机器学习——异常检测

> 在生产生活中，由于设备的误差或者人为操作失当，产品难免会出现错误。然后检查错误对人来说又是一个十分琐碎的事情。利用机器学习进行异常值检测可以让人类摆脱检错的烦恼。

检测算法

- 1.选定容易出错的nn个特征$\{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}$ {x1(i),x2(i),…,xn(i)}作为变量。
- 2.计算m个样本的平均值和方差。

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

μj=1m∑i=1mxj(i)

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2$$

σ2=1m∑i=1m(xj(i)−μj)2

-

3.给定监测点xx.计算p(x)p(x)

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$$

p(x)=∏j=1np(xj;μj,σj2)

-

4.如果$p(x) < \epsilon$p(x)<ϵ,则为异常值；反之，不是。

开发和评价一个异常检测系统

异常检测算法是一个非监督学习算法，意味着我们无法通过结果变量判断我们的数据是否异常。所以我们需要另一种方法检测算法是否有效。当我们开发一个系统时，我们从有标签（知道是否异常）的数据入手，从中找出一部分正常数据作为训练集，剩余的正常数据和异常数据作为交叉检验集和测试集。

具体评价方法如下：

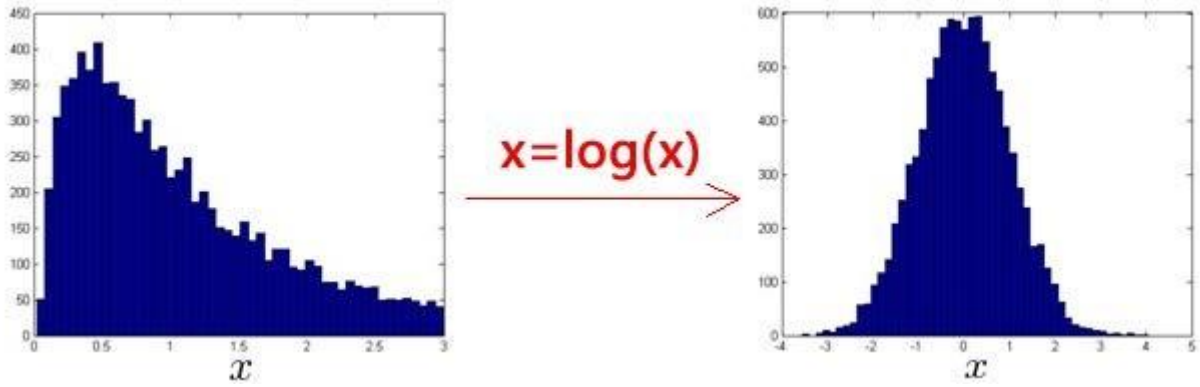
- 根据测试集数据，估计出特征的平均值和方差，构建p(x)p(x)函数
- 对于交叉检验集，尝试使用不同的ϵϵ最为阈值，并预测数据是否异常，根据F1值或者查准率与查全率的比例来选择ϵϵ
- 选出ϵϵ后，针对测试集进行预测，计算异常检验系统的F1值或者查准率与查全率之比

异常检测与监督学习对比

异常检测	监督学习
大量的正常值（y=0）和少量的异常值(y=1)	大量的正向类（y=0）和少量的负向类(y=1)
异常数据太少，只能根据少量数据进行训练	有足够多的正向和负向数据以供训练
举例：1.欺诈行为检测；2.生产废品检测；3.检测机器运行状态	举例：1.邮箱过滤器；2.天气预报；3.肿瘤分类

分布的处理

- 对于高斯分布的数据，直接运用以上算法就好。
- 但是对于非高斯分布的数据，虽然也可是使用上面的算法，但是效果不是很好，所以我们尽量将非高斯分布转化成（近似）高斯分布，然后再进行处理。
- 数据整体偏小，可以求$\ln(x)$ ln(x)或者$x^a, 0 < a < 1$xa,0<a<1
- 数据整体偏大，可以求e^x ex或者$x^a, a > 1$xa,a>1



误差分析

在误差分析中，如果我们可以发现我的选定的变量是否合适，进而进行相应的改正。如左图所示，异常点xx对应的概率很高，显然这种分布方式不能很好地识别出异常值。所以我们尝试增加变量或者改变变量的类型来识别异常值。如右图所示，通过增加一个变量，我们能够更好地识别异常点。所以，误差分析对于一个问题来说还是很重要的。