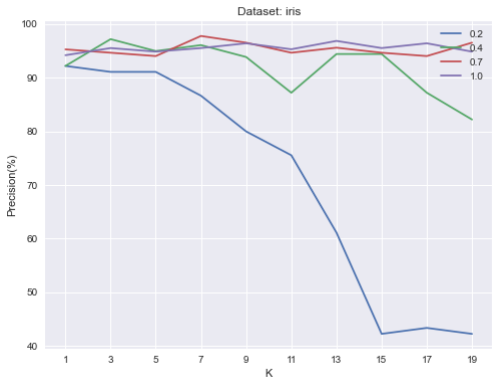
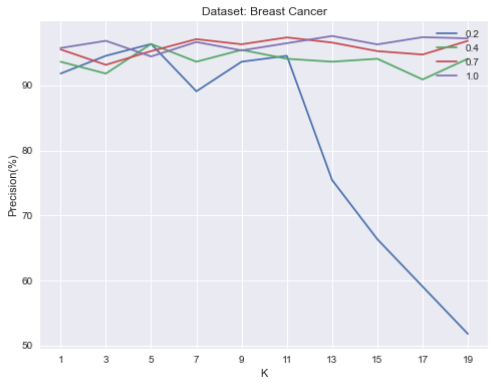
实验1：实验环境与基本概念熟悉

（1）K-近邻算法（KNN）实验

(b) 调整参数

对于iris，breast cancer数据集，我分别实验了取样率sample rate为0.2， 0.4， 0.7， 1.0，K值取2~20的组合情况的分类精度。有如下结果：





实验数据分析：可以看到，对于两组数据集，在取样率比较低的情况下(sample rate=0.2)，K值偏大的时候(例如K=13)，蓝色线条预测的准确率出现了明显下跌；在取样率正常的情况下(sample rate>=0.7)的时候，K值在2~20之间都能达到一个比较高的准确率，但绿色线条，紫色线条在K比较小的区间 (如1~5)之间的时候，测试集上的准确率有上升的趋势。

原因分析：在数据量比较少，而K又比较大的情况下，我们的模型每次都相当于只比较所有点中各个类别点数量的相对多少，模型很简单，显然是出现了“欠拟合”的情况。而当数据量比较大，K比较小的时候，我们的模型相当于每次只选择离判别点最近的点的类别，模型很“复杂”（每次都要计算最近的点），出现了“过拟合”的情况，在测试集上的泛化能力不好，因此出现了绿色线条上K比较小的时候准确率比较低的情况。

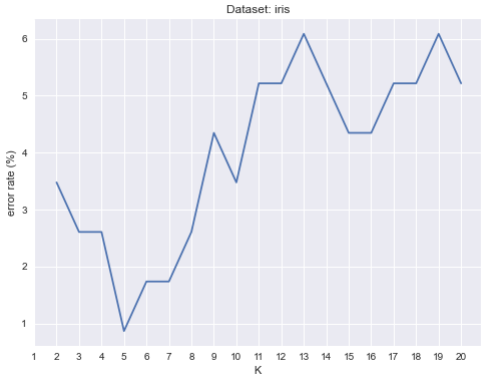
参数选取建议：在保证数据量的情况下，选择一个既不太大，又不太小的K，一般为5-15。避免出现过拟合和欠拟合。

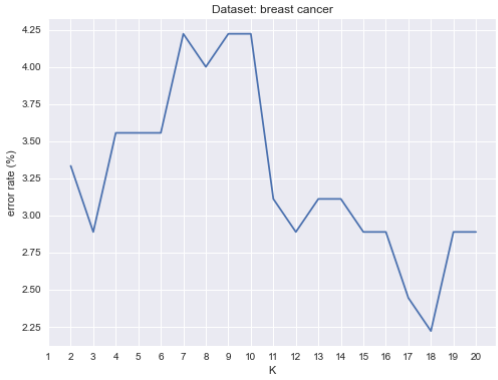
(c) 改进

改进1：提供自适应的参数选择策略

我们可以使用k-fold 交叉验证的办法，将训练集五等分，然后每次在四份数据上训练，在最后的那份数据集上做交叉验证，取多次交叉验证结果的平均值，作为某个K的平均误差率，选择最小的平均误差率的K。

分别在iris数据集和breast cancer数据集上实验，有如下结果：





可以发现，在iris数据集上，选择K=5的时候，有比较好的误分类率，在breast cancer数据集上，选择K=18的时候，有比较好的误分类率。

因此，接下来的算法改进部分我们分别使用K=5和K=18测试两个数据集。

改进2：使用kd-tree算法改进算法效率

我们可以利用点与点之间的局部距离信息（把小区间内的点组织成树），使用kd-tree算法，来改进knn的算法效率。我们使用iris数据集来验证我们的改进：

brute force算法：

i. brute force

```
brute_X_train=train_data[:, :-1]
brute_y_train=train_data[:, -1]
brute_X_test=test_data[:, :-1]
brute_y_test=test_data[:, -1]
```

```
knn = KNN(brute_X_train, brute_y_train, K)
knn.score(brute_X_test, brute_y_test)
```

```
It takes 0.0685s for function[score] to calculate precision over test set
0.9866666666666667
```

kd-tree算法：

ii. kd-tree

```
kd_train_data=list(map(lambda x:[tuple(x[:-1]), x[-1]], train_data))
kd_test_data =list(map(lambda x:[tuple(x[:-1]), x[-1]], test_data))
```

```
kd_tree=kdtree(kd_train_data)
kd_tree.score(kd_test_data, K)
```

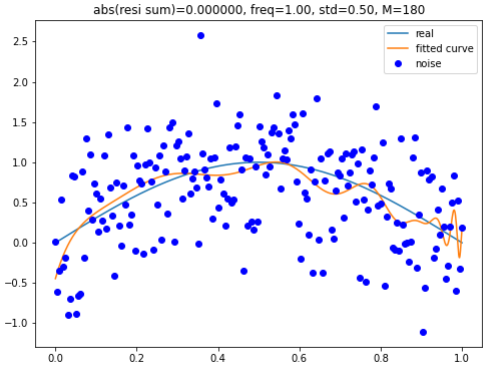
```
It takes 0.0299s for function[score] to calculate precision over test set
0.9866666666666667
```

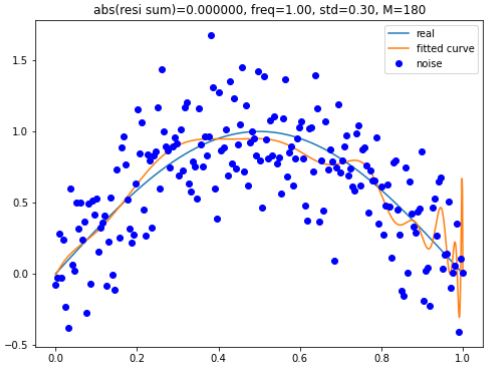
可以看到，在保证准确率的情况下，我们算法使用的时间缩短为了原来的1/3。

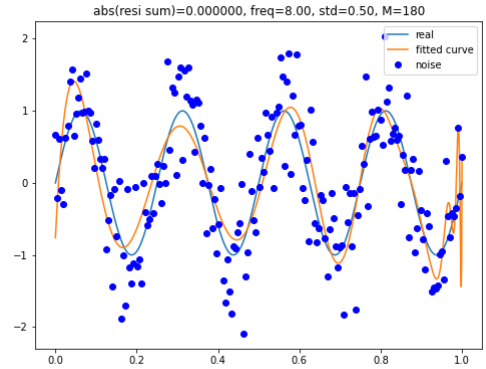
(2) 基于高阶多项式回归的欠拟合和过拟合分析实验

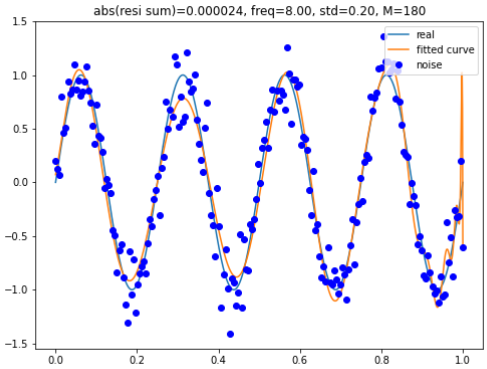
在这个实验中，通过循环组合不同的条件，可以得到以下一些“过拟合”，“欠拟合”很明显的图像（具体参数标注在了图像标题上）：

过拟合：

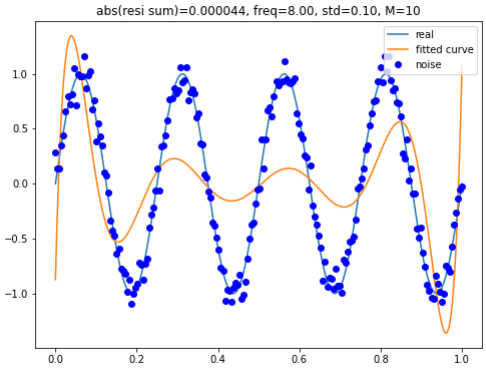


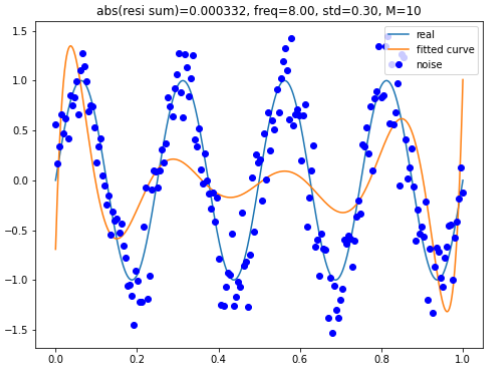


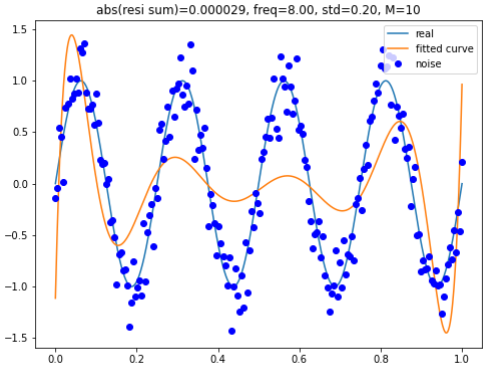


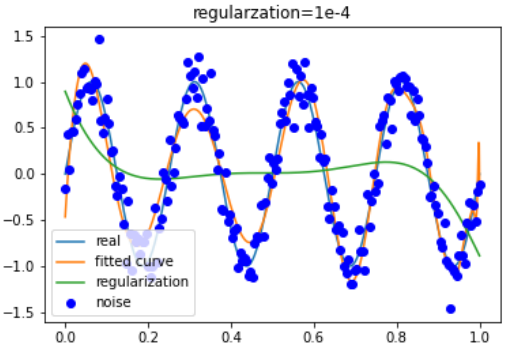


欠拟合：









拟合效果与条件的关系： 从上面八幅图中可以看出，当样本方差很大（样本噪音比较大），频率比较低，模型比较复杂（拟合的多项式阶数高）的时候，比较容易出现过拟合，尤其是当模型很复杂的时候。M=180几乎必然出现过拟合。

而当样本方差比较小（训练样本接近正弦曲线），频率比较高（样本密集），模型比较简单的时候，比较容易出现欠拟合。

直观感受： 从最后一幅图中我们可以看出，通过加上正规化项之后，模型直接从过拟合变成了欠拟合。因此这幅图提示我们可以通过调整正规化项因子λ的大小来调整模型的拟合情况。

避免过拟合：

1. 适当增大正规化因子λ
2. 对数据进行预处理，使样本各个特征方差尽可能一致
3. 减少额外的特征项，使模型更简单
4. 根据KNN实验知道，尽可能地获得更多的数据

避免欠拟合：

1. 适当减小正规化因子λ
2. 适当增加模型的复杂度，比如加入更多的特征等

本次实验所有代码和图片资源见： **MLPR实验01**