三、线性模型

1、引言

真实数据集中不同维度的数据通常具有高度的相关性，这是因为不同的属性往往是由相同的基础过

程以密切相关的方式产生的。在古典统计学中，这被称为——**回归建模**，一种参数化的相关性分析。

一类相关性分析试图通过其他变量预测单独的属性值，另一类方法用一些潜在变量来代表整个数

据。前者的代表是 **线性回归**，后者一个典型的例子是 **主成分分析**。本文将会用这两种典型的线性相关分

析方法进行异常检测。

需要明确的是，这里有两个重要的假设：

假设一：近似线性相关假设。线性相关假设是使用两种模型进行异常检测的重要理论基础。

假设二：子空间假设。子空间假设认为数据是镶嵌在低维子空间中的，线性方法的目的是找到合适

的低维子空间使得异常点(o)在其中区别于正常点(n)。

基于这两点假设，在异常检测的第一阶段，为了确定特定的模型是否适合特定的数据集，对数据进

行探索性和可视化分析是非常关键的。

3、线性回归

在线性回归中，我们假设不同维度的变量具有一定的相关性，并可以通过一个相关系数矩阵进行衡

量。因此对于特定的观测值，可以通过线性方程组来建模。在实际应用中，观测值的数量往往远大于数

据的维度，导致线性方程组是一个超定方程，不能直接求解。因此需要通过优化的方法，最小化模型预

测值与真实数据点的误差。

线性回归是统计学中一个重要的应用，这个重要的应用往往是指通过一系列自变量去预测一个特殊

因变量的值。在这种情况下，异常值是根据其他自变量对因变量的影响来定义的，而自变量之间相互关

系中的异常则不那么重要。这里的异常点检测主要用于数据降噪，避免异常点的出现对模型性能的影

响，因而这里关注的兴趣点主要是正常值(n)。

而我们通常所说的异常检测中并不会对任何变量给与特殊对待，异常值的定义是基于基础数据点的

整体分布，这里我们关注的兴趣点主要是异常值(o)。

广义的回归建模只是一种工具，这种工具既可以用来进行数据降噪也可以进行异常点检测。

3.1 基于自变量与因变量的线性回归

3.1.1 最小二乘法

为了简单起见，这里我们一元线性回归为例:

变量Y为因变量，也就是我们要预测的值； 为一系列因变量，也就是输入值。系数

为要学习的参数。假设数据共包含个样本，第个样本包含的数据为和，带入式

(1) 如下式所示:

3.1.2 梯度下降法

数据集

监督学习一般靠数据驱动。我们通常收集一系列的真实数据，例如多栋房屋的真实售出价格和它们

对应的面积和房龄。我们希望在这个数据上面寻找模型参数来使模型的预测价格与真实价格的误差最

小。在机器学习术语里，该数据集被称为训练数据集（training data set）或训练集（training set），

通常还应该有一个用于防止过拟合的交叉验证集和一个用于评估模型性能的测试集(test set)。一栋房屋

被称为一个样本（sample），其真实售出价格叫作标签（label），用来预测标签的两个因素叫作特征