哈爾濱工業大學

计算建模实验报告

题　　目	基于RANSAC和最小二乘的直线和曲线拟合
学　　院	计算机科学与技术
专　　业	计算机科学
学　　号	1190200122
学　　生	袁　野
任课教师	刘绍辉

哈尔滨工业大学计算机科学与技术学院

2021.9

实验五:基于 RANSAC 和最小二乘的直线和曲线拟合

注意：请按照大家阅读文献的格式进行撰写，确保文档格式的规范性！

一、 实验内容或者文献情况介绍

1、 直线拟合

根据直线方程$ax+by+c=0$，产生随机点(x_i,y_i)，然后增加随机噪声成为(x_i+n_i,y_i+m_i)，根据这些点，拟合直线$ax+by+c=0$中的参数。

如果有一系列平行直线$ax+by+c_1=0$，$ax+by+c_2=0$，$ax+by+c_3=0$，然后对直线上的点添加类似的噪声，拟合这些平行直线。

2、 曲线拟合

自己设计曲线方程，例如圆方程，椭圆方程，然后添加适当的噪声(例如高斯噪声)，然后分别采用 ransac 和最小二乘方法进行拟合。

如果添加一些外点，拟合效果如何？是否有方法改进！

二、 算法简介及其实现细节

1、 拟合方式

1.1 RANSAC 拟合

a、直线拟合

随机选择两点(确定一条直线所需要的最小点集)；由这两个点确定一条线 l；

根据阈值 t，确定与直线 l 的几何距离小于 t 的数据点集 S(l)，并称它为直线 l 的一致集；

重复若干次随机选择，得到直线 l1,l2,…, ln 和相应的一致集 S(l1),S(l2),…,S(ln)；

使用几何距离，求最大一致集的最佳拟合直线，作为数据点的最佳匹配直线。

b、 一般模型

确定求解模型 M，即确定模型参数 p，所需要的最小数据点的个数 n。

由 n 个数据点组成的子集称为模型 M 的一个样本；从数据点集 D 中随机地抽取一个样本 J，由该样本计算模型的一个实例 Mp(J)，确定与 Mp(J)之间几何距离<阈值 t 的数据点所构成的集合，并记为 S(Mp(J))，称为实例 Mp(J)的一致集；

如果在一致集 S(Mp(J))中数据点的个数# S(Mp(J))>阈值 T, 则用 S(Mp(J))重新估计模型 M，并输出结果；如果# S(Mp(J)) < 阈值 T，返回到步骤 2；

经过 K 次随机抽样，选择最大的一致集 S(Mp(J))，用 S(Mp(J))重新估计模型 M，并输出结果。

1.2 最小二乘法拟合

最小二乘法是由勒让德在 19 世纪发现的，形式如下式：

$$标函数 = \sum （观测值 - 理论值）^2$$

观测值就是我们的多组样本，理论值就是我们的假设拟合函数。目标函数也就是在机器学习中常说的损失函数，我们的目标是得到使目标函数最小化时候的拟合函数的模型。举一个最简单的线性回归的简单例子，比如我们有m个只有一个特征的样本：$(x_i, y_i)(i = 1,2\ldots,m)$

样本采用一般的$h_\theta(x)$为n次的多项式拟合，$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_n x^n, \theta(\theta_0, \theta_1, \theta_2, \ldots, \theta_n)$为参数

最小二乘法就是要找到一组$\theta(\theta_0, \theta_1, \theta_2, \ldots, \theta_n)$使得$\sum_{i=1}^{n}(h_\theta(x_i) - y_i)^2$(残差平方和)最小，即，求

$$min \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2$$

最小二乘法的代数法解法就是对θ_i求偏导数，令偏导数为0，再解方程组，得到θ_i。矩阵法比代数法要简洁，下面主要讲解下矩阵法解法，这里用多元线性回归例子来描述：

假设函数$h_\theta(x_1, x_2, \ldots, x_n) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_n x^n, \theta(\theta_0, \theta_1, \theta_2, \ldots, \theta_n)$的矩阵表达方式为：

$$h_\theta(\boldsymbol{X}) = \boldsymbol{X}\theta$$

其中，假设函数 $h_\theta(\boldsymbol{X}) = \boldsymbol{X}\theta$ 为 $m * 1$ 的向量，θ 为 $n * 1$ 的向量，里面有 n 个代数法的模型参数。$\boldsymbol{X}$ 为 $m * n$ 维的矩阵。n 代表样本的个数，m 代表样本的特征数。

损失函数定义为 $J(\theta) = \frac{1}{2}(\boldsymbol{X}\theta - \boldsymbol{Y})^T(\boldsymbol{X}\theta - \boldsymbol{Y})$，其中 $\boldsymbol{Y}$ 是样本的输出向量，维度为 $m * 1$。$\frac{1}{2}$ 在这主要是为了求导后系数为1，方便计算。

根据最小二乘法的原理，我们要对这个损失函数对 θ 向量求导取0。结果如下式：

$$\frac{\partial}{\partial\theta} J(\theta) = \boldsymbol{X}^T(\boldsymbol{X}\theta - \boldsymbol{Y}) = 0$$

对上述求导等式整理后可得：

$$\theta = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

2、 具体实现

2.1 拟合单条直线

首先我们确定一条直线为 $y = 5x + 13$，那么我们均匀随机生成 100 个范围在$(0,20)$的浮点数作为数据集的横坐标，然后我们将其带入直线公式得到对应的 100 个纵坐标，这样我们就的到了 100 个直线上的点坐标。然后我们将这些点的横坐标和纵坐标的数字均添加一个方差为 2 的高斯噪声，得到最终的数据集。

然后分别通过最小二乘方法和 ransac 对数据进行拟合，在最小二乘

方法中我们需要将样本特征数设为 2。然后我们将数据代入后即可。在 ransac 中，由于是拟合直线，因此每条直线由两个数据点模拟，返回结果即可。

以上代码为 5.1.py。

2.2 拟合平行直线

由于三条直线平行，因此我们直接对数据集进行拟合即可，拟合出来的结果的斜率即为三条直线的斜率。斜率确定后，我们只需要根据当前斜率下各点的截距进行排序分类即可。拟合部分与单条直线无异。

以上代码为 5.2.py

2.3 用最小二乘方法拟合曲线

这里拟合的是 $y = sin(x)$。生成数据集的方式与直线拟合类似，首先随机生成 100 个 $y = sin(x)$ 上的点，然后对这些点的横纵坐标分别加上方差为 0.1 的高斯噪声。

在拟合过程中，样本特征数的不同会导致不同的拟合结果。样本特征数对拟合结果的影响我们将在第三部分讨论。

以上代码为 5.3.py

2.4 用 ransac 方法拟合曲线

此处我们拟合的是$(x-3)^2+(y-3)^2=1$。生成数据集时，首先生成 100 个圆上的点，随后对这些点的横纵坐标分别加上方差为 0.1 的高斯噪声。

由于三点确定一个圆，因此在 ransac 的过程中，我们需要随机三个数据点来产生圆，并统计所有数据点分别到圆心的距离与原半径的差值在阈值 th 内的点数，最终确定模型。

有关终止阈值 K，采用自适应算法：

对内点比例作最保守估计$w = w_0$（如$w_0 = 0.1$，这意味着在数据点集中可能有 90%的外点），应用公式$K = \frac{logz}{\log(1-w^n)}$得到抽样次数 K 的初始值$K_0$，其中 z 取值为 0.02，t 取值为 0.1。

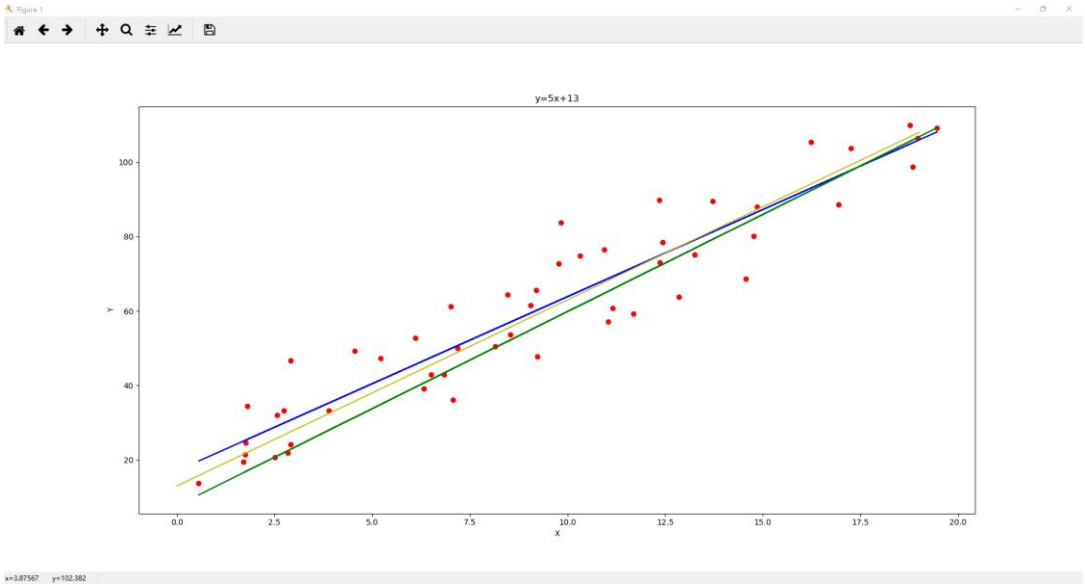
抽样并更新w_0, K_0；令当前抽样的一致集所含数据点占整个数据点的比例为，若$w > w_0$，则更新$w = w_0$，并且应用公式$K = \frac{logz}{\log(1-w^n)}$更新抽

样数K_0；否则，保持原来的w_0, K_0；

如果抽样次数已达到或超过 K_0，则终止抽样；否则，返回上一步。

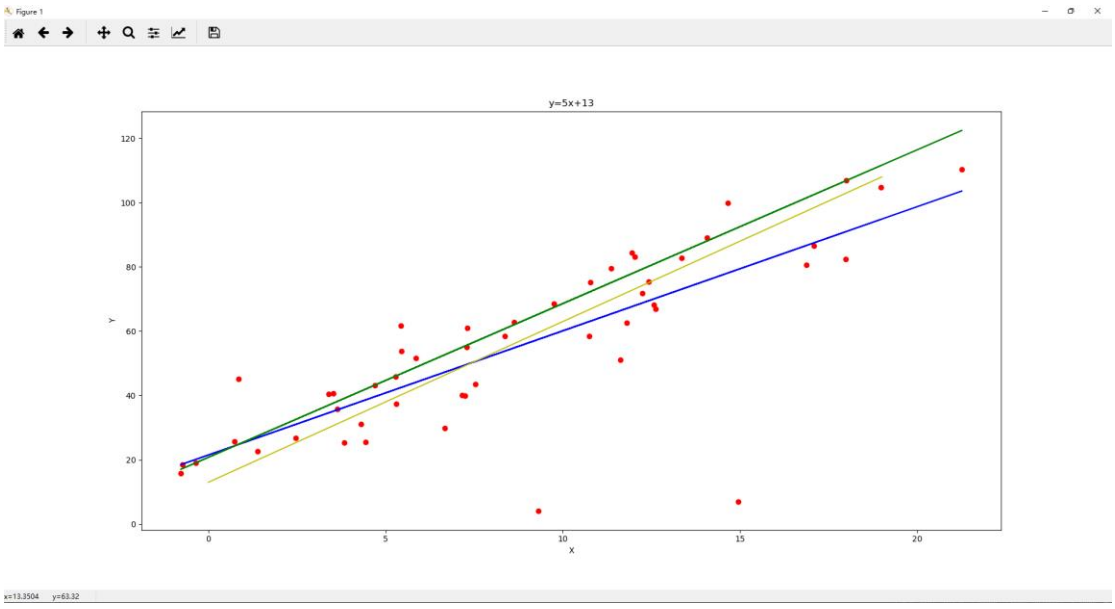
三、 实验设置及结果分析（包括实验数据集）

1、拟合单条直线



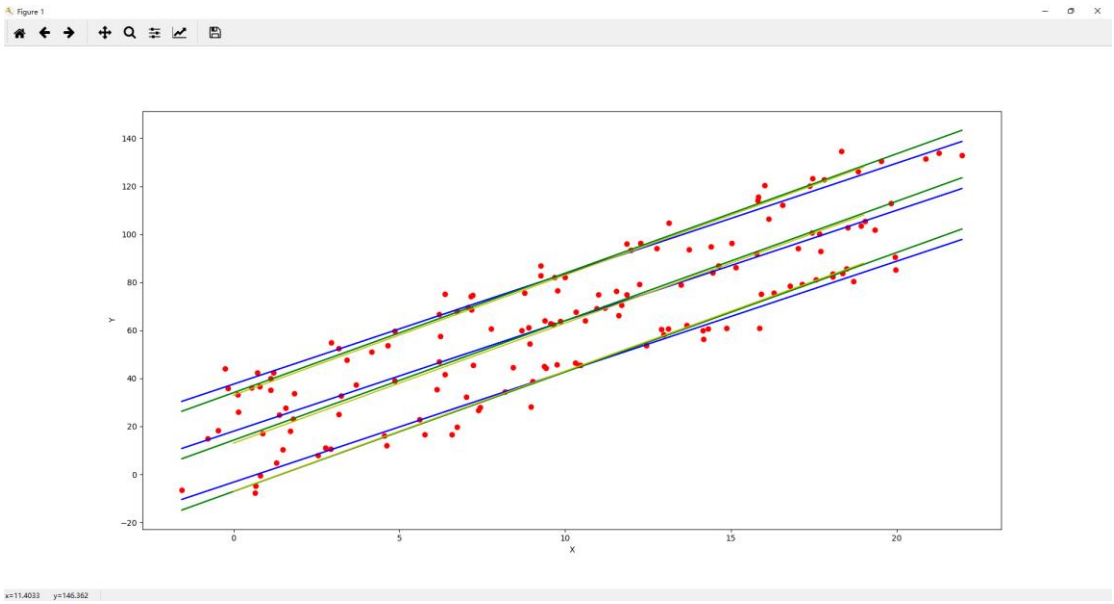
红点为随机产生的数据集，黄色线条为$y = 5x + 13$, 蓝色线条为最小二乘方法拟合的直线，绿色线条为 ransac 拟合的线条。

加入三个外点后的效果如下：



显然 ransac 拟合出来的直线受外点的影响要小于最小二乘方法拟合出来的直线，因此在由外点存在的时候，ransac 的效果是更好的。

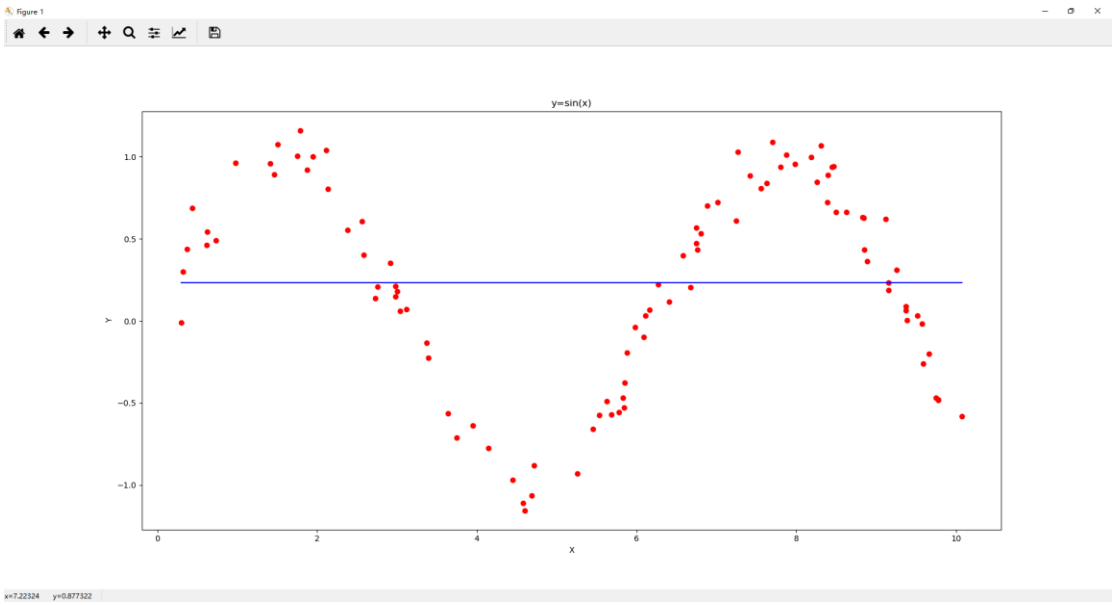
2、拟合平行线



拟合的直线为$y = 5x - 7$、$y = 5x + 13$、$y = 5x + 33$。两种拟合方法拟合出的三条直线如图所示，黄色为标准直线，蓝色为最小二乘方法拟合的直线，绿色为 ransac 拟合出的直线。

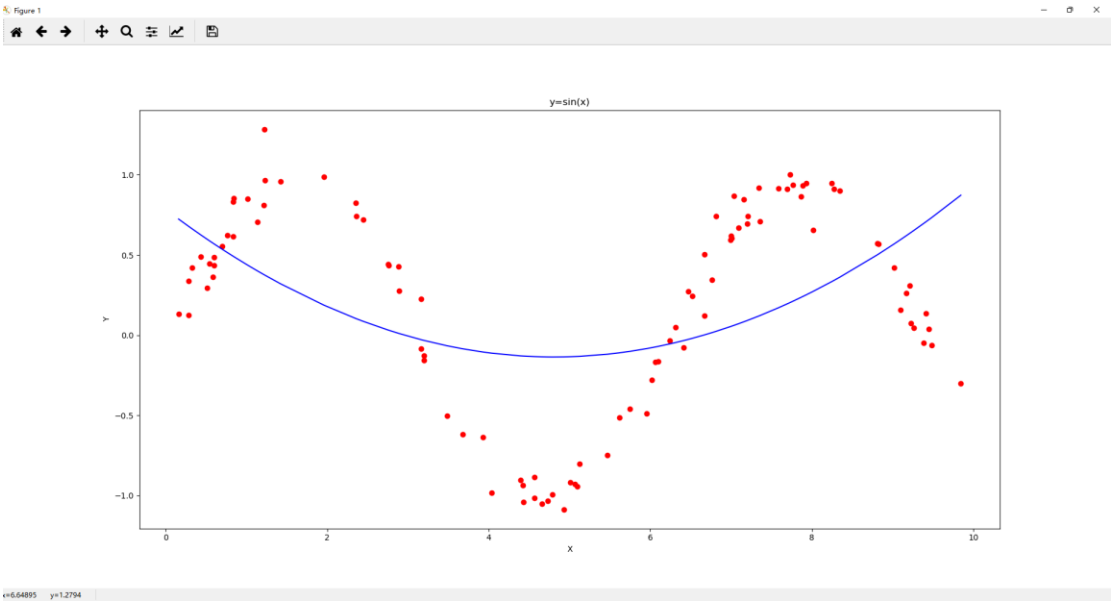
3、最小二乘方法拟合曲线

特征数为 1：



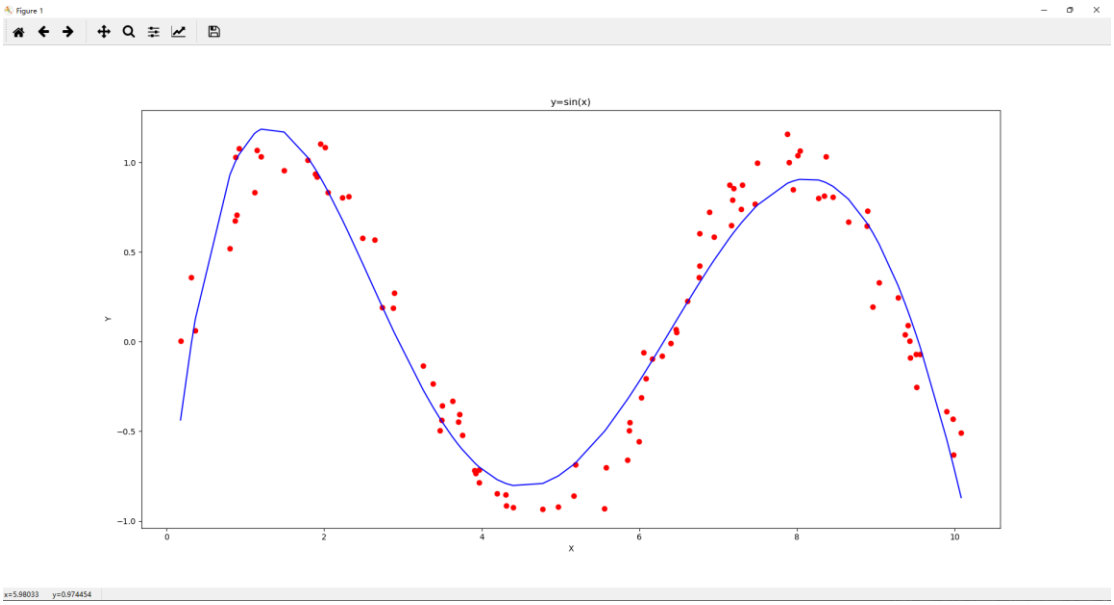
显然为欠拟合。

当特征数为 3 时：



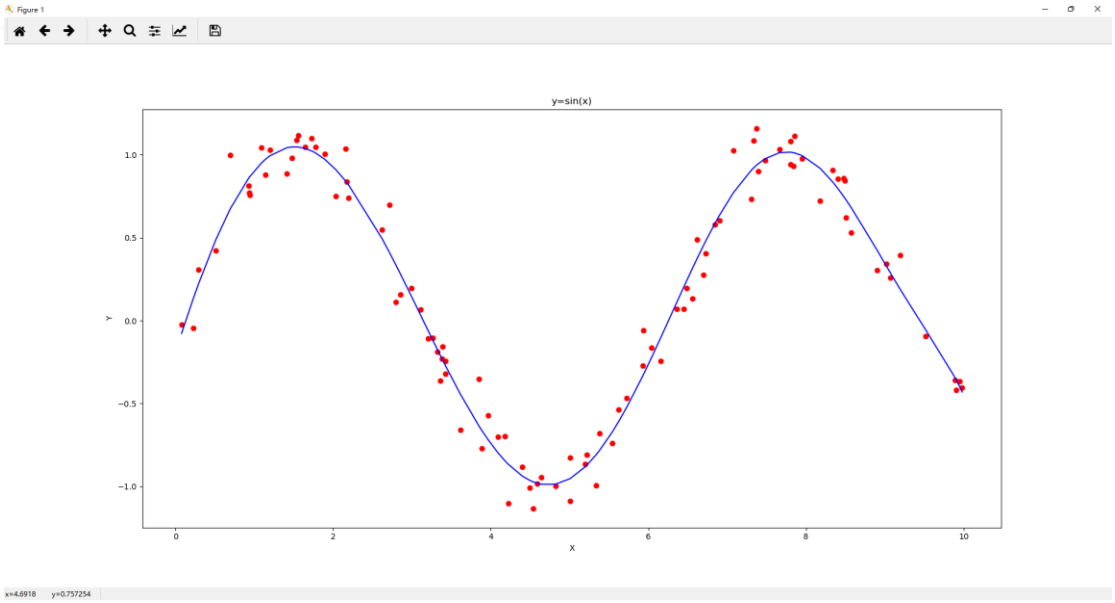
仍为欠拟合。

当特征数为 6：



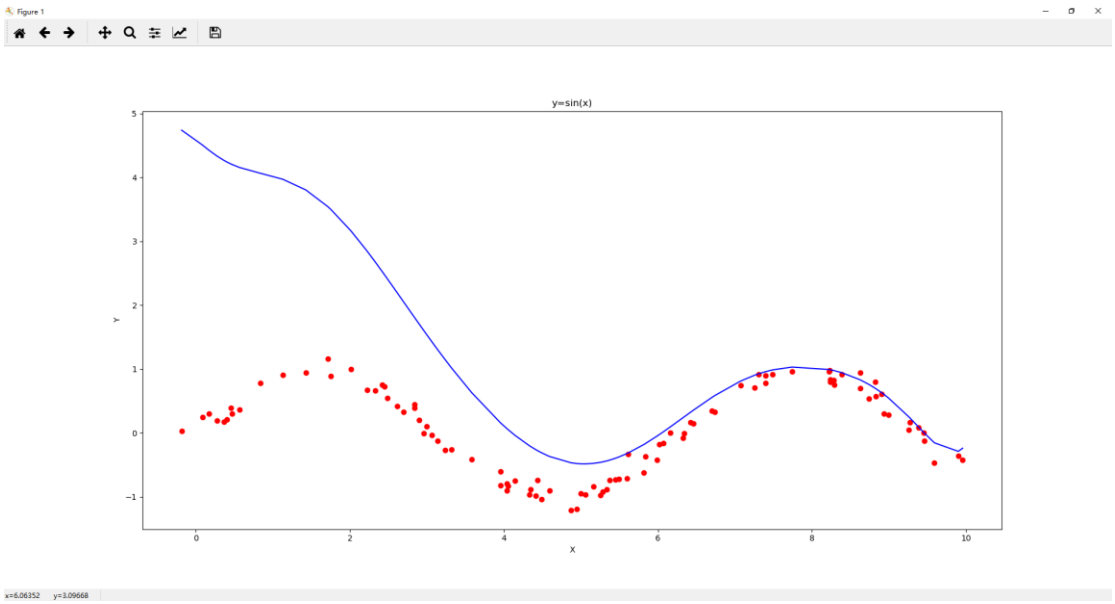
此时的拟合结果较前两次已有了较好的提升，但仍然为欠拟合。

当特征数为 10 时：



此时的拟合结果最好。

特征数为 14:

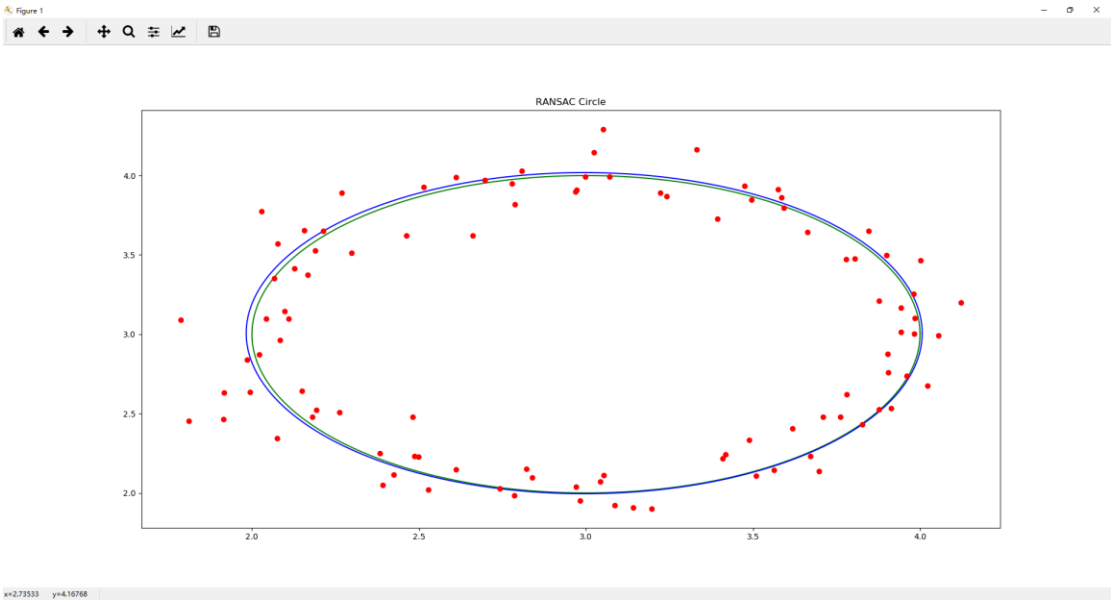


此时的拟合结果明显偏离预期，此时由于特征数过多，已经过拟合。

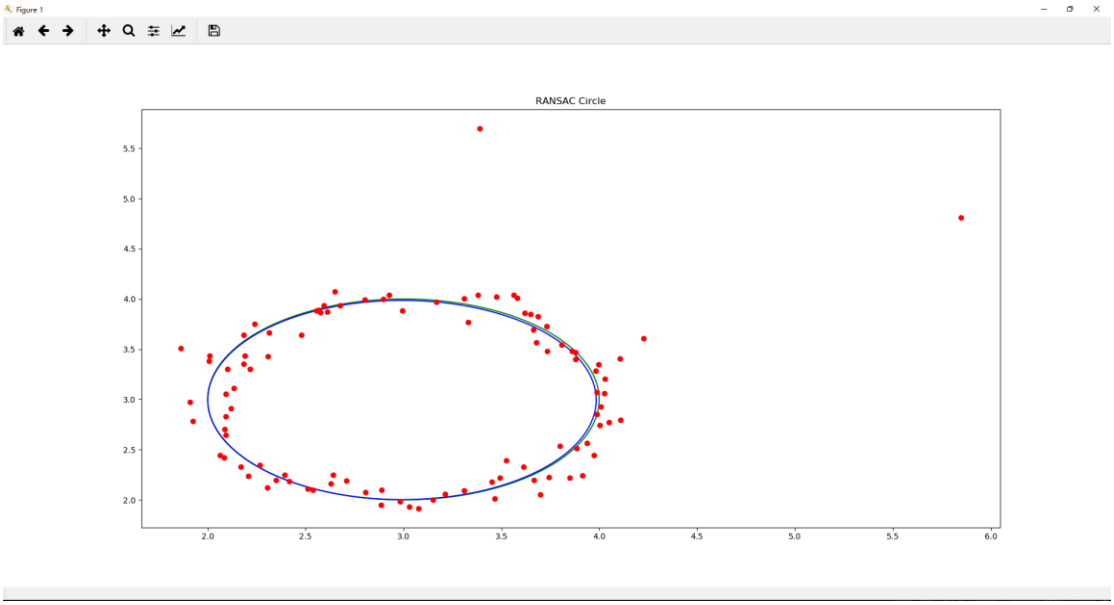
4、ransac 拟合曲线

拟合结果如下：

我们可以看出在拟合过程中加入阈值的自适应函数后会得到比较好的拟合效果。



当我们加入外点之后拟合结果如下：



可以发现加入外点之后，拟合效果基本不会受到影响，说明有外点情况下 ransac 的拟合效果受影响较小。

四、 结论

对于直线或者曲线产生的随机点，加入高斯噪声后，用最小二乘法和 RANSAC 算法都可以进行比较好的拟合。

对于外点比较多的拟合，RANSAC 算法效果更好。

对于最小二乘法，多项式阶数过低会导致欠拟合，过高会导致过拟合，需要找到合适的正则项阶数。正则项和增加样本点的数目可以有效防止过拟合。

五、 参考文献

《机器学习》周志华.北京.清华大学出版社