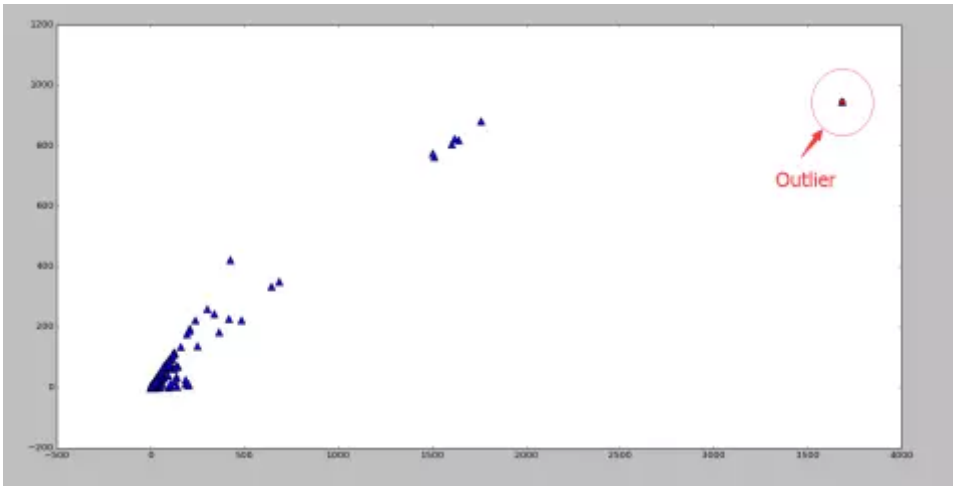
异常点检测算法（一）

原创 2016-06-23 张戎 数学人生

异常点检测（又称为离群点检测）是找出其行为很不同于预期对象的一个检测过程。这些对象被称为异常点或者离群点。异常点检测在很多实际的生产生活中都有着具体的应用，比如信用卡欺诈，工业损毁检测，图像检测等。

异常点（outlier）是一个数据对象，它明显不同于其他的数据对象，就好像它是被不同的机制产生的一样。例如下图红色的点，就明显区别于蓝色的点。相对于蓝色的点而言，红色的点就是异常点。



一般来说，进行异常点检测的方法有很多，最常见的就是基于统计学的方法。

（一）基于正态分布的一元离群点检测方法

假设有 n 个点 $(x_1, \ldots, x_n)$，那么可以计算出这 n 个点的均值 μ 和方差σ。均值和方差分别被定义为：

$$\mu = \sum_{i=1}^{n} x_i/n,$$

$$\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2/n.$$

在正态分布的假设下，区域 $\mu \pm 3\sigma$ 包含了99.7% 的数据，如果某个值距离分布的均值 μ 超过了 3σ，那么这个值就可以被简单的标记为一个异常点（outlier）。

（二）多元离群点的检测方法

涉及两个或者两个以上变量的数据称为多元数据，很多一元离群点的检测方法都可以扩展到高维空间中，从而处理多元数据。

（1）基于一元正态分布的离群点检测方法

假设 n 维的数据集合形如 $\vec{x_i} = (x_{i,1}, ..., x_{i,n}), i \in \{1, ..., m\}$，那么可以计算每个维度的均值和方差 $\mu_j, \sigma_j, j \in \{1, ..., n\}$. 具体来说，对于 $j \in \{1, ..., n\}$，可以计算

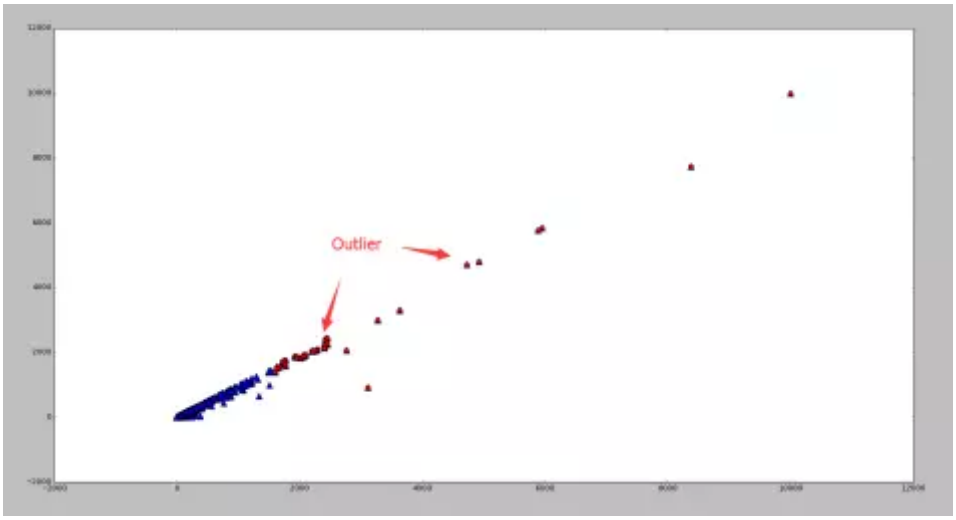
$$\mu_j = \sum_{i=1}^{m} x_{i,j}/m$$

$$\sigma_j^2 = \sum_{i=1}^{m} (x_{i,j} - \mu_j)^2/m$$

在正态分布的假设下，如果有一个新的数据 $\vec{x}$，可以计算概率 $p(\vec{x})$如下：

$$p(\vec{x}) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2})$$

根据概率值的大小就可以判断 x 是否属于异常值。运用该方法检测到的异常点如图，红色标记为异常点，蓝色表示原始的数据点。



（2）多元高斯分布的异常点检测

假设 n 维的数据集合 $\vec{x} = (x_1, ..., x_n),$，可以计算 n 维的均值向量

$$\vec{\mu} = (E(x_1), ..., E(x_n))$$

和 $n \times n$ 的协方差矩阵：

$$\Sigma = [Cov(x_i, x_j)], i, j \in \{1, ..., n\}$$

如果有一个新的数据 $\vec{x}$，可以计算

$$p(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}))$$

根据概率值的大小就可以判断 $\vec{x}$ 是否属于异常值。

（3）使用 Mahalanobis 距离检测多元离群点

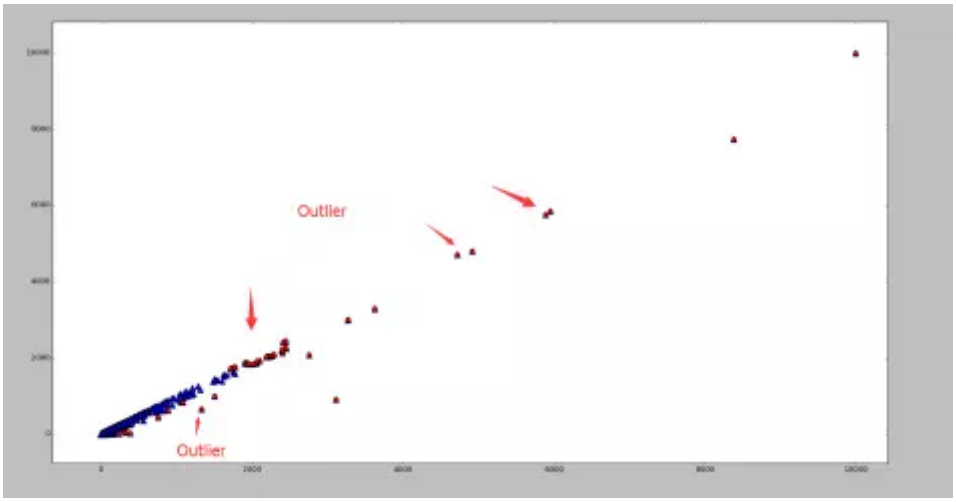
对于一个多维的数据集合 D，假设 $\overline{a}$ 是均值向量，那么对于数据集 D 中的其他对象 a，从 a 到 $\overline{a}$ 的 Mahalanobis 距离是

$$MDist(a,\overline{a}) = \sqrt{(a-\overline{a})^T S^{-1}(a-\overline{a})},$$

其中 S 是协方差矩阵。

在这里，$MDist(a,\overline{a})$ 是数值，可以对这个数值进行排序，如果数值过大，那么就可以认为点 a 是离群点。或者对一元实数集合 $\{MDist(a,\overline{a})|a \in D\}$ 进行离群点检测，如果 $MDist(a,\overline{a})$ 被检测为异常点，那么就认为 a 在多维的数据集合 D 中就是离群点。

运用 Mahalanobis 距离方法检测到的异常点如图，红色标记为异常点，蓝色表示原始的数据点。



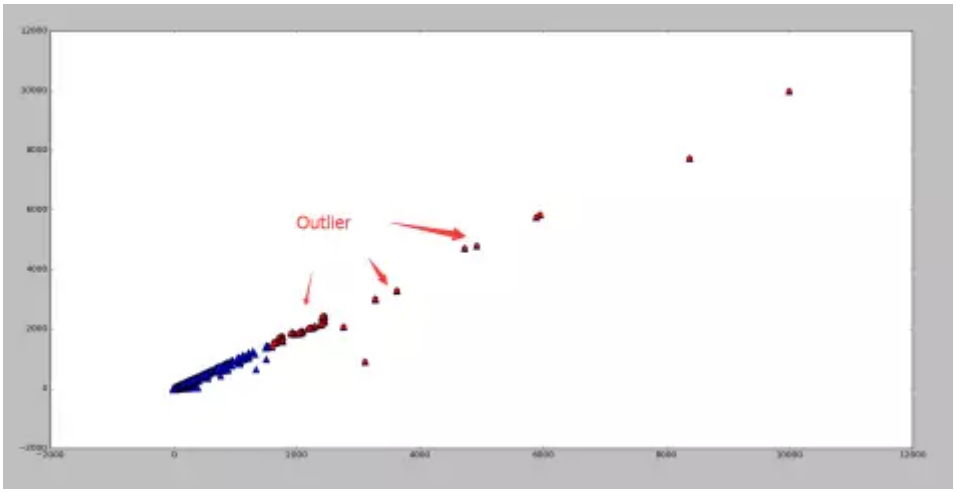
（4）使用 χ^2 统计量检测多元离群点

在正态分布的假设下，χ^2 统计量可以用来检测多元离群点。对于某个对象 $\mathbf{a}$，χ^2 统计量是

$$\chi^2 = \sum_{i=1}^{n}(a_i - E_i)^2/E_i.$$

其中，a_i 是 $\mathbf{a}$ 在第 i 维上的取值，E_i 是所有对象在第 i 维的均值，n 是维度。如果对象 $\mathbf{a}$ 的 χ^2 统计量很大，那么该对象就可以认为是离群点。

运用 χ^2 统计量检测到的异常点如图，红色标记为异常点，蓝色表示原始的数据点。



END

相关文章推荐：

1. 量子计算（一）
2. 特征工程简介
3. 转行数据挖掘和机器学习
4. 聚类算法（一）

欢迎大家关注公众账号数学人生
（长按图片，识别二维码即可添加关注）


数学人生