性有定量的评估方法后，会筛选出那些不那么显著的特征（如本例中的“是否高富帅”），然后去分析考核指标这么低的原因，启发你引入新的特征（如本例中的“是否品德良好”、“是否有潜力”）将原有特征拆分重组，可能会有更好的效果。而这些生成的新特征，又要经过特征有效性分析来最终评估。如此反复迭代。

10. 特征有效性分析



我们用X2来表示新特征，与上面的思路类似，我们计算X2的平均互信息：

$$I(Y,X_2) = H(Y) - H(Y|X_2) = 0.03114$$

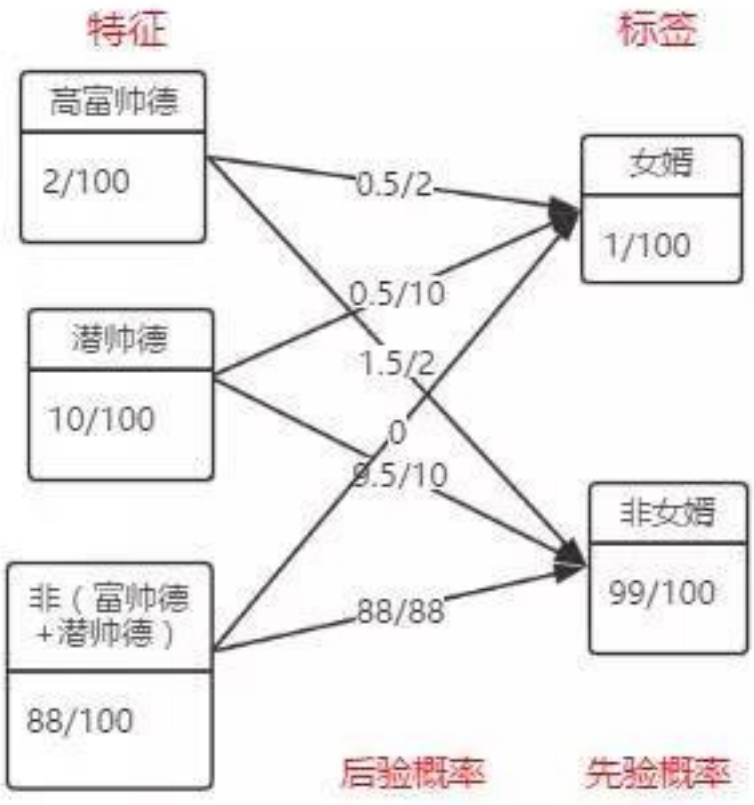
与之前的平均互信息

I（Y,X）=0.01230比起来，有了显著提高。可见新特征X2比之前的特征X更有效。

11. 剧情四:韩妈妈给名单分级

> 在跟韩梅梅聊完之后，韩妈妈转念一想:“为什么非要有一份相亲名单?可以把这12个人再分成两类，第一类是高富帅的，先相亲。这些觉得不合适后再考虑剩下的10个人啊。”

12. 特征有效性分析



我们继续分析，用X3来表示新特征，与上面的思路类似，我们计算X3的平均互信息：

$$I(Y,X_3) = H(Y) - H(Y|X_3) = 0.03593$$

与之前的平均互信息

I（Y,X2）=0.03114比起来，又有了一定的提高。可见新特征X3比之前的特征X2更有效。

韩妈妈真是为女儿的相亲操碎了心。

13. 剧情五：韩妈妈问计赵媒婆

> 韩妈妈思索完之后抑制不住内心的兴奋，想找人倾诉。这时她正好在路上碰见了赵媒婆。赵媒婆在韩妈妈的老闺蜜圈中享有盛誉，相亲非常有经验。 赵媒婆听了韩妈妈的诉说后，微微一笑，说：“你这个名单不够专业。”韩妈妈大为诧异。赵媒婆继续说：“高、富、帅三个特征本来就是相互独立的三个特征，你硬生生地绑在一起，多少大好青年被你给甩掉了。后面的潜力股啊、人品端正啊什么的都类似。”韩妈妈恍然大悟:“真是这样啊。” 赵媒婆说：“其实你这里最大的问题是这些特征的评估都是拍脑袋决定，没有充分的现实数据做支撑，很可能会犯错误的。”韩妈妈暗暗点头，心生佩服。 赵媒婆接着说：“还有一个

> 问题，你准备了两份名单，也就是把人群分成了三份，你算平均互信息只能评价整体的，具体到每一份人群你怎么对他们评价？”韩妈妈想了想，说：“我们可以直接用相对于某个具体条件的信息熵啊。”赵媒婆说：“何苦这么麻烦呢？”韩妈妈听她话里有话，打算继续问下去。

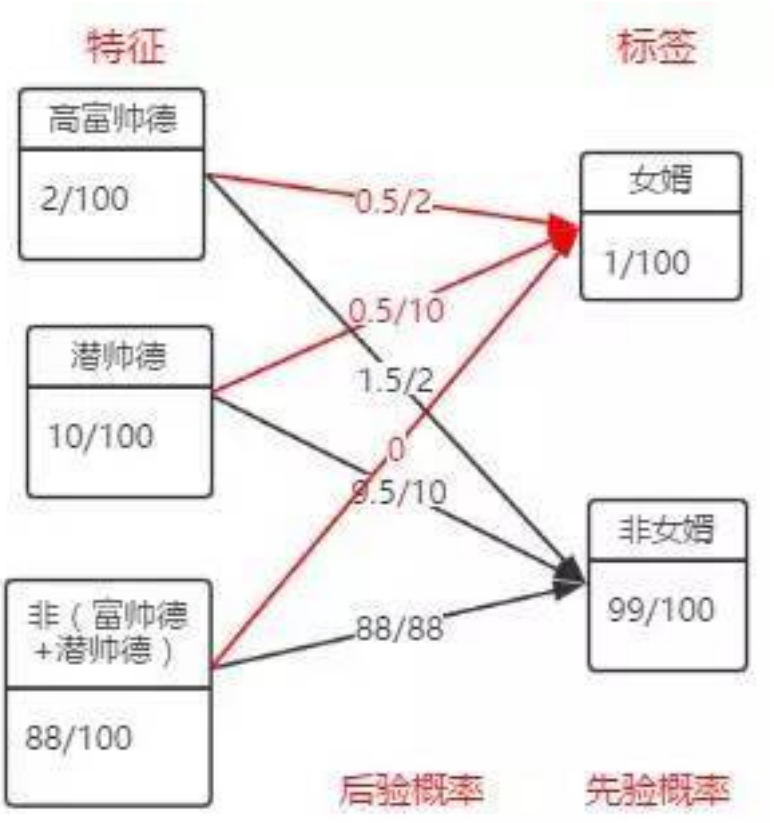
14. 评价特征选项的两个方法

在赵媒婆最后一个问题中，韩妈妈所说的其实是可以计算以下三个值来评估具体的特征选项：

$$H(Y|X_3 = \text{“高富帅德”})$$
$$H(Y|X_3 = \text{“潜帅德”})$$
$$H(Y|X_3 = \text{“不是高富帅德且不是潜帅德”})$$

而这三个值在之前计算条件熵H(Y|X3)的过程中就已经计算出来了。所以比较起来应该很方便。

但其实更简单的方法用他们相对于所需要标签的后验概率评价。如下图红色的部分，比较大小就可以找出评价较好的特征。



显然“高富帅德”的评分最高（0.25），“潜帅德”的评分次之（0.05），“不是高富帅德且不是潜帅德”评分最差（0）。符合韩妈妈的

预期。

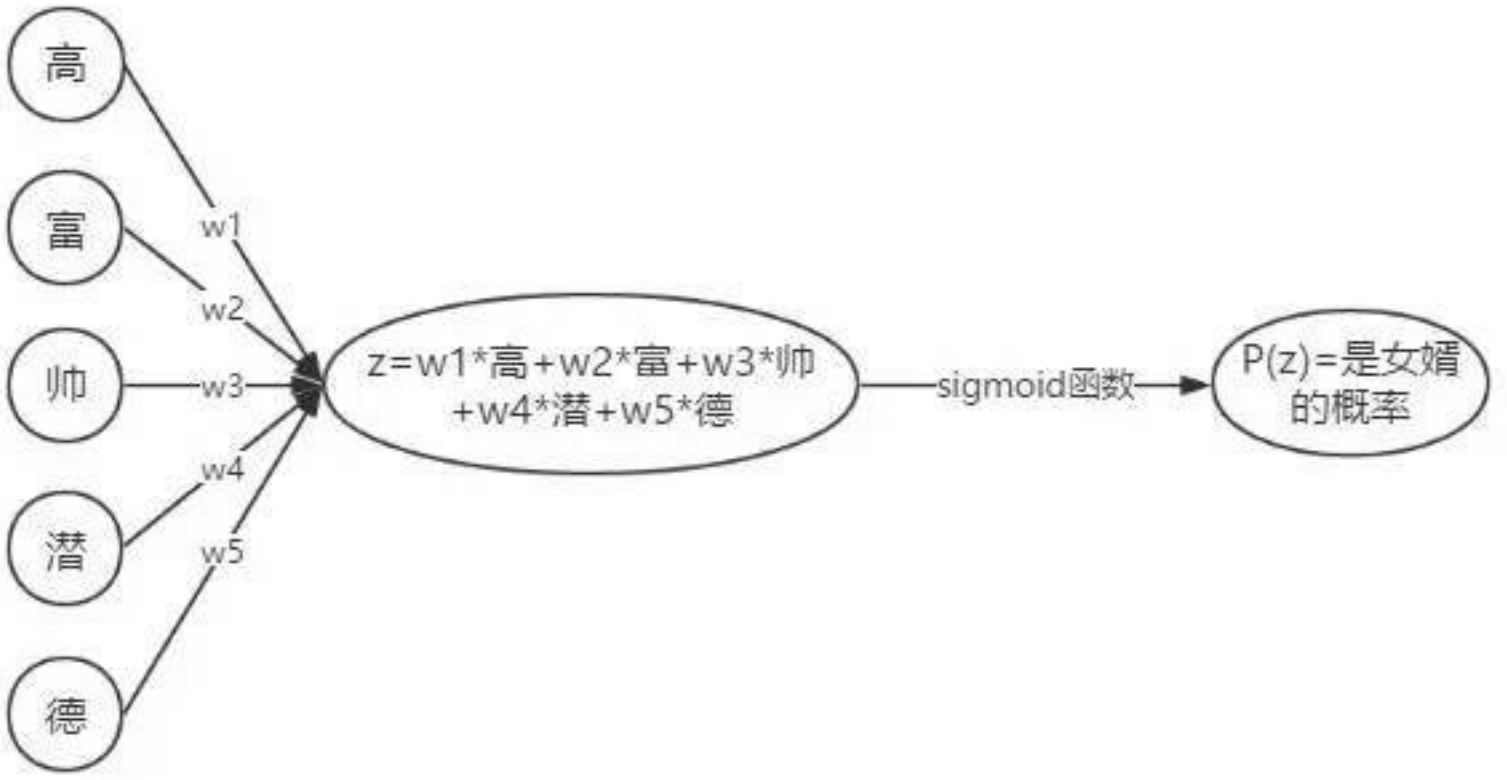
然而，赵媒婆的想说的并不是这种方法，而是逻辑回归.....

15. 剧情六：赵媒婆的数据库

> 赵媒婆不等韩妈妈说话，就直接拿出了自己的神器:一个平板电脑。然后打开她的相亲数据库，点了点鼠标，一张巨大的表展现出来。韩妈妈目瞪口呆：“现在媒婆都用高科技了？”赵媒婆傲娇地说:“那是。”这张大表是她这么多年来全国各地相亲介绍的所有男生信息，分别标注了每个男生的升高、年龄、年薪、长相特点、教育经历、工作经历、是否海归、工作年限、工作公司、工作地点、出身地、是否有户口、是否公务员、具体职业、行业、性格倾向等等信息。 她还有一张女生信息表，另外一张男生女生相亲情况表（相亲成功、相亲不成、继续发展、未接触）。媒婆一一给韩妈妈解释这些信息。韩妈妈连连惊呼。 赵媒婆接着说道：“我们可以从里面找出跟你女儿情况相近的一些女生信息，再把跟她们相过亲的男生找出来，把其中相亲成功的归为一类，剩下的归为另一类。然后假设男生的每个特征对相亲成功都有贡献，贡献的权重为wi。我们用逻辑回归的方法可以求出这些权重，把这些权重大的特征挑出来，你再用它们来找女婿就方便了。”韩妈妈说：“逻~辑~什么？”赵媒婆说：“高科技了，你不懂的。不过给我干儿子写了个博客来介绍，你可以看看。”

16. 特征筛选与特征工程工作流

呃，我们什么时候成赵媒婆的干儿子了?先不管这些。逻辑回归并不是什么高科技，在前面的文章里已有简单的解释。我们在这里就补充说明一下为什么可以用权重来衡量特征的贡献。以下是一个典型的逻辑回归过程：



我们期望P(z)的概率越大越好，sigmoid函数是个单调递增函数，所以z越大越好，在所有特征都归一化的前提下，显然是权重wi越大越好。因此与wi对应的特征就是我们要寻找的显著特征。而那些权重小的特征就可以先不考虑了。这就完成了一个最简单的特征筛选的过程。

当然，这里所说的权重大可以指的是权重的绝对值很大，比如特征“富”的权重是-100，是一个很小的数，但这也就意味着“不富”的权重会很大，以至于显著影响我们的z的结果。所以这也是一个显著特征。

需要补充一下的是，在工程实践中，权重的幅度和正则化也有关系。L1正则化会把特征拉稀疏，会产出一部分0特征。而不是0的那些特征，是有作用的特征。所以L1正则化其实具备一定的特征选择（feature selection）的作用。尤其是很高维空间的feature，用L1正则化，其实能帮助做一下feature selection的。而L2正则化，则会把各个维度的权重拉平均一些，抑制住各个维度权重幅度的方差。但是抑制归抑制，最后的权重还是会有大小差异，就像上文说的，绝对值大的权重，对应的特征区分度好一些。

对于那些不够显著的特征，我们需要分析一下这个特征的具体情况是怎样，是否需要对其进行重新拆分与重组，拆分重组后新的特征又可以进行特征有效性分析。如此不断迭代反复，就可以挑选出比较理想的特征了。

我们用以下整个工作流大致展现这个过程。由于很多内容没有展开，我们先把名字写进去，在后续的文章中继续扩展。

17. 剧情七：韩妈妈新名单尘埃落定

> 在韩妈妈与赵媒婆的尽心鼓捣下，最终生成了一个只有4个人的相亲名单。其中只剩下一名高富帅，另外三人中有一人正是李雷。韩妈妈拿着新名单给女儿看，韩梅梅沉默半晌，心想李雷在四人名单中怎么也能存在，莫非这也是缘分？

18. 小结

本文中主要讲了一些特征有效性分析的方法，包括用互信息，平均互信息，条件熵，后验概率，逻辑回归权重等方法对特征与标签的相关性进行了评估。有了这些评估做基础，可以筛选出显著的特征，并对对不显著的特征进行分析、拆分和重组，最终形成新的特征并反复迭代。本文略过了一些特征预处理的方法，并对特征有效性评估的阐述不够充分，我们将在接下来的部分予以讨论。

下篇

1、剧情一：挑螃蟹的秘密

> 李雷与韩梅梅的关系发展得不错，趁国庆休假一起来天津玩。今天，李雷十分神秘地请韩梅梅去一家餐馆吃螃蟹。韩梅梅大失所望，这个餐馆很不起眼，感觉就像路边的老食堂。菜单都用粉笔写在黑板上，一点都不高档。一看价格，满黄螃蟹120块钱一只!这也太贵了。 李雷看到了韩梅梅的神情，笑着解释道：“这家店老板有一个绝活——会看螃蟹。他能保证120块的螃蟹就是满黄。如果拆开来不是，这个螃蟹就不要钱，再换一个。靠着老板的绝活，这家店已经是几十年的老店了，在当地非常有名气。郭德纲、赵丽蓉这些天津社会名流都来这家店吃过螃蟹。”韩梅梅将信将疑。拆开螃蟹，饱满的蟹黄喷薄欲出。韩梅梅边吃边惊叹：“从没有吃个这么好吃的螃蟹！”李雷接着说:“老板的绝活密不外传，几十年来都自己上货。虽说是一个大老板，一年到头满身海鲜味。而且他也不开分店。”韩梅梅说：“那是，这么高明的绝活只有他自己知道才能挣钱啊。”这时，韩梅梅拂面而笑，突然想考一考自己的相亲对象，说：“李大码农，你不是做机器学习的吗？如果要你去用机器学习挑满黄的螃蟹，你怎么做？”

2. 初步划定特征的范围，获取特征

李雷早就想过这个问题了。长期的职业素养让他对任何事情都想用机器学习的方法去鼓捣。李雷的基本思路是这样的，我们尽可能观察螃蟹更多的特征，从中找出与“螃蟹满黄”最相关的特征来，帮助我们去判断。当然特征有非常多，我们可以先头脑风暴一下：

1. 一些直观的特征：包括蟹壳的颜色和光泽度、钳子的大小、肚脐的形状、螃蟹腿的粗细和长度、眼睛的大小和颜色光泽、螃蟹的品种、重量、体积、腰围等等.....
2. 一些需要在互动过程中观察到的特征：螃蟹钳子的力量，对外界刺激的反应，用筷子触碰螃蟹眼睛后的反应，螃蟹行动的速度.....
3. 还有一些外部环境的特征： 收获螃蟹的季节，培养螃蟹的水域.....

> 韩梅梅插话到：“这么多特征我头都大了，你还有完没完?”

其实，如果真要穷举出所有的特征可能永远也举不完。但是我们**目的很明确**——判断螃蟹是否是满黄。所以我们只关心跟这个问题（“标签”）相关的特征，它们只占所有特征中很小一部分。怕就怕一些糊涂的需求方连目的都不明确就要求一通乱搞，即便出来了一堆结果，也不知道有什么用。

头脑风暴完之后，很重要的一点就是找到对这个问题有长期经验的人，虚心向他们学习。人脑其实是一个很好的特征筛选器，这些经验可以给我们非常多的指导和启发，极大地减少我们试错的工作量。比如我们可以直接去找海鲜市场问螃蟹贩子，去田间地头找螃蟹养殖户，去海鲜饭店去问有经验的采购员和厨师.....他们的最一线的经验是特征工程中的宝贵财富。

但这里需要考虑将经验转换成可量化的指标，才能便于机器学习。比如人们可能会说螃蟹很“活跃”、很“精神”，或者很“慵懒”。这些特征需要转换成一些可量化指标去衡量，具体怎么转换也有很大学问。

接下来要考虑的问题是对这些特征的可用性进行简单的评估。比如：

1. 特征获取、描述难度
2. 数据的规模
3. 特征的准确率
4. 特征的覆盖率

5. 其他

我们通过明确目标，头脑风暴，咨询专家，特征量化，可用性评估等流程，就基本划定了特征范围。

3. 剧情二：“特征预处理”的门道

> 李雷说完，便拿出自己的平板，给韩梅梅看自己某个项目中搜集的初始特征。这些特征被放在一张巨大的表里。 韩梅梅看着这些密密麻麻的数字，心想：看李雷说得头头是道，但还是没告诉我怎么挑，不能让他轻易绕过去。 于是她说：“我看你这些特征数据有大有小，有些就是几万上下浮动，有些仅仅是小数点后好几位的微小变化，有些就是在0或1这两种可能中变化，有些连值都没有。你这些数据能用吗？”
> 李雷说：“不能，要转换成标准件。”韩梅梅：“标准件？”

4. “特征标准件”

如果把机器学习过程当做一个加工厂的话，那输入的数据（特征、标签）就是原材料，输出的模型和判断结果就是产品。并不是胡乱扔进去任何原材料都能加工出合格产品的。原材料需要一个“预处理”过程才能方便地被算法处理。这些预处理后的数据，李雷起了个不够规范的名字，叫做“特征标准件”。

以二分类问题为例，不同的算法对“特征标准件”的要求是不同的。比如逻辑回归和神经网络，比较喜欢归一化之后在[-1,1]区间内浮动的特征。而贝叶斯方法，喜欢因子化之后的{0,1}分布的二元特征，每个特征只有“是”和“不是”两种可能的状态。

5. 连续特征与非连续特征

特征可以分为两类：“连续特征”和“非连续特征”。

“身高”、“体重”、“成绩”、“腰围”、“长度”、“宽度”、“体积”、“速度”等等，都是连续特征。连续特征能够比较方便地进行归一化。归一化的统一公式如下：

$$x*=x-\mu S$$

μ为所有样本数据的均值，x−μ的步骤叫做去均值化

1. 当S=xmax−xmin时，经过处理的数据在区间[-1,1]之间。
2. 当S=σ（所有样本的标准差）时，经过处理的数据符合标准正态分布，即均值为0，标准差为1

另一方面：“是否高富帅”、“是否白富美”、“螃蟹的品种”、“螃蟹所在的水域”、“收获螃蟹的季节”等等，都是非连续特征。非连续特征能够比较方便地进行因子化，或者它本身就是二元特征。方法如下：

> 特征“收获螃蟹的季节”：{春，夏，秋，冬} **因子化**后的结果为：

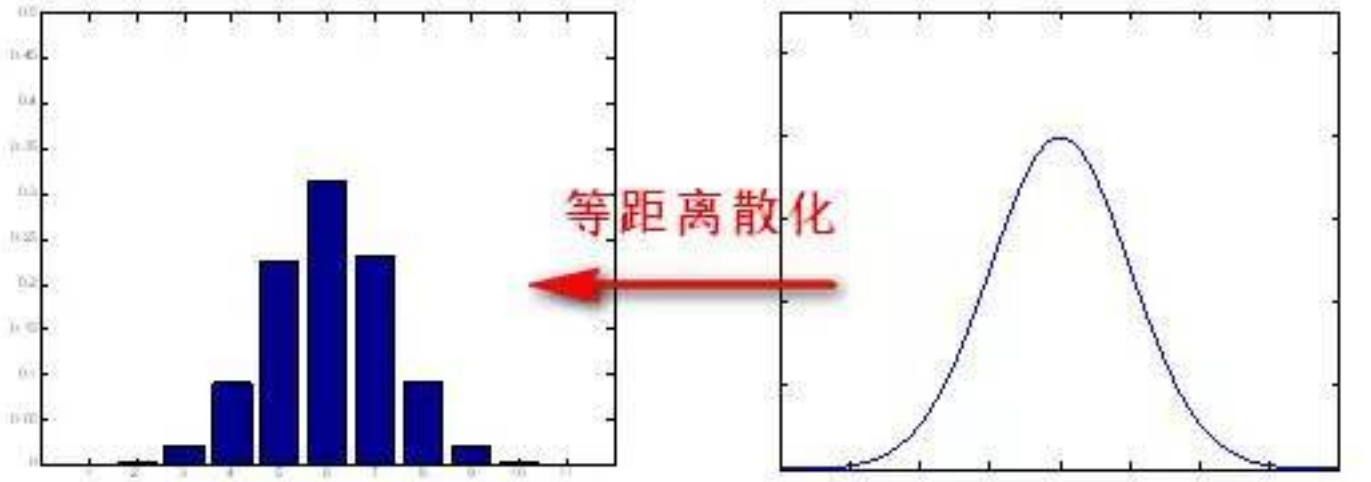
- 特征“是否春”：{是，否}
- 特征“是否夏”：{是，否}
- 特征“是否秋”：{是，否}
- 特征“是否冬”：{是，否}

6. 两类特征的相互转化

连续特征可以当非连续特征来用，非连续特征可以当连续特征来用。

连续特征可以离散化非连续特征。比如“年龄”作为一个连续特征，假设它的取值范围是[0,100]。我们可以中间切一刀，比如选择60（岁）。大于等于60岁的就叫做“老年”，小于60岁的就是“非老年”，这样就转化成了一个二元特征了。怎么选择离散的分离边界也很有学问。

如果我们中间切两刀甚至更多刀，比如18（岁）和60（岁）。大于等于60岁的就叫做“老年”，18岁到60岁之间的就叫做“中青年”，小于18岁就叫做“未成年”。然后再把这3类因子化成3个二分类就够了：“是否老年”、“是否中青年”和“是否未成年”。



非连续特征因子化成二元特征{0,1}后可以直接当做[0,1]之间的连续特征来用。

7. 去除特征之间的共线性

我们在对离散特征因子化过程中细分到二元特征为止即可。那对于二元特征本身能否因子化成两个特征？比如以下例子：

> 特征“螃蟹的性别”：{公，母}，可否转换为：

- 特征“是否公螃蟹”：{是，否}
- 特征“是否母螃蟹”：{是，否}

这是不行的，因为这两个特征的信息完全一样，也叫做共线性。计算这两个特征之间的条件熵：

> H(“是否公螃蟹”|“是否母螃蟹”)=0

也可以用计算条件熵的方法去衡量两类离散特征的差异性，方便去除共线性关系的特征。

连续特征也有着共线性的情况，比如同一个品种的螃蟹腿的“长度”和“粗细”是共线性关系。也就是说，如果我们知道螃蟹腿的长度是x厘米，那么螃蟹腿的直径就是kx厘米，k是一个稳定的常数。因此我们只需要螃蟹腿的“长度”这一个特征就够了。那么连续特征的共线性如何去除?

可以计算两个变量(x,y)的相关系数：

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{cov(x,y)}{\sqrt{cov(x,x) \times cov(y,y)}}$$

r_{xy}的取值范围是[-1,1]，如果是0则统计独立，如果接近1则强相关。

可以计算这些数据的**协方差矩阵**，进而求出相关系数矩阵。就可以比较任意两个特征了。

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

$$R = \begin{bmatrix} r_{11} & r_{12} & \Lambda & r_{1p} \\ r_{21} & r_{22} & \Lambda & r_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ r_{p1} & r_{p2} & \Lambda & r_{pp} \end{bmatrix}$$

其中 $$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}}$$

既然协方差矩阵都求了，那就干脆用主成分分析（PCA）吧，这样更省事。得到主成分，线性相关的那些量就直接被舍弃了。我们在前文《深度学习与计算机视觉系列(7)_神经网络数据预处理，正则化与损失函数》 对PCA有相关论述。

感兴趣的同学可以试试把上述离散二元特征当做连续变量使用，构造几个数据，计算其相关系数并进行主成分分析。发现其相关系数就是-1，主成分分析后自动就变成一个主成分了。可见PCA对于连续特征与非连续特征都是去除共线性的通用方法。

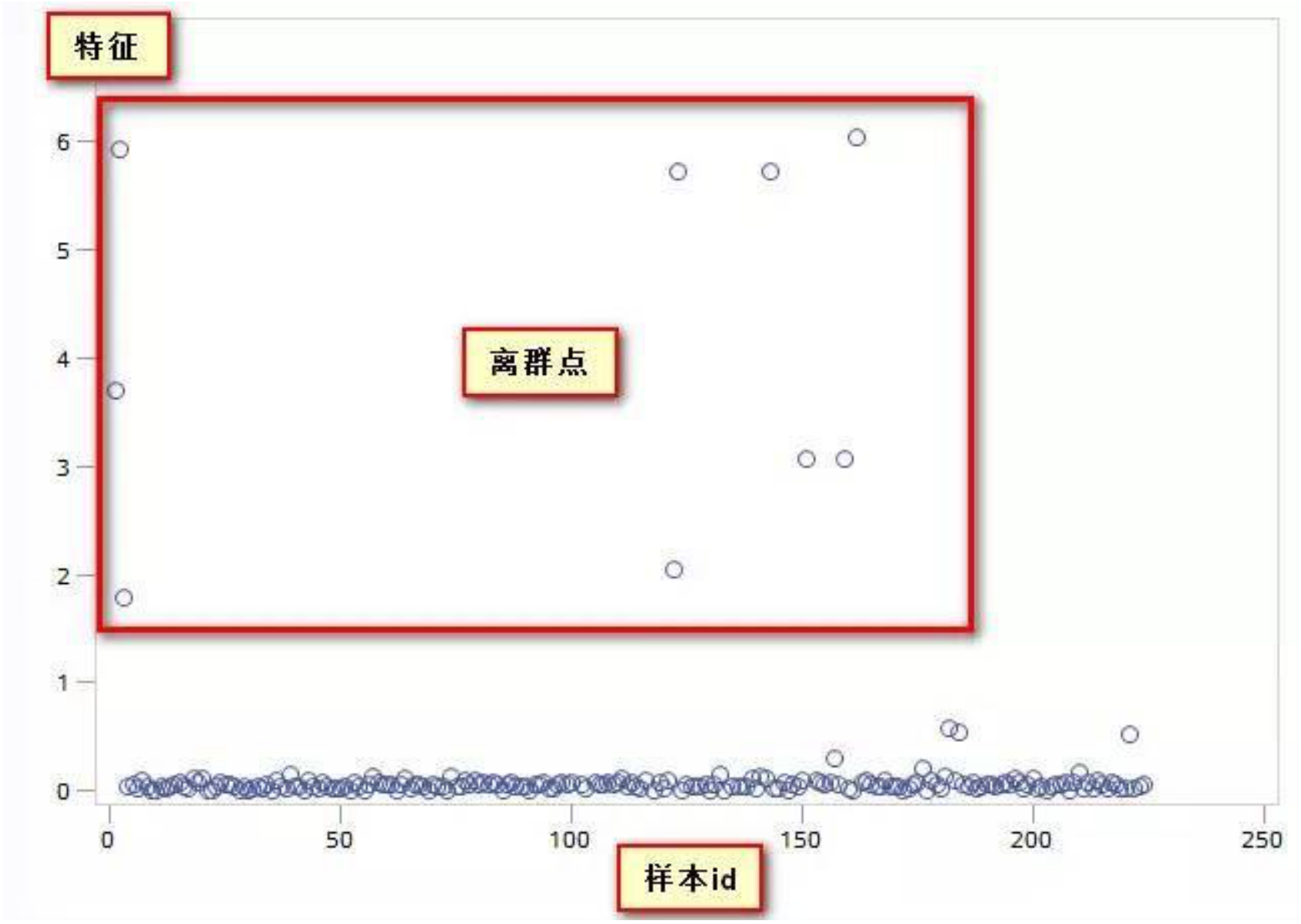
8. 缺失值的处理

这个问题现在才讲，但实际过程中应该在前期去处理。掌握以下三点就够了：

1. 如果某个特征的缺失值比较多：可能就直接舍弃。
2. 如果缺失值不是很多，而且是连续特征：可以考虑用回归方法去拟合，或者直接用众数、中位数、平均数等具体的值去替代即可。
3. 如果缺失值不是很多，而且是非连续特征：可以尝试把缺失值当做一个新的类目去处理，可能也揭示了一定的客观现实。

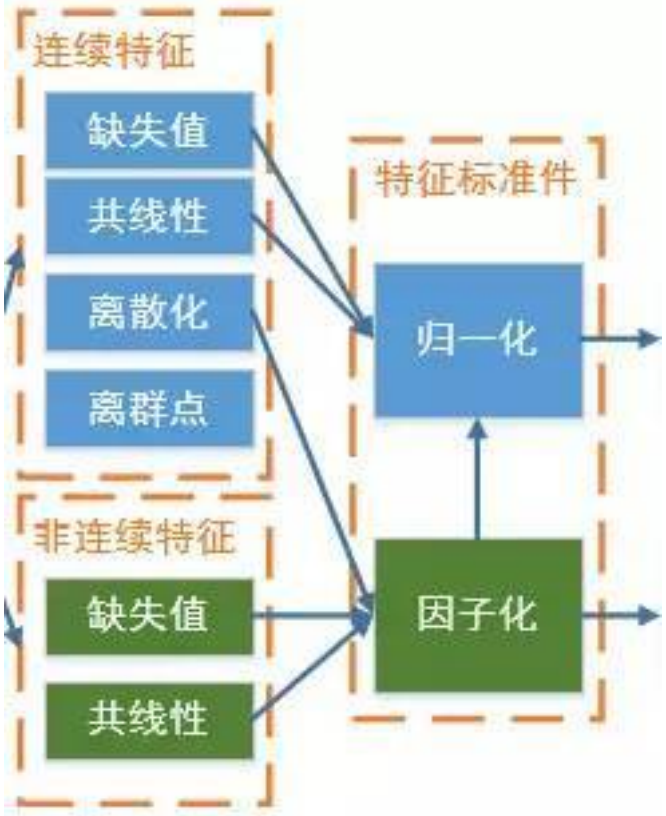
9. 离群点的分析

对于连续特征，最好看看其在样本中的分布。如果某些值偏离了主要聚集区域，可能需要单独抽出来分析，里面可能包含了更多的信息，可以这样画图方便观察：



10. 特征预处理小结

特征的预处理步骤比较多，相互之间的关系比较复杂。我们画了一张图以揭示它们之间的关系：



11. 剧情三：李雷另辟蹊径挑螃蟹

> 韩梅梅长叹一口气：“终于听你叨逼叨逼说完了。”李雷说：“没办法啊，这块工作其实挺多的。我还要好多没说完.....”“你打住”，韩梅梅赶紧说，“我算服了你了。但是李大码农，你还没有跟我说你怎么靠这些特征挑螃蟹呢。”李雷说：“不急，用逻辑回归.....”韩梅梅说：“不要用逻辑回归，我已经从赵媒婆那知道了。你换个方法，用非连续特征来做。”韩梅梅存心想刁难她的相亲对象。 李雷说：“那可以用贝叶斯。”

12. 用贝叶斯方法挑螃蟹

我们的标签用Y=“是满黄”来表示，相应的Y=“不是满黄”。Xi表示所有离散化的二元特征，如X1=“是河蟹”，X2=“是秋季收货”，X3=”钳子的力量大”.....。于是在已知这些特征的情况下，该螃蟹“是满黄”的概率如下：

$$P(Y|X_1, X_2, X_3 \dots) = \frac{P(X_1,X_2,X_3 \dots|Y) \times P(Y)}{P(X_1,X_2,X_3 \dots)}$$

其实，可以直接判断P(Y|X1,X2,X3...)是否大于1/2即可。因为

P(X1,X2,X3...)这一项算起来比较麻烦，我们用以下方法直接把它约掉。

先求出螃蟹“不是满黄”的概率：

$$P(\overline{Y}|X_1, X_2, X_3 \dots) = \frac{P(X1,X2,X3...|\overline{Y})\times P(\overline{Y})}{P(X1,X2,X3...)}$$

再两式相处，得到：

$$\frac{P(Y|X1,X2,X3...)}{P(\overline{Y}|X1,X2,X3...)} = \frac{P(X1,X2,X3...|Y)\times P(Y)}{P(X1,X2,X3...|\overline{Y})\times P(\overline{Y})}$$

这样就约去了P(X1,X2,X3...)。只需要判断

$$\frac{P(Y|X_1,X_2,X_3...)}{P(\overline{Y}|X_1,X_2,X_3...)}$$

是否大于1即可。但是，工程上用除法不太方便，两边同时取对数log，得到：

$$log\frac{P(Y|X1,X2,X3...)}{P(\overline{Y}|X1,X2,X3...)} = log\frac{P(X1,X2,X3...|Y)}{P(X1,X2,X3...|\overline{Y})} + log\frac{P(Y)}{P(\overline{Y})}$$

> 左边是螃蟹“是满黄”的**逻辑发生比，只需要判断其是否大于0即可。**

到目前为止，以上都是等价变换。

接下来我们引入贝叶斯方法中常用的条件独立假设：

$$P(X_1, X_2, X_3 \ldots | Y) = P(X_1|Y) \times P(X_2|Y) \times P(X_3|Y) \ldots$$
$$P(X_1, X_2, X_3 \ldots | \overline{Y}) = P(X_1|\overline{Y}) \times P(X_2|\overline{Y}) \times P(X_3|\overline{Y}) \ldots$$

将它们带入上式，就变成了：

$$log\frac{P(Y|X_1,X_2,X_3...)}{P(\overline{Y}|X_1,X_2,X_3...)} = log\frac{P(X_1|Y)}{P(X_1|\overline{Y})} + log\frac{P(X_2|Y)}{P(X_2|\overline{Y})} + log\frac{P(X_3|Y)}{P(X_3|\overline{Y})} + \ldots$$
$$+log\frac{P(Y)}{P(\overline{Y})}$$

于是我们得到了一个简单的求和式，只需要判断等式右边求和的结果是否大于0即可。而最关键的就是右边每一项都非常好求！假如训练集中所有的满黄螃蟹收集在一起，统计每一个特征出现的次数，除以满黄螃蟹的总数，就是其相应的条件（后验）概率了。再统计该特征在非满黄螃蟹集合中的条件（后验）概率，二者相除再取对数即可。

13. 用贝叶斯方法进行特征有效性分析

等式右边作为一个求和式，其中每个求和项

$$log\frac{P(X_i|Y)}{P(X_i|\overline{Y})}|$$

的绝对值越大，其对结果的影响越强烈，相对应的特征就是显著特征。而绝对值比较小的特征就是非显著特征，剔除掉也不会很明显地影响结果。这就完成了一个特征筛选的过程。

我们再分析一下各个求和项的结构，里面的概率部分是后验概率，是特征相对于标签的后验概率。

14. 贝叶斯与逻辑回归之间的关系

我们继续看看这个求和项，是不是很像逻辑回归中的求和项？我们如果拿二元特征当做连续变量采用逻辑回归方法。其判别式如下：

$$z = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + b; \text{其中} x_i \in \{0, 1\}$$

二者的表达式惊人地相似！莫非

$$log\frac{P(X_i|Y)}{P(X_i|\overline{Y})} = w_i$$

，二者一模一样?

感兴趣的同学可以自己举个例子试一下，发现还是有区别的，二者求出来的权重不一样。产生这样差别的原因是什么呢?

想必大家都猜到了。就是贝叶斯方法引入的两个条件独立假设。正因为这两个条件独立假设，贝叶斯方法直接跳过了逻辑回归中反复迭代用梯度下降法才能求出的各个权重。

因此贝叶斯方法与逻辑回归的区别就是贝叶斯方法引入了一个更强的附加假设，而且可以直接通过统计结果求权重，而不必用梯度下降法。

所以有些情况下贝叶斯方法求出来的结果不好，就可以考虑考虑是不是条件独立假设的原因。

因此，可以说“在某种假定下，可以证明：与朴素贝叶斯分类方法一样，许多神经网络和曲线拟合算法输出最大的后验假定。”——韩家炜：《数据挖掘：概念与技术（第三版）》

15. 剧情四：李雷露馅儿了

> 韩梅梅听完，十分感慨地说：“难怪机器学习能挑出正确的结果，难怪

> 赵媒婆用机器学习方法从这么多人中能把你挑出里来。你还是有两下子嘛。”“废话，她是我干妈”，李雷志得意满，不小心说漏嘴。 韩梅梅：“什么？！”李雷后悔不已，尴尬地陪着笑脸说道：“梅梅，我错了，我不该瞒你这么久。”到了这个地步，李雷只能和盘托出了。

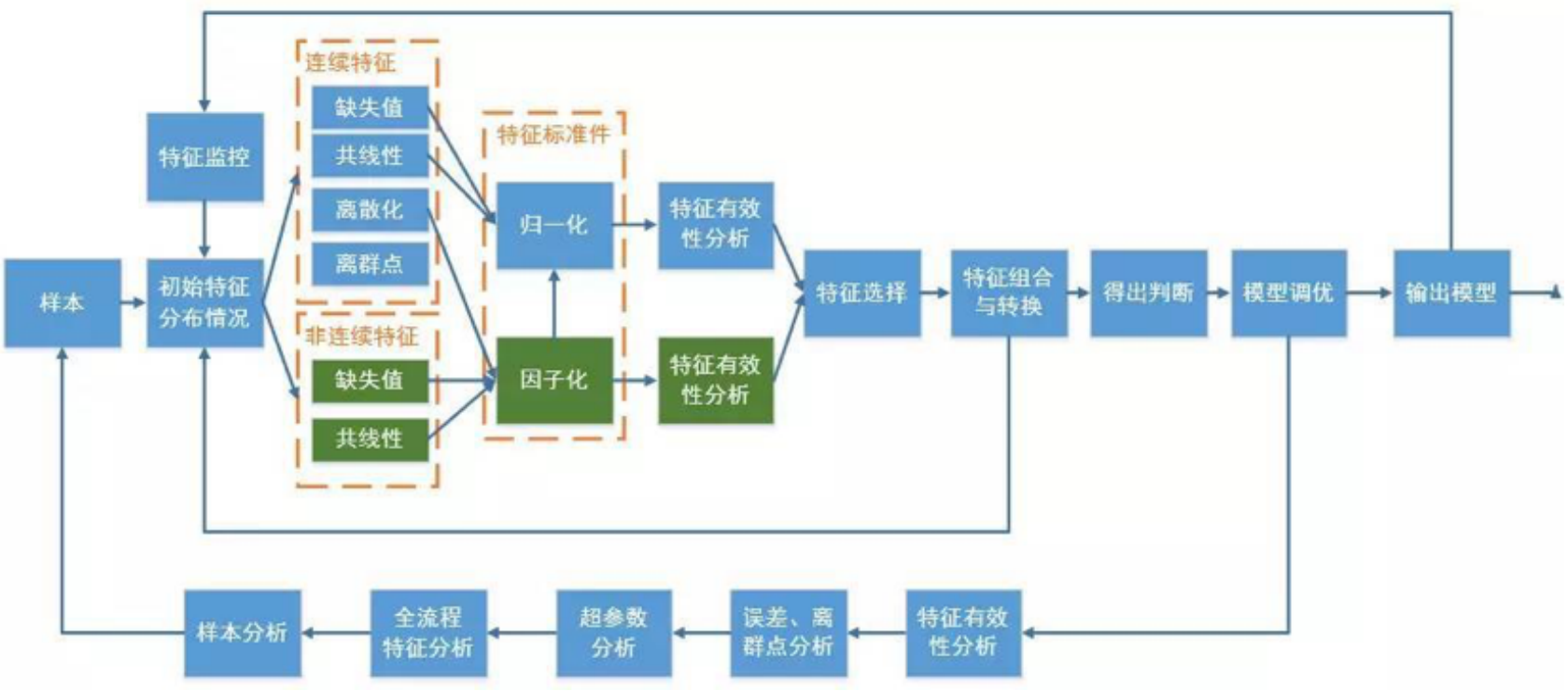
16. 数据VS算法

其实李雷早就知道韩妈妈要挑选相亲名单，如果按她的标准，李雷根本没法进入名单中。而李雷也猜想她会去找赵媒婆。他就早早地联系赵媒婆，跟她推销他的机器学习方法。赵媒婆终于被李雷忽悠动心了。李雷就帮她开发那个相亲算法。但其实赵媒婆的样本数量不够，特征数量却非常多，肯定会过拟合。李雷就跟她说他会多找一些相亲的数据。李雷能从哪里找啊，只能发动周围的同学，让他们找他们观察到的情侣案例。而在这些群体中，恰好中学、大学是同学的情侣比率非常高，而且很多男方是码农。而李雷刚好符合这个条件，李雷的评分就非常高了。

因为样本选择本来就代表性不足，没能覆盖更多的青年群体，所以还是过拟合，只是偏向了李雷这边的概率而已。

可见，做机器学习虽然看起来比较炫酷的是算法，但真正关键的是数据。数据决定了你结果的上限，而算法只是尽可能逼近这个上限。而这点李雷并没有告诉赵媒婆。

对样本情况的分析要在特征优化过程中尤其注意。整个流程图如下：



17. 特征选择的局限性

而且，李雷并不觉得感情这样复杂的东西能够用赵媒婆那些量化的指标衡量好的。房子、车子、学历、文凭这些并不能衡量两个人之间的感情。一些非常重要的特征是难以量化的，比如两个人的“三观”、两个人对待感情的态度、两个人相互相处的独一无二的经历、两个人刻骨铭心的情感体验、那种两个人相信能够一辈子都在一起的笃定的感觉.....这些至关重要的特征极其复杂，却非常难以量化。所以对于这类问题，机器学习的能力还是很有限的。

18. 剧情五：尾声

> 韩梅梅听完李雷，既生气，又好笑，还有一点点小感动：这小子为了感情还是蛮拼的。 一段沉默之后，韩梅梅笑着对李雷说：“好了好了，我不怪你了。”李雷长舒一口气。 韩梅梅继续说：“问个挑螃蟹的问题。你刚才选了这么多特征。为什么不考虑用B超直接照一下，看看里面什么东西不就成了吗？”李雷一听，犹如当头一棒，整个脑子都被草泥马占满了：“我去，这么简单的方法我怎么想不到？！”韩梅梅这时已经笑得肚子痛了，根本说不上话。 李雷吐槽到：“梅梅，你太厉害了。我觉得机器永远也学不到的两样东西就是人类的情感和脑洞啊！”

19. 后记

其实博主也没有丧心病狂到抓只螃蟹去照B超，只是自己被这个想法逗乐了，大家开心就好哈。O(∩_∩)O~~

如果真要死磕，据说B超的穿透力比较弱，对骨骼、空气等很难达到深部，因此难以成像。但是通过声波的回声来判断，也是一个思路。就像有些人可以通过拍打西瓜听声音来判断它甜不甜的道理一样。

如果不用机械波而用电磁波，比如X射线，估计哪怕能看到螃蟹满黄顾客也不会吃了。顾客也会担心放射残留的。CT应该好些，但是贵呀。一套设备下来，螃蟹估计也不止120块钱了吧。没玩过CT，不知道成本多少.....总之还是要考虑获取特征的成本的。

感谢作者授权转载

作者介绍：

龙心尘和寒小阳：从事机器学习/数据挖掘相关应用工作，热爱机器学习/数据挖掘

『我们是一群热爱机器学习，喜欢交流分享的小伙伴，希望通过“ML学分计划”交流机器学习相关的知识，认识更多的朋友。欢迎大家加入我们的讨论群获取资源资料，交流和分享。』

联系方式：

龙心尘 johnnygong.ml@gmail.com

寒小阳 hanxiaoyang.ml@gmail.com

原文发布于微信公众号 - 大数据文摘（BigDataDigest）