数据挖掘实验报告

成绩:

姓名	沈建鑫	学号	18051624
实验时间	12.25	实验地点	四教北 404
实验名称	异常检测		

一、实验过程与问题

请列举实验过程中遇到的问题与对应的解决方案。

1.在对实验要求的数据集进行可视化时，对于一维数据的可视化比较难以实现，想到了以时间为横坐标来进行数据可视化操作。

2.在以日、月、年为单位聚合降水量数据时，一开始想错了，想的是先都按原始数据集来计算变化量，然后基于这个来作日月年的聚合。但仔细一想就会发现问题：日降水变化量按月、年聚合并没有什么实际意义。所以后来修改代码，将其调整为首先对原始数据集做日、月、年的聚合，再在这个基础上分别对各自做降水量变化差值，这时得到的就是（当月总量-上月总量）、（当年总量-上年总量），这时的数据就有意义了。

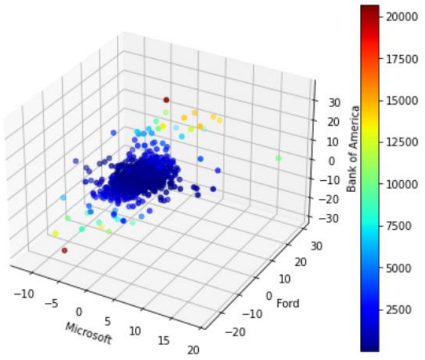
二、实验结果与分析

请回答实验代码文件（ipynb 文件）中的思考题

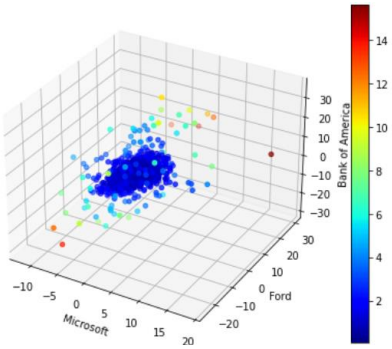
思考：尝试对比分析使用参数的异常检测模型与基于距离的异常检测方法的特点（包括但不限于方法理论、原理、实验结果等）

答：使用参数的异常检测模型需要假设大多数数据的实例符合某些概率分布，然后找出不符合分布的数据来检测异常，该方法非常注重原始数据的概率分布正确性，在原始数据概率分布判断正确的情况下可以得到比较好的结果，而在原始数据不符合该概率分布时，难以得到正确的结果。该方法需要模型来进行检测。

而基于距离的异常检测方法不同于基于参数的方法，而是通过计算不同样本间的距离来检测异常。该方法成立的假设是：若一个数据对象和大多数点距离都很远，那这个对象就是异常。该方法相比于参数方法，首先是比较简单，因为确定一个距离度量比确定一个分布要简单的多，但是计算复杂度较高，同时对 k 值和阈值十分敏感，对于密度不均的点很难得到好的结果。



基于参数的方法



基于距离的方法

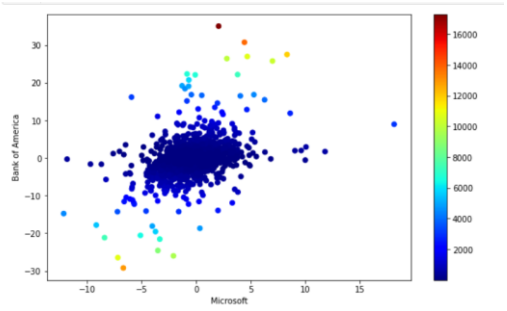
结合实验中的案例，我们可以看到两种不同的方法对于样本集中的异常检测结果确有不同，基于参数的方法在概率分布方向上的影响更大，而基于距离的方法在样本密度间的差异更大。

思考： 对以下数据集使用参数模型，并简述你的观察结果与相关分析

1.对本 tutorial 中的股票集“stocks.csv”中的一个或者两个维度进行异常分析

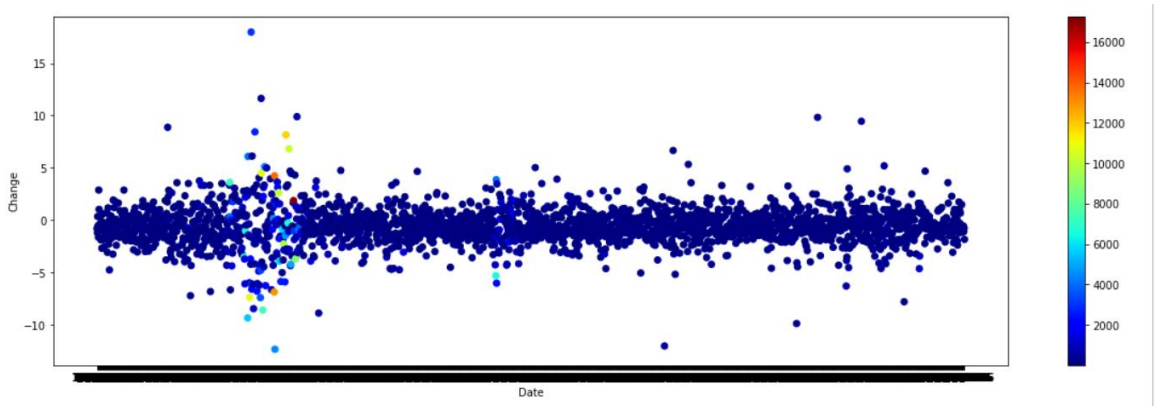
2.对“tutorial4_Data Preprocessing”中的降水量据集“DTW_prec.csv”进行不同尺度（例如年月日）的聚合预处理，然后再对降水量及其变化进行异常分析

答：1.



二维的结果

通过基于高斯分布的参数模型，我们采用马氏距离对 MSFT 和 BAC 这两个维度的信息的股价变化百分比进行了异常检测。可以看到，当数据差异在 BAC 维度上时，异常检测比较明显，而在 MSFT 维度上时，异常检测不明显，这符合高斯概率分布，由此可以看出我们的实验结果符合基于模型的异常检测预期。

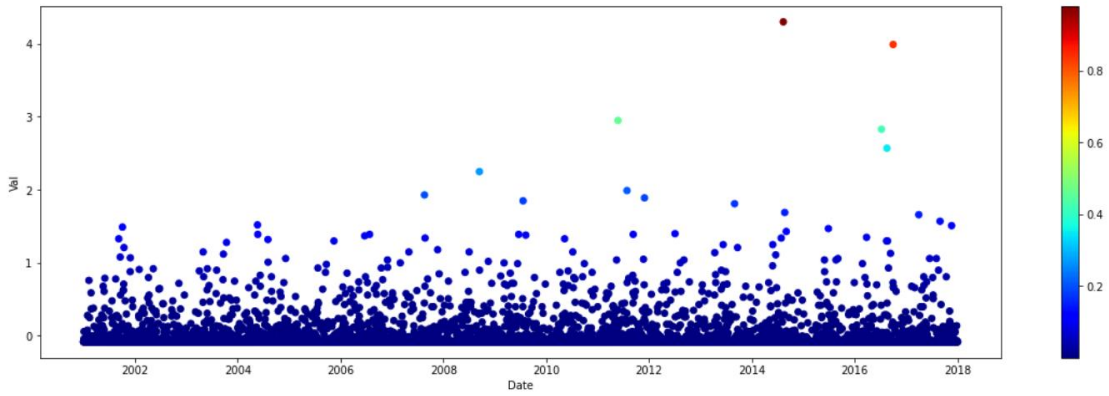


一维的结果

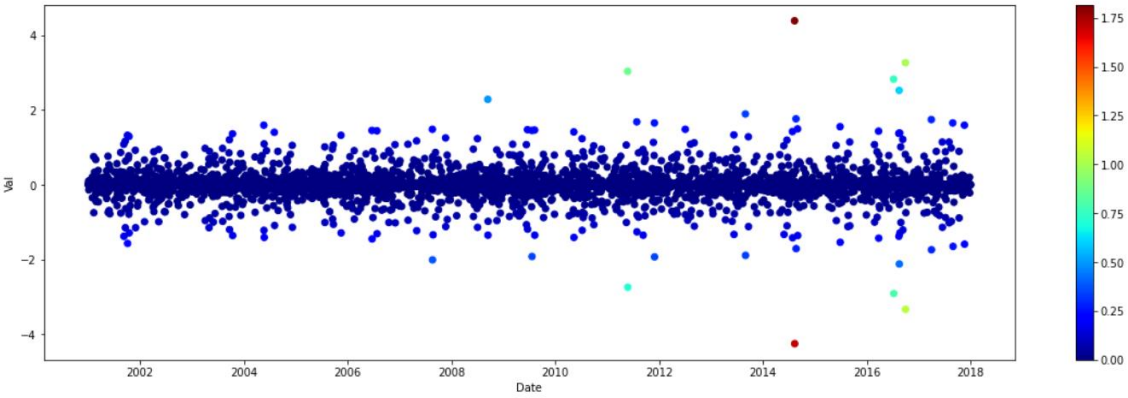
相对二维的模型，一维我们只采用了 MSFT 的数据来进行异常检测，为了便于显示我们用了时间作为横坐标来进行可视化。可以看到，在某一段时间内股价变化的异常点检测比较明显，但结果似乎有点奇怪，可见在一维的情况下数据可能不符合高斯分布或者不适用于马氏距离度量。

2.

以日聚合



降水量值异常检测



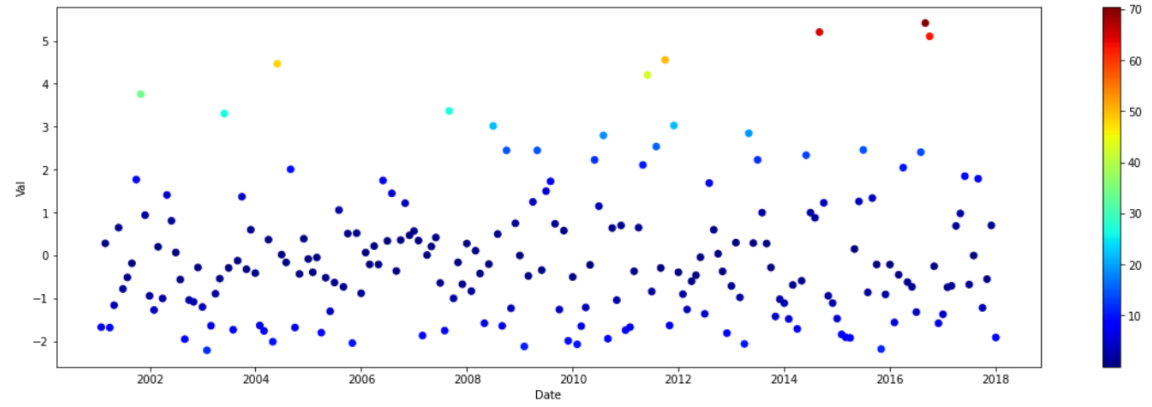
变化量异常检测

DATE	PRCP	Anomaly score
8/11/2014	4.38	1.817051
8/12/2014	-4.25	1.710790

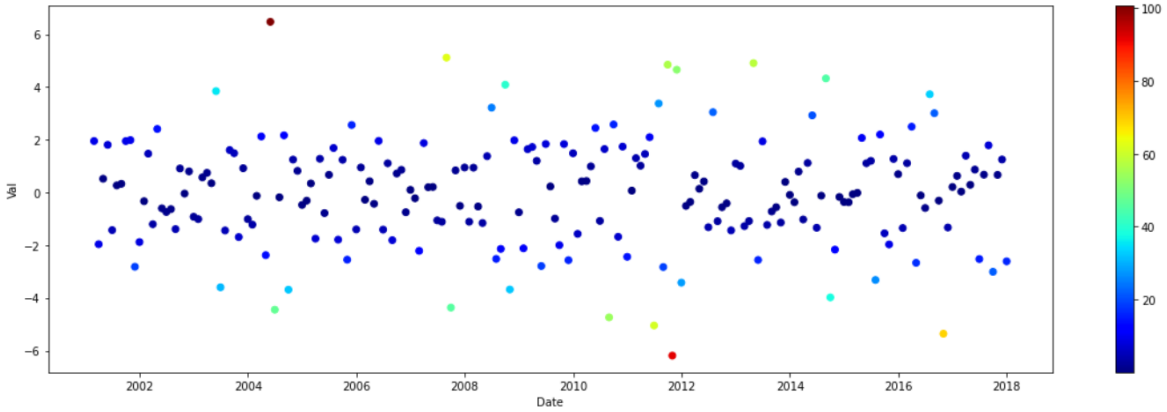
变化量异常值输出

根据上图可视化结果我们可以看到当以日作为聚合标准时，降水量变化出现了两个比较明显的异常点，对照降水量值的可视化结果，我们可以发现是恰好是一个降水量值很高的异常点位置，因此结合实际，我们可以推断那天应该是有台风登录等剧烈降水气候的出现，从而导致了数据异常，当然也不排除检测仪器错误，具体的还是要结合当时的实际情况再下结论。

以月聚合



降水量值异常检测



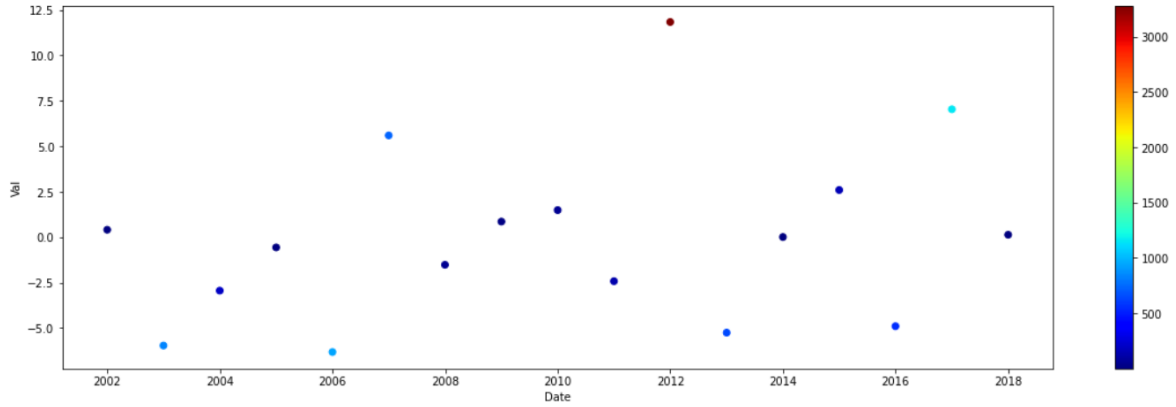
降水量变化量异常检测

DATE	PRCP	Anomaly score
2004-05-31	6.561182	178.665862
2011-10-31	-6.108818	154.878705

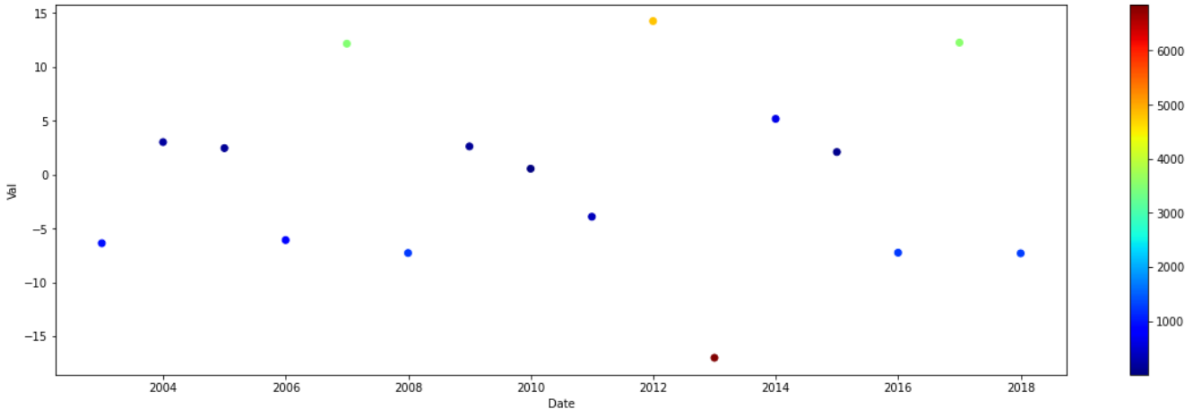
变化量异常值输出

从可视化结果中我们可以看到，在以月聚合的结果中，据集在中间部分的点按照参数模型的估计都判定为正常点，而最大的异常点出现在了 2004-05-31 和 2011-10-31。

以年聚合



降水量值异常检测



降水变化量异常检测

	PRCP	Anomaly score
DATE		
2012-12-31	-16.9725	21052.769198
2011-12-31	14.2975	14939.559939

变化量异常值输出

在以年聚合的结果中，我们可以观察到降水量变化的异常点分布在 2012 年左右。

由上述几个实验结果我们可以看出，虽然聚合度不同，但异常检测的点基本都分布在左边两边的上边缘点上，这个结果符合高斯概率分布模型。

思考：请查阅相关资料或者进行搜索，分析

1.马氏距离的优缺点

2.列举一下马氏距离适用的场景

答：1.马氏距离是数据的协方差距离，其优点是不受量纲的影响，也就是说马氏距离与原始数据的单位无关，可以排除变量之间的相关性干扰，可以应对高维线性分布的数据中各维度间非独立同分布的问题。缺点：不可避免地会夸大微小变量的作用，并且由于协方差矩阵不稳定，不一定能够顺利计算出马氏距离，要求样本数大于样本的维数。

2.马氏距离适用的场景：根据马氏距离的特点，首先就是可以用它来进行异常检测算法的实现，从影像中将异常信息从影响背景和噪声中分离出来，比较出名的有 RX 异常检测算法。RX 算法假设数据空间白化且服从高斯分布，在此基础上通过分析窗口的统计量（均值和方差），并与设定的阈值比较判断是否为异常值。可以看出，RX 算子实际上计算的是待检测点光谱与背景窗口均值向量之间的马氏距离。

其次可以用它在 SLAM 中计算实际观测特征和地图特征之间的距离，从而能够较为准确的选出最可能的关联。

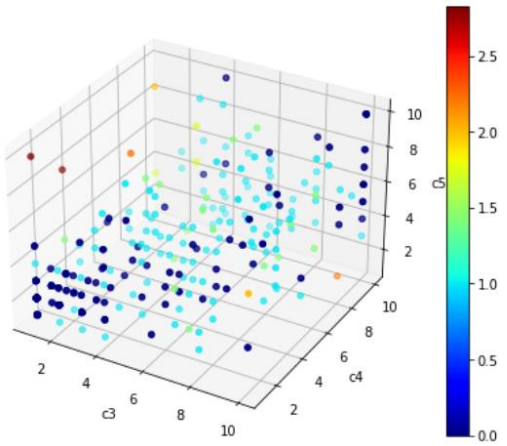
思考：“所有 3 只股票的收盘价格下跌”是什么时候出现的，请尝试通过网络搜索等方式分析为何当时出现了“所有 3 只股票的收盘价格下跌”的情况

答：根据实验报告中的可视化结果分析，我们可以得出所有 3 只股票都是在 2008 年 10 月的时候收盘价格下跌，经过网络查询后发现在当时美国第四大投资银行雷曼兄弟向美国政府申请破产保护，一场自大萧条以来最为严重的金融海啸迅速向全球蔓延。金融危机的出现使得所有股票都进行了下跌，可见数据分析结果与实际符合。

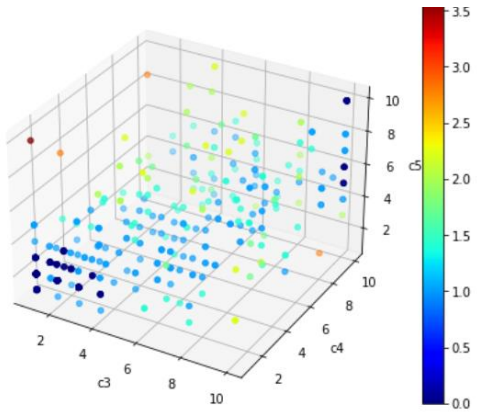
思考：对以下数据集使用基于距离的异常方法，并简述你的观察结果与相关分析

对“tutorial4_Data Preprocessing”中的癌症据集“breast-cancer-wisconsin.data”中的一个或者多个维度进行异常分析

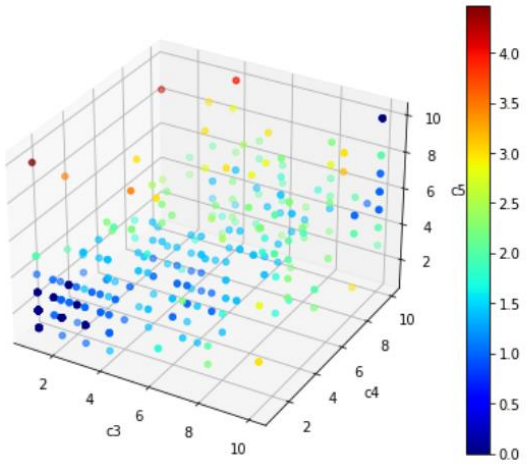
答：首先我们读取 breast-cancer-wisconsin.data 数据集中的数据，因为数据集中存在有“？”的数据，所以我们先通过数据清洗把这些错误数据清洗掉，然后对其中的第四第五第六列数据进行异常检测可视化分析，结果如下图：



k=2



k=4



k=8

根据上图中的结果我们可以看出，不同的 k 值对癌症数据集的异常点结果检测还是有比较大的不同，虽然它们对特别异常的点的检测效果都差不多，但在非异常点的判定上就有了很大的区别，当然这和基于距离的算法原理有关，具体如何划分比较合适还是要根据实际应用来定义，如果觉得稍有变化就有癌症风险，那我们理应调大 k 值，而如果觉得稍有变化可以判定为正常，那么可以相应调小 k 值来使得异常点不那么敏感。

思考：请查阅相关资料或者进行搜索，分析

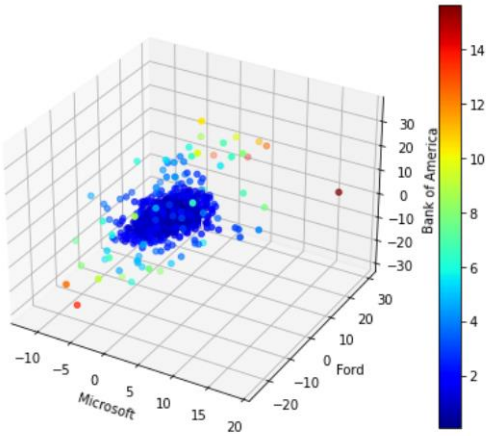
1.马氏距离和欧式距离的特点对比

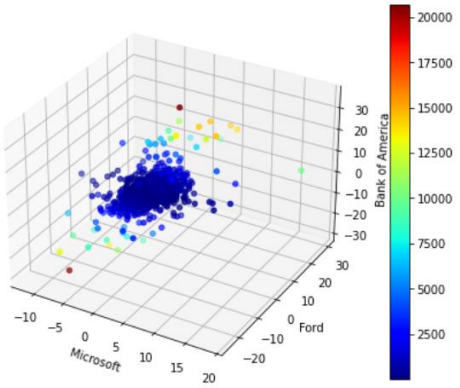
2.马氏距离和欧式距离的在本教程的异常检测的结果的对比分析

答：1.欧氏距离是通常采用的距离定义，它是在多维空间中两个点的真实距离，它将样本的不同属性之间的差别等同看待。而马氏距离计算是建立在总体样本的基础上的，对于同样的样本放在不同的样本集中得到的两个样本之间的马氏距离一般是不同的。总的来讲，欧氏距离关注的是两个样本之间的差异，而马氏距离关注的是两个样本在当前样

本集中的差异，具体哪个合适还是要根据实际情况而定。

2.





欧氏距离　　马氏距离

从上图中可以看出，在本教程的异常检测结果中，欧氏距离和马氏距离的异常检测结果对聚集位置点的效果都差不多，都属于异常值较低的点，而对于高异常点，虽然二者都评判为高异常点，但异常度的还是有不小的差距。欧式距离的结果就是完全按照绝对距离来进行计算，因此距离越远的点异常值越大。而马氏距离会结合各种特征的联系，从而使得绝对距离最远的点可能并不是异常值最大的点。同时，我们观察左图可以看到，在欧氏距离作为评判标准的前提下，有些点虽然离聚集区比较近，但仍然被评判为了异常点，这时候理论上应该采用马氏距离来排除其他变量的干扰。

三、意见和建议

（如有，请写出对本次实验的具体意见和建议，包括但不限于教学内容、实验内容、教学 PPT 等）