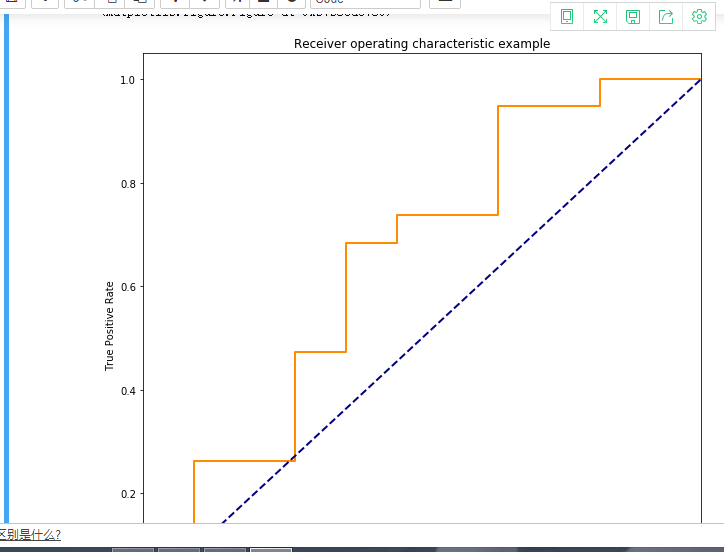
**IRIS数据集的训练模型**

**1.ROC曲线（ROC curve）**

ROC曲线是一种模型评价指标，根据学习器的预测结果对样例进行排序，按此顺序逐个把样本作为正例进行预测，每次计算出两个重要量的值（TPR、FPR），分别以它们为横、纵坐标作图。与PR曲线使用查准率、查全率为纵、横不同，ROC 曲线的纵轴是“真正例率”（真阳性率，TTR），横轴是“假正例率”（假阳性率，FPR）。

通过鸢尾花数据得到的ROC曲线，可以看出其假正例和假反例总数较少。



1. **learning 曲线**

学习曲线是通过画出不同训练集大小时训练集和交叉验证的准确率，诊断模型的偏差和方差，以及增大训练集是否可以减小过拟合。

模型用GSVC做对比，模型选择方法中需要用到 learning\_curve 和交叉验证方法 ShuffleSplit。

画学习曲线就是调用 sklearn.model\_selection 的 learning\_curve，

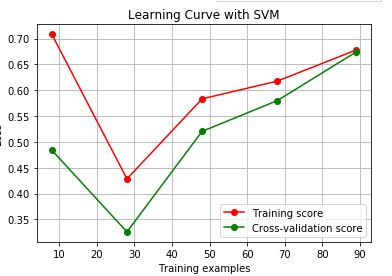
学习曲线返回的是 train\_sizes, train\_scores, test\_scores，

画训练集的曲线时，横轴为 train\_sizes, 纵轴为 train\_scores\_mean，

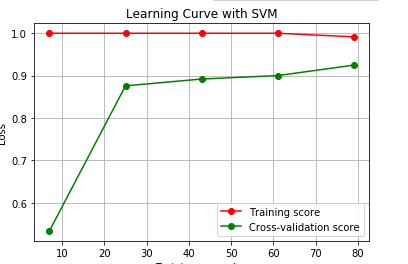
画测试集的曲线时，横轴为 train\_sizes, 纵轴为 test\_scores\_mean：

通过改变超参数和gamma的值，改变模型状态。

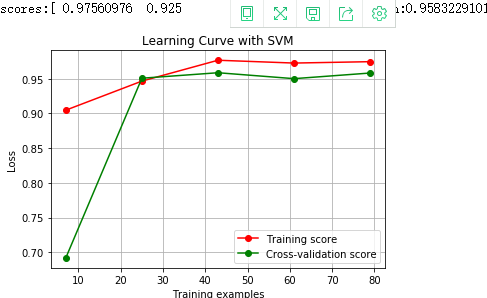






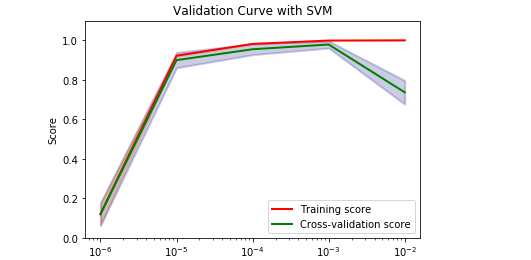






1. **验证曲线**

通过验证曲线**判定过拟合与欠拟合**。验证曲线是一种**通过定位过拟合与欠拟合等诸多问题的方法**，帮助**提高模型性能的有效工具**，绘制的是**准确率与模型参数之间的关系**。validation\_curve函数帮助实现。如果训练分数和验证分数都很低，估计量是欠拟合的。如果训练分数高而验证分数低，估计量是过度拟合，否则估计量是好的。



通过代码获得鸢尾花数据集的验证曲线，从图看出训练得分总比验证得分高，验证得分增长到最高点后由于过拟合而开始骤降。

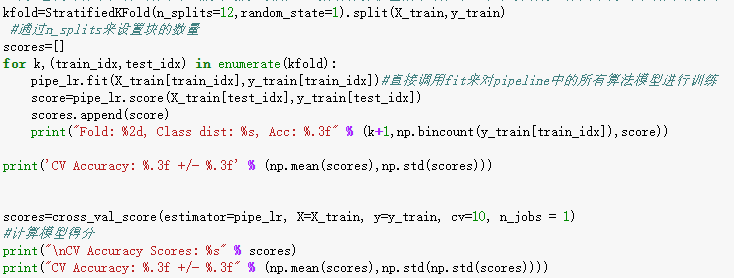
1. **k-fold cross validation**

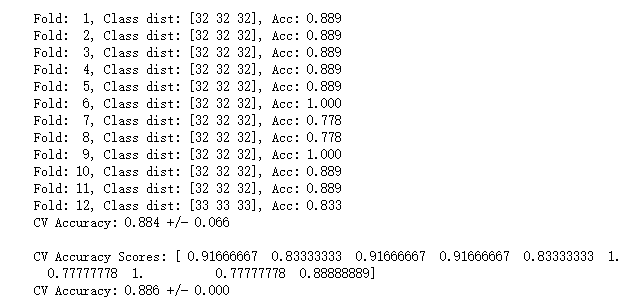
一般情况K折交叉验证用于模型调优，找到使得模型泛化性能最优的超参值。找到后，在全部训练集上重新训练模型，并使用独立测试集对模型性能做出最终评价。K折交叉验证使用了无重复抽样技术的好处：每次迭代过程中每个样本点只有一次被划入训练集或测试集的机会。

实现K折交叉验证：

使用数据集中的类标y\_train初始化sklearn.cross\_validation模块下的StratifiedKFold迭代器，通过n\_folds参数设置块的数量。

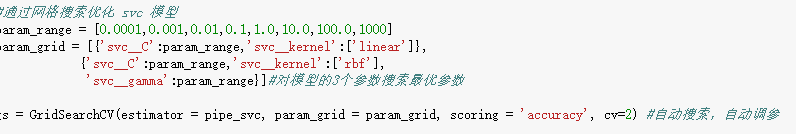
使用kfold在k个块中迭代时，使用train中返回的索引去拟合流水线，通过pipe\_lr流水线保证样本都得到适当的缩放。使用test索引计算模型的准确率，存在scores中。





1. **参数优化**

网格搜索其实是一种暴力搜索参数的方法，它通过我们指定不同的超参列表进行穷举搜索，并计算每一个超参组合对于模型性能的影响，来获取最优的超参组合。



嵌套交叉验证：在外围循环中，使用分层k折交叉验证将训练数据划分为训练快和测试块，训练块用于传入内部循环进行训练和超参数调优，测试块用于评估模型的最终表现。使用嵌套交叉验证，估计的真实误差与在测试集上的结果几乎一致。



结合网格搜索和交叉验证不仅可以对算法的超参数进行调优，而且还可以根据不同算法的表现来选择最适合的机器学习算法。

1. **分类**

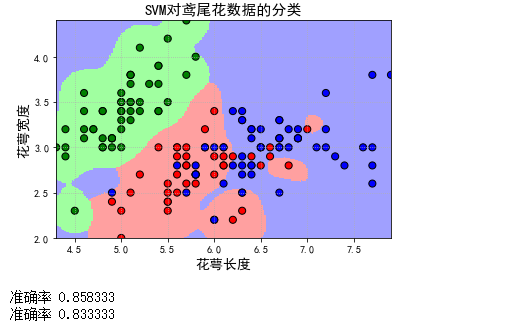
**1.SVM分类：**

数据SVM分类器构建：



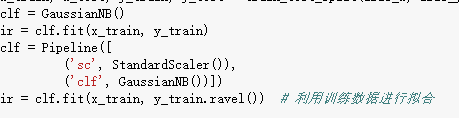
gamma值越大，训练集的拟合就越好，但是会造成过拟合，导致测试集拟合变差

gamma值越小，模型的泛化能力越好，训练集和测试集的拟合相近，但是会导致训练集出现欠拟合问题，从而准确率变低，导致测试集准确率也变低。

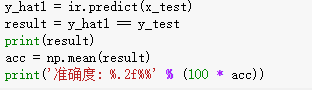


1. **朴素贝叶斯分类**

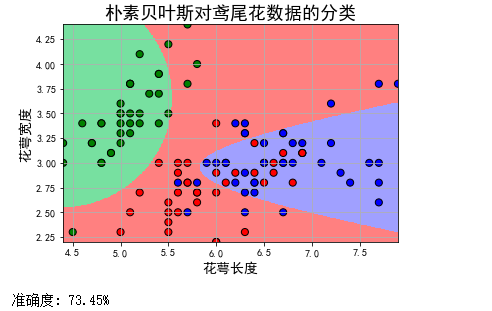
调用高斯朴素贝叶斯实现训练：



对结果进行评估：



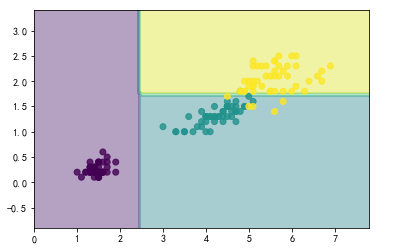
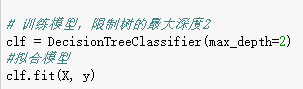
得到图：



1. **决策树分类**

Sklearn机器学习包中，决策树实现类是DecisionTreeClassifier，能够执行数据集的多类分类。输入参数为两个数组X[n\_samples,n\_features]和y[n\_samples],X为训练数据，y为训练数据的标记数据。

构建树模型：



树的深度不一样得到的结果不一样。

