链滴

特征的表现方式

作者：moloee

原文链接：https://ld246.com/article/1520075021952

来源网站：链滴

总的来说，数据并不总是以规范的特征向量的形式呈现的，一般是按数据库记录、 协议缓冲区或其他何形式。所以必须要根据这些数据来创建特征向量。

实践中，机器学习从业人员将大约75%的时间花在特征工程方面了； 特征就是我们要的东西。

有以下几种方式或者几点需要注意：

1. 用one-hot的方式去编码 枚举值
2. 一个特征至少应该具有非0值，而且在数据集中出现过多次
3. 特征最好具有清晰明确的意义
4. 特征值不应随时间而变化
5. 特征值不应具有非理性的离群值，去除离群点，采用分箱技术

下面简单对上面几点进行介绍

1. 如果有可以直接作为特征值的还好，比如age=20, 够买次数=15等，但对于有些指标如street="wall street", brand="qiaopai" 等，就不太好规范化，可以用one-hot的方式编码。用一组特征向量, [0,0,0 ... , 1,0,0]这样，对可能出现的枚举值，当前值为1，其他为0来有效处理
2. 如果某个特征都是0值，或者绝大多数都是0值，则可能不是一个好的特征，在预处理的时候可以去掉
3. 明确的定义，比如age=25，而不是用小时或者秒来计算
4. 有些特征的值可能是从上游模型传导下来的，这时候应该尽量保证特征的值不会随时间改变或者漂移
5. 去除离群点、完全不合理的点，对于特征的训练没有价值；分箱技术是类似直方图的方式，虽然纵没有直接的可比性或者可编码的方式，可以用直方图分割区间，每个区间有类似的特性，然后再用one hot的方式编码。