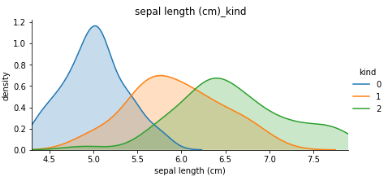
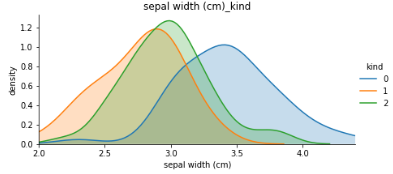
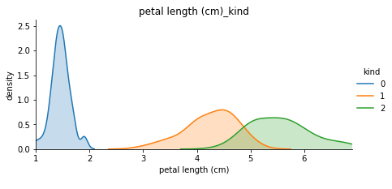
**说明文档**

1. **import packages and dataset**导入包和scikit-learn鸢尾花数据集
2. **Exploratory Data Analysis** 探索性数据分析

根据鸢尾花分类结果，分析各个特征的分布



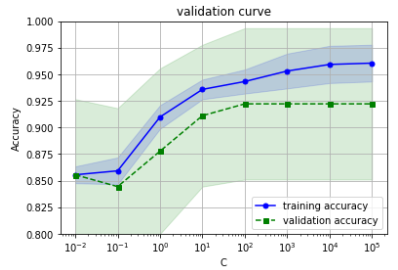




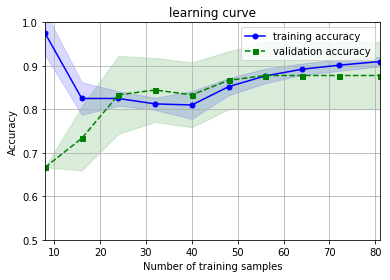
1. **data reduction** 数据降维

均值归一化，提高逻辑回归模型计算速度，方便SVM模型计算

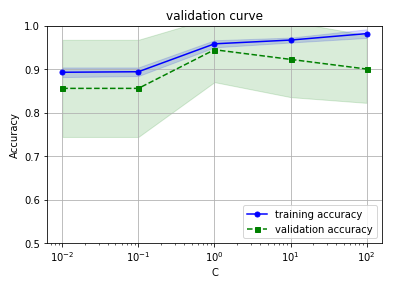
1. **Model Training** 模型训练
   1. **逻辑回归**
      1. 嵌套交叉验证选最优参数
      2. 画验证曲线选最优参数



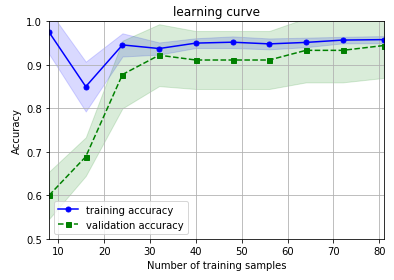
* + 1. 画学习曲线判断是否过拟合或欠拟合



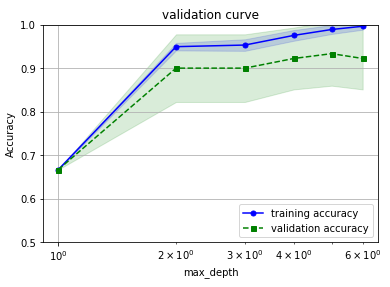
* 1. **支持向量机**
     1. 嵌套交叉验证选最优参数
     2. 画验证曲线选最优参数



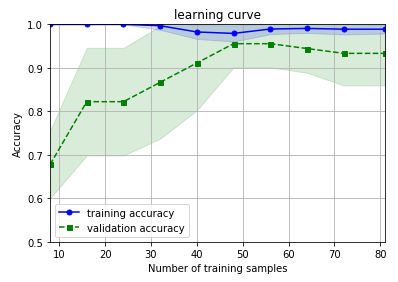
* + 1. 画学习曲线判断是否过拟合或欠拟合



* 1. **决策树**
     1. 嵌套交叉验证选最优参数
     2. 画验证曲线选最优参数



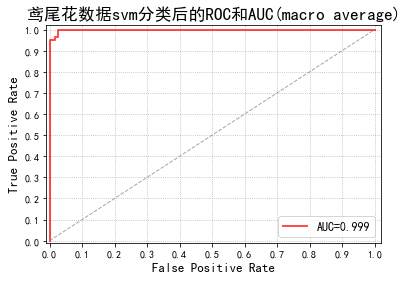
* + 1. 画学习曲线判断是否过拟合或欠拟合

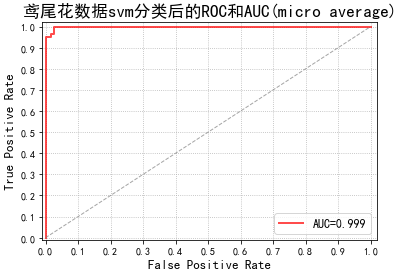


1. **Draw the ROC curve** 画ROC曲线

以测试成绩最优和学习曲线没过拟合或欠拟合的模型SVM为例，画ROC曲线

分别以宏平均和微平均为基础画ROC曲线

macro average

micro average

RPC曲线表现的很好，在误诊率很低的情况下就达到极高的召回率

另外，由于原数据集3种分类分布均匀，所以以宏平均和微平均为基础画的ROC曲线没什么区别

1. **harvest**收获

## 10折交叉验证得到的模型比5折交叉验证得到的模型(exactly parameter)有更好的泛化能力

## 训练集学习曲线先下降后上升的可能原因是数据集太小，导致学习曲线刚开始，训练出的模型受干扰太大

## 多分类画ROC curve，有两种方法1.macro average 2.micro average