10 | 特征预处理

2018-06-26 王天一

机器学习40讲 进入课程 >

讲述：王天一
时长 17:04 大小 7.91M

华盛顿大学教授、《终极算法》（The Master Algorithm）的作者佩德罗·多明戈斯曾在 Communications of The ACM 第 55 卷第 10 期上发表了一篇名为《机器学习你不得不知的那些事》（A Few Useful Things to Know about Machine Learning）的小文，介绍了 12 条机器学习中的“金科玉律”，其中的 7/8 两条说的就是对数据的作用的认识。

多明戈斯的观点是：数据量比算法更重要。即使算法本身并没有什么精巧的设计，但使用大量数据进行训练也能起到填鸭的效果，获得比用少量数据训练出来的聪明算法更好的性能。这也应了那句老话：**数据决定了机器学习的上限，而算法只是尽可能逼近这个上限。**

但多明戈斯嘴里的数据可不是硬件采集或者软件抓取的原始数据，而是经过特征工程处理之后的精修数据，**在他看来，特征工程（feature engineering）才是机器学习的关键。**通常来说，原始数据并不直接适用于学习，而是特征筛选、构造和生成的基础。一个好的预测模

型与高效的特征提取和明确的特征表示息息相关，如果通过特征工程得到很多独立的且与所属类别相关的特征，那学习过程就变成小菜一碟。

特征的本质是用于预测分类结果的信息，特征工程实际上就是对这些信息的编码。机器学习中的很多具体算法都可以归纳到特征工程的范畴之中，比如使用 L_1 正则化项的**LASSO 回归**，就是通过将某些特征的权重系数缩小到 0 来实现特征的过滤；再比如**主成分分析**，将具有相关性的一组特征变换为另一组线性无关的特征。这些方法本质上完成的都是特征工程的任务。

但是今天，我将不会讨论这些，而是把关注点放在算法之外，看一看**在特征工程之前，数据的特征需要经过哪些必要的预处理（preprocessing）。**

特征缩放（feature scaling）可能是最广为人知的预处理技巧了，它的目的是**保证所有的特征数值具有相同的数量级。**在有些情况下，数据中的某些特征会具有不同的尺度，比如在电商上买衣服时，身高和体重就是不同尺度的特征。

假设我的身高 / 体重是 1.85 米 /64 公斤，而买了同款衣服的两个朋友，1.75 米 /80 公斤的穿 L 号合适，1.58 米 /52 公斤的穿 S 号正好。直观判断的话，L 码应该更合适我。可如果把（身高，体重）的二元组看作二维空间上的点的话，代表我自己的点显然和代表 S 码的点之间的欧式距离更近。如果电商不开眼的话，保不齐就会把 S 码推荐给我。

实际上，不会有电商做出这么弱智的推荐，因为他们都会进行特征缩放。在上面的例子中，由于体重数据比身高数据高出了一个数量级，因此在计算欧式距离时，身高的影响相比于体重是可以忽略不计的，起作用的相当于只有体重一个特征，这样的算法自然就会把体重相近的划分到同一个类别。

特征缩放的作用就是消除特征的不同尺度所造成的偏差，具体的变换方法有以下这两种：

标准化（standardization）：$x_{st} = \dfrac{x - mean(x)}{stdev(x)}$

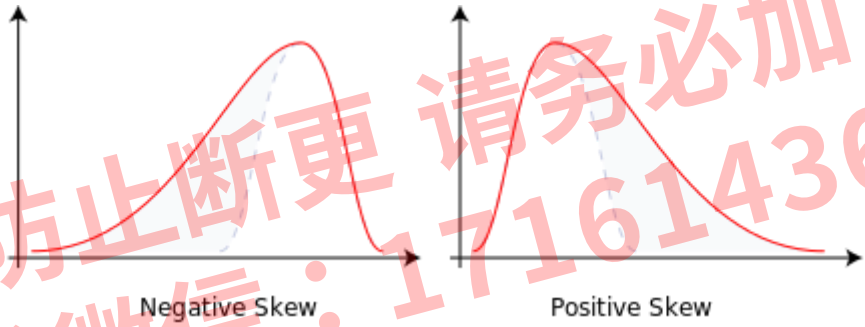
归一化（normalization）：$x_{norm} = \dfrac{x - \min(x)}{\max(x) - \min(x)}$

不难看出，**标准化的方法用原始数据减去均值再除以标准差，不管原始特征的取值范围有多大，得到的每组新数据都是均值为 0，方差为 1**，这意味着所有数据被强行拉到同一个尺度

之上；**归一化的方法则是用每个特征的取值区间作为一把尺子，再利用这把尺将不同的数据按比例进行转换，让所有数据都落在 [0, 1] 这个范围之内。**虽然实现方式不同，但两者都能够对数据做出重新标定，以避免不同尺度的特征产生不一致的影响，可谓殊途同归。

除了尺度之外，数据的偏度也是值得关注的一个问题。**偏度（skewness）是用于描述概率分布非对称性的一个指标。**下图给出了两个分别具有**负偏度**和**正偏度**的概率分布示意图，从中可以看出具有偏度的分布的形状都是类似的：一侧是瘦高的形状，占据了概率分布的大部分，另一侧则是比较长的拖尾。

想要理解这个图形所表示的概率分布，只要把正偏度的图形想象成你所在单位的工资分布就可以了：左侧的瘦高形状表示了拿着低工资的绝大部分普通员工，右侧的拖尾则表示了工资更高、但人数更少的中层领导和高级主管。无论机关、事业单位还是企业，工资的分布大抵都是这样。



不同偏度的概率分布（图片来自维基百科）

数据服从有偏分布意味着什么呢？意味着数据当中可能存在着**异常点**（outlier）。30 个维秘模特的体重应该近似地服从正态分布，而正态分布是无偏的对称分布。可是如果把其中一个模特的体重换成相扑运动员的体重，这个数据集的均值就会产生明显的上升，数据的直方图也会朝新均值的反方向产生明显的偏移。这时，偏度就体现为少量异常点对样本整体的拉拽作用，类似于用一个董事长和 99 个普通工人计算平均工资产生的喜剧效果。

面对偏度较大的数据，第一反应就应该是检查是否有异常点存在。一般来说，如果少量数据点和其他数据点有明显区别，就可以认为是异常点。在处理异常点时，首先要检测这些数据的**可靠性**，判断异常取值是不是由错误或者失误导致，比如那个混进维秘模特里的相扑选手。

如果异常点本身并没有问题，需要考虑的下一个问题就是异常点和正常点是否**来源于不同的生成机制**，从而具有不同的概率分布。如果对异常点所在的分布的采样数据较少，就不足以

体现出分布的特性，导致单个数据点看起来显得突兀。

对于像决策树这类对异常点比较敏感的算法来说，不管来源如何，异常点都需要被处理。最直接的处理办法就是**将异常点移除**，但当数据集容量较小时，这种一刀切的方式会进一步减少可用的数据，造成信息的丢失，这时就需要采用名为“**空间标识**”（spatial sign）的数值处理方法。

空间标识方法先对所有的数据点进行前面提到的标准化处理，再用样本向量的 2 范数对样本中的所有特征进行归一化，其数学表达式可以写成

$$x_{ij}^{*}=\frac{x_{ij}}{\sum_{j=1}^{N}x_{ij}^{2}}$$

式中的 N 是数据的维度。显然，**空间标识算法将所有的数据点都映射到高维空间的球面上，这个映射和标准化或者归一化的不同之处在于它处理的对象并不是所有样本的同一个特征，而是同一个样本的所有特征，让所有样本呈现一致的尺度。**

当然，即使在没有异常点的情况下，数据依然可能呈现出有偏的分布，这在数字图像处理中并不罕见。有偏分布的一个明显特点是最大值和最小值之间相差较大，通常可以达到**20 倍**或者更高。

这种数据尺度的不一致即使出现在单个特征上也不是一件好事情，对它进行修正，也就是**对数据进行去偏度处理的常用方法就是取对数变换（log transformation）**，也就是对特征取值取对数。最大值和最小值之间的 20 倍差距经过对数变换后变为 $\log_2 20=4.3$，这就在一个可以接受的范围内了。除了对数之外，**求平方根和求倒数也是移除偏度的常见处理方式。**

异常点也好，尺度不一致的数据也好，它们至少还都是完整的数据。可有些时候，一个样本里的某些特征会压根儿没有取值，而是一片空白，这种情况被称为**缺失值**（missing values）。

数据缺失的可能原因多种多样，在这里就不做展开了，关键还是在于如何处理这些缺失值。最简单粗暴的办法依然是将不完整的数据全部删除，对小数据集来说这依然不是好办法。更主动的处理方式是**给这些缺失值进行人为的赋值（imputation）**，就像数值计算或者信号处理中的插值方法一样。

人为赋值相当于在机器学习中又嵌套了一层机器学习，里层的机器学习被用于估计未知的属性值，也要使用训练数据。最常用的赋值算法是**K 近邻算法**：选取离具有缺失值的样本最近的 K 个样本，并以它们对应特征的平均值为缺失值赋值。此外，**线性回归**也可以用来对缺失值进行拟合。但可以确定的是，不管采用什么方法，人为赋值都会引入额外的不确定性，给模型带来的性能造成影响。

会做加法也要会做减法，缺失的数据需要添加，多余的数据也要删除。**在模型训练之前移除一些特征有助于增强模型的可解释性，也可以降低计算中的开销。**如果两个特征之间的相关性较强，或者说具有**共线性**（collinearity），这时就可以删除掉其中的一个，这正是**主成分分析**的作用。

除此之外，如果某个特征在绝大多数数据中的取值都是相同的，那这个特征就没有存在的意义，因为它体现不出对于不同分类结果的区分度。这就像在学校里，老师给所有同学的出勤都打满分，这部分平时分是拉不开成绩差距的。

什么样的特征不具备区分度呢？这里有两个经验性的标准：**一是特征取值的总数与样本数目的比例在 10% 以下**，这样的特征在 100 个样本里的取值数目不超过 10 个；**二是出现频率最高的特征取值的出现频率应该在出现频率次高的特征取值频率的 20 倍以上**，如果有 90 个样本的特征取值为 1，4 个样本的特征取值为 2，其余取值的样本数目都在 4 个以下，这样的特征就可以被删除了。

今天我和你分享了在模型训练之前对数据特征进行预处理的一些指导性原则，其要点如下：

特征缩放可以让不同特征的取值具有相同的尺度，方法包括标准化和归一化；

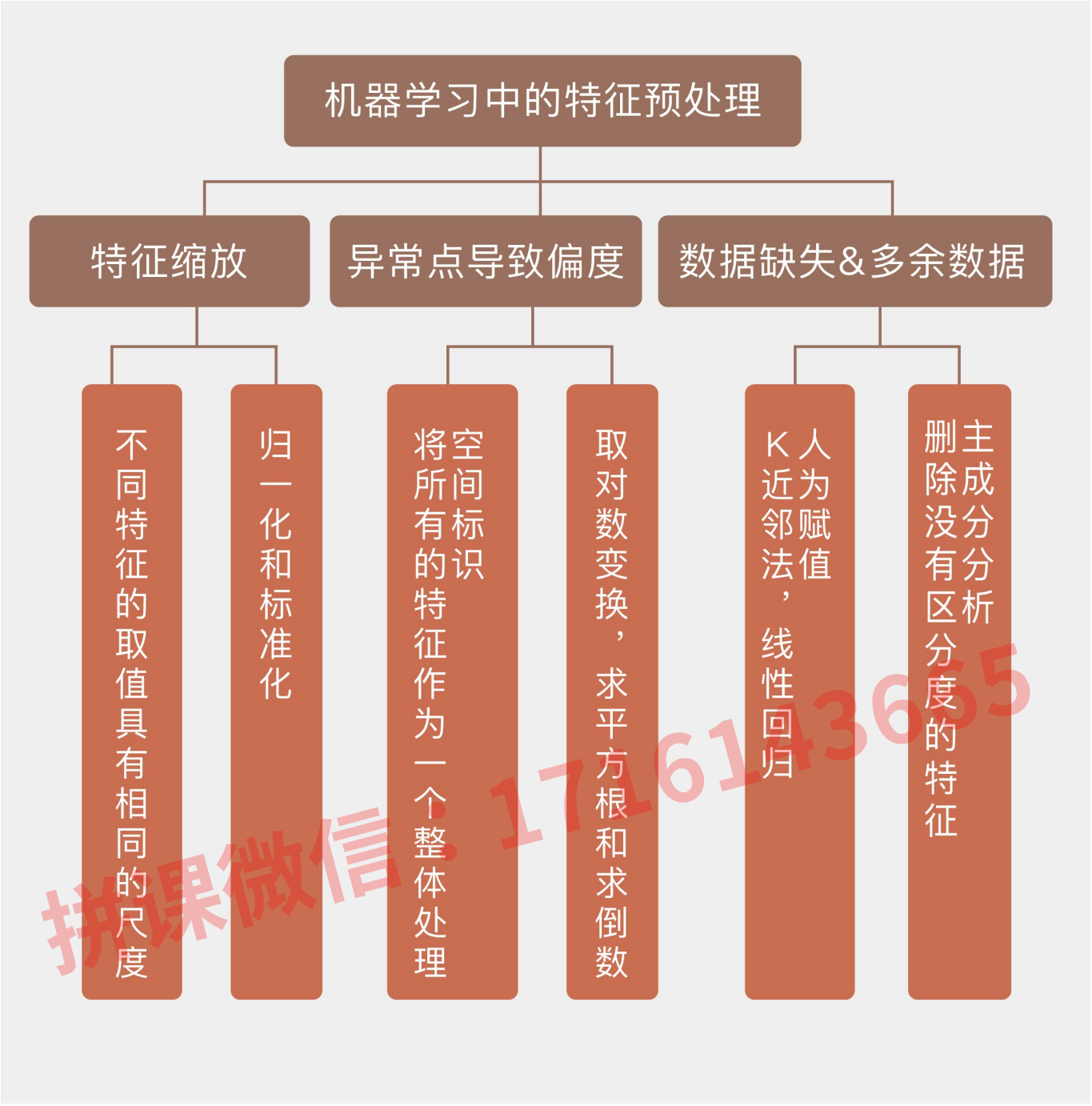
异常点会导致数据的有偏分布，对数变换和空间标识都可以去除数据的偏度；

k 近邻方法和线性回归可以用来对特征的缺失值进行人为赋值；

删除不具备区分度的特征能够降低计算开销，增强可解释性。

这里介绍的特征预处理技巧可以说是挂一漏万。那么在实际的任务当中，你遇到过哪些不理想的特征数据，又是如何处理的呢？

欢迎分享你的经历。



上一篇　09 | 实验设计

下一篇　11 | 基础线性回归：一元与多元

精选留言 (12)

写留言

曾珍
2018-09-16

1

空值我是用独热编码的方式，好想处理结果比线回归填充好一点

展开

作者回复: onehot确实比较常用

林彦
2018-06-26

1

特征尺度不一致还是挺常见的。用的是文中提到的标准化方法。缺失值的K近邻和插值方法

以前实践中只知道信号处理里有插值的函数，其他领域还没用过。

展开

作者回复: 各个学科的思想其实都是相通的。

暴走的carr...

2019-01-13

对于处理缺失值，以前我只知道用平均值或众数来代替，现在学会了，还能内嵌一个机器学习算法来处理缺失值，突然高端了好多

Daryl

2019-01-10

有个入门的问题，麻烦帮我解答下。
1:对训练集标准化/归一化/pca后，是否也要对测试集执行同样操作?
2:如果同样的操作，是直接对测试集transferm()，还是fit_transferm()?
3:标准化/归一化/pca 怎么针对数据集选择用哪种方式?

Kevin.zh...

2018-12-26

作业：

前段时间在通过爬虫程序获取了原始数据，在数据清洗的阶段，发现有很多的缺失数据，还有重复数据，重复数据之前没有使用pandas，就直接用的SQL筛选，对于缺失数据，我采用的笨办法，就是直接观察是哪个特征缺失，然后进行最原始的人工赋值替换操作，说实话，工作量大还不靠谱！边做心里还边打鼓！我不知道如何采用线性回归和K近...

展开

黄海娜

2018-11-25

老师，空值怎么用独热编码的方式呀?

展开

作者回复: 空值是指属性没有取值？这属于不完整数据了吧。那就丢弃或者人为估计一个值赋给它。

Y024
2018-10-11

https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
文中所提小文链接

作者回复: 感谢分享

五岳寻仙
2018-09-23

老师好！在删除不具备区分度的特征时，老师讲到：

什么样的特征不具备区分度呢？这里有两个经验性的标准：一是特征取值的总数与样本数目的比例在 10% 以下，这样的特征在 100 个样本里的取值数目不超过 10 个；二是出现频率最高的特征取值的出现频率应该在出现频率次高的特征取值频率的 20 倍以上，如果有...
展开

作者回复: 这个问题应该这么理解：性别这个特征本来就只有2个可能的取值，所以相当于每个值都取到了。如果说某个特征可能的取值范围是所有的正整数，但数据里只有1 2 3这三个，这才是文章里所说的情况。
另一个角度看，在性别这个特征上，如果100个数据里98个是男的，这样的特征也没什么意义。

Geek_40512...
2018-06-27

在用随机森林模型的时候，我们能知道每棵树在不同layer的具体特征变量名字吗?

作者回复: 当然可以，每棵树每个节点用来分裂的特征都是随机选择的。

我心飞扬
2018-06-27

做标准化之后有负数不能log了 是不是先log

展开

作者回复: 有负数时可以通过减最小值再加1把所有数据变成正数；也可以取绝对值做对数，再对负数得到的结果乘以-1。但这些都是纯数学的处理，线性操作可能会破坏数据的统计特性，所以还是选择其他的方法吧。
既然已经做了标准化，数据的尺度就应该基本一致了，为什么还要做对数呢？

我心飞扬
2018-06-27

请问空间标识和log的方法是要一起用吗？还是说。有负数就不能用log，这时候怎么办？如果统一把他加成正数，这样合理吗？会不会对分析产生一些误导呢。

展开

作者回复: 空间标识是把异常点拉成正常的，log是处理单个特征取值范围过大的，两个解决的不是一个问题。所以结合起来用原则上可以，关键还是在于要达到什么目的。

rkq@geekba...
2018-06-26

关于特征缩放我有一个问题：如果我的模型是普通的线性回归，需要对特征做缩放处理吗？我的理解是不需要，因为最终学得的参数就会体现出特征的缩放。不知道对不对？

展开

作者回复: 是的，线性回归并不直接基于距离，所以缩放与否计算出的参数和误差会有区别，但对整体趋势不会有太大的影响。