**实验4 机器学习实验**

班级： 学号： 姓名：

**一、实验目的**

掌握逻辑回归、支持向量机和K-means聚类等常用机器学习方法的原理，熟悉分类问题和聚类问题的求解流程，理解输入特征个数和不同超参数对机器学习算法的训练和测试结果的影响。

**二、实验原理**

逻辑回归（Logistic Regression）是一种有监督学习方法，属于分类算法，可以处理二元分类以及多元分类，预测结果是离散的分类，例如判断这封邮件是否是垃圾邮件，以及用户是否会点击此广告等。逻辑回归只是对线性回归的计算结果加上了一个Sigmoid函数（Logistic函数），将数值结果转化为0到1之间的概率，根据这个概率可以做预测，例如概率大于0.5，则这封邮件就是垃圾邮件，或者肿瘤是否是恶性的等等。

支持向量机(Support Vector Machine)也是一种有监督学习方法，它尝试寻找一个最优决策边界，使距离两个类别最近的样本最远，从而对分类问题提供良好的泛化能力。支持向量机是由Cortes和Vapnik于1995年首先提出的，在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

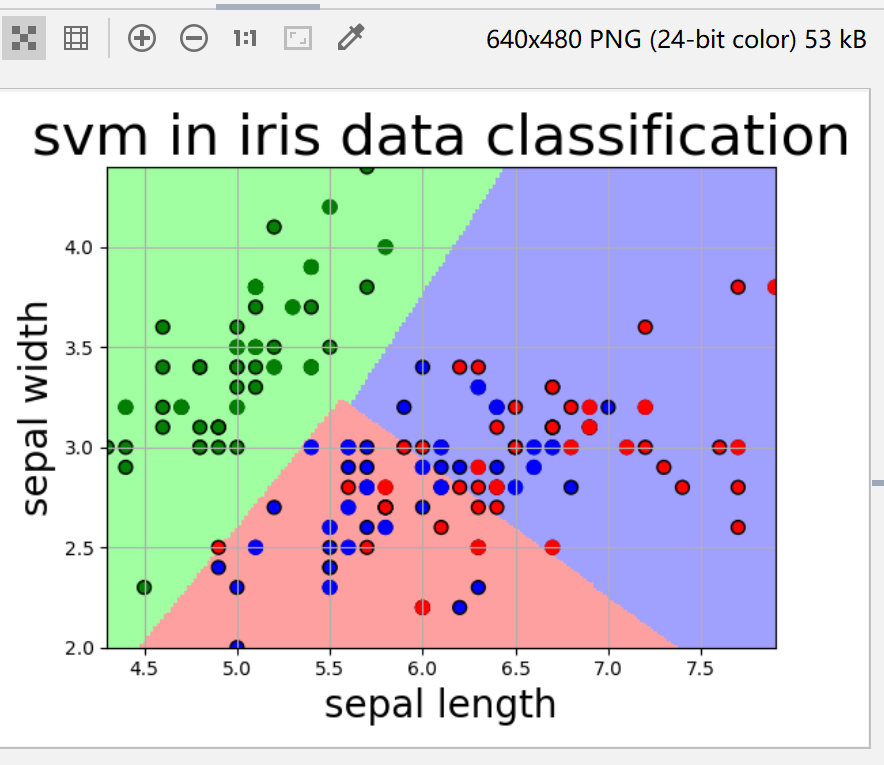
1. means（又称k-均值或k-平均）聚类是一种无监督学习方法。K-means聚类算法的思想是首先随机确定k个中心点作为聚类中心，然后把每个数据点分配给最邻近的中心点，分配完成后形成k个聚类，计算各个聚类的平均中心点，将其作为该聚类新的类中心点，然后重复迭代上述步骤直到分配过程不再产生变化。

**三、实验结果**

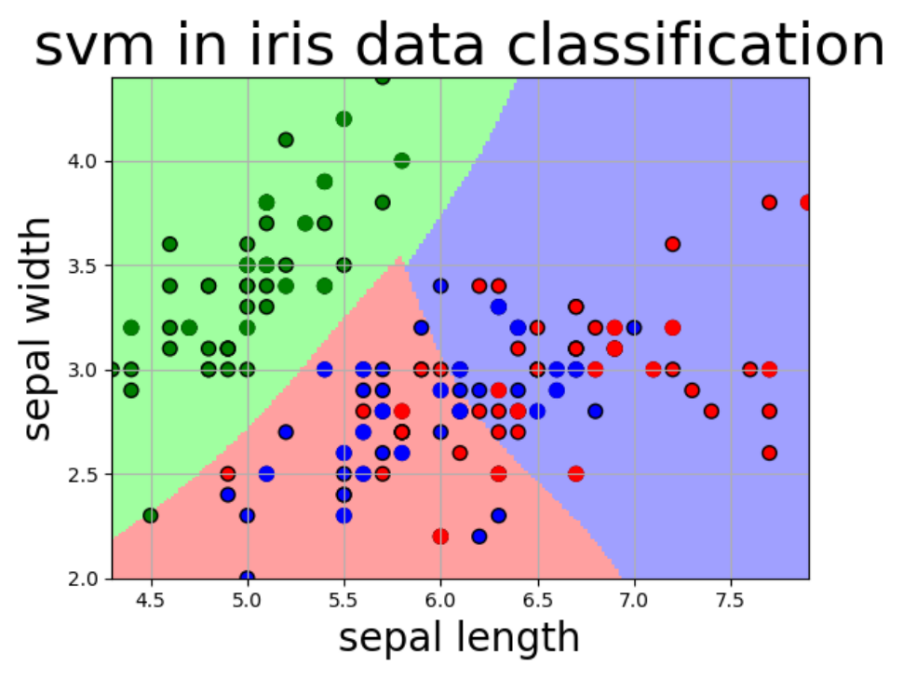
表1 鸢尾花分类结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 分类算法 | 算法参数 | 输入特征个数 | 输入特征名称 | 训练集的准确率 | 测试集的准确率 |
| 逻辑回归 | C=1(正则化系数λ的倒数，默认值为1）  multi\_class='ovr'（默认值） | 2 | 萼片长度萼片宽度 | 0.705 | 0.622 |
| 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.952 | 0.889 |
| 支持向量机  （线性核） | C=0.5（惩罚参数，默认值为1.0）  decision\_function\_shape='ovr' | 2 | 萼片长度萼片宽度 | 0.819 | 0.778 |
| 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.981 | 0.978 |
| 支持向量机  （高斯核） | C=0.5（惩罚参数，默认值为1.0）  gamma=1（高斯核的宽度）  decision\_function\_shape='ovr' | 2 | 萼片长度萼片宽度 | 0.829 | 0.778 |
| 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.981 | 0.978 |

支持向量机（线性）只输入两个特征的分类区域图：



支持向量机（高斯核）只输入两个特征的分类区域图：



逻辑回归只输入两个特征的分类区域图：

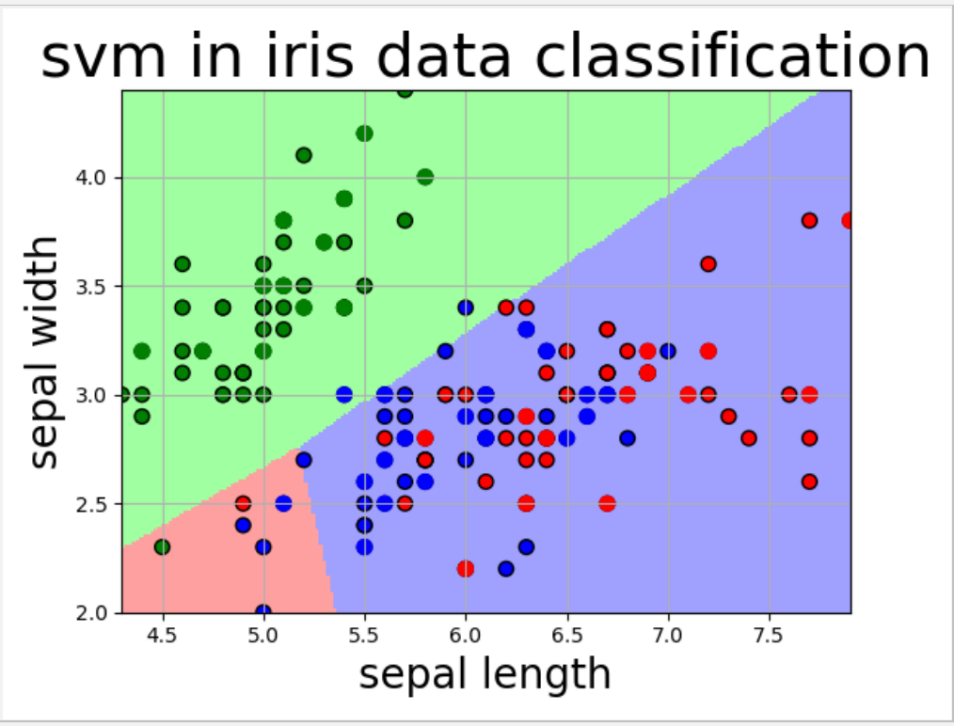


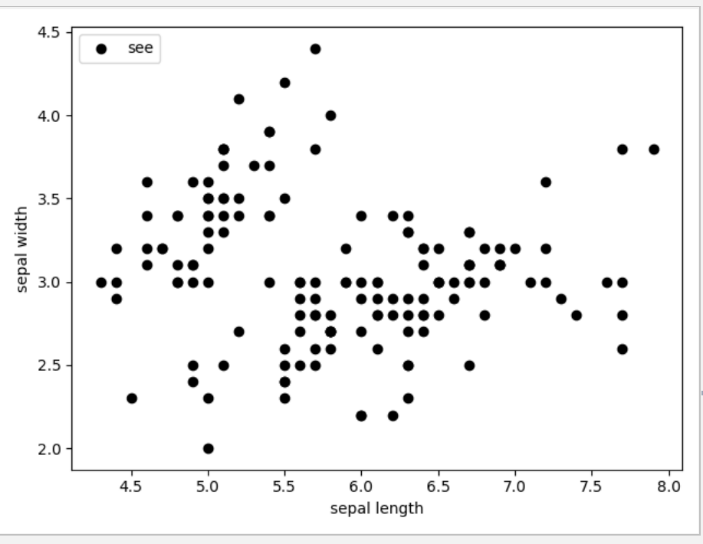
表2 支持向量机（高斯核）的分类结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 分类算法 | 算法参数 | 输入特征个数 | 输入特征名称 | 训练集的准确率 | 测试集的准确率 |
| 支持向量机（线性核） | C=0.5 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.981 | 0.978 |
| C=0.8 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.981 | 1.000 |
| C=1.0 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.981 | 1.000 |
| C=5.0 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.981 | 0.978 |
| C=10.0 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.971 | 0.978 |
| 支持向量机（高斯核） | C=0.8  gamma=0.1 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.971 | 0.978 |
| C=0.8  gamma=1 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.981 | 0.956 |
| C=0.8  gamma=10 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 1.000 | 0.956 |
| C=1  gamma=10 | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 1.000 | 0.933 |

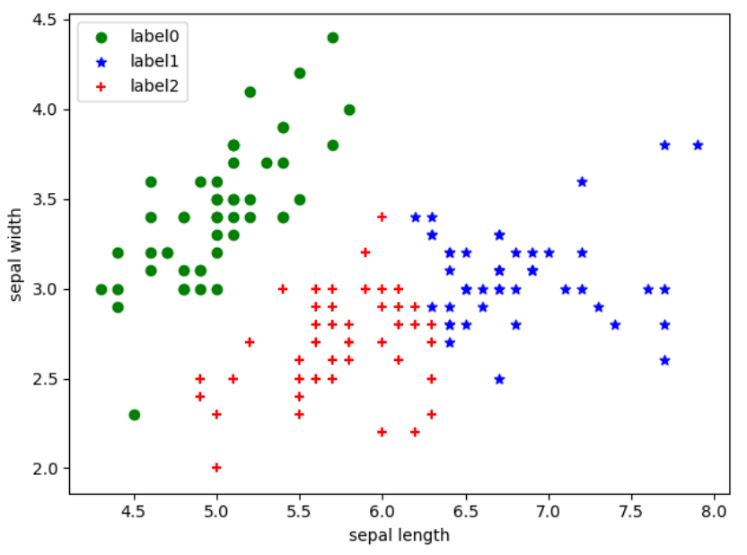
表3 基于k-均值的鸢尾花聚类结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 初始聚类 | 输入特征个数 | 输入特征名称 | 轮廓系数 | 正确率 |
| 3 | 2 | 萼片长度萼片宽度 | 0.445 | 0.1 |
| 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.553 | 0.893 |
| 2 | 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.681 | 0.02 |
| 4 | 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.497 | 0.153 |

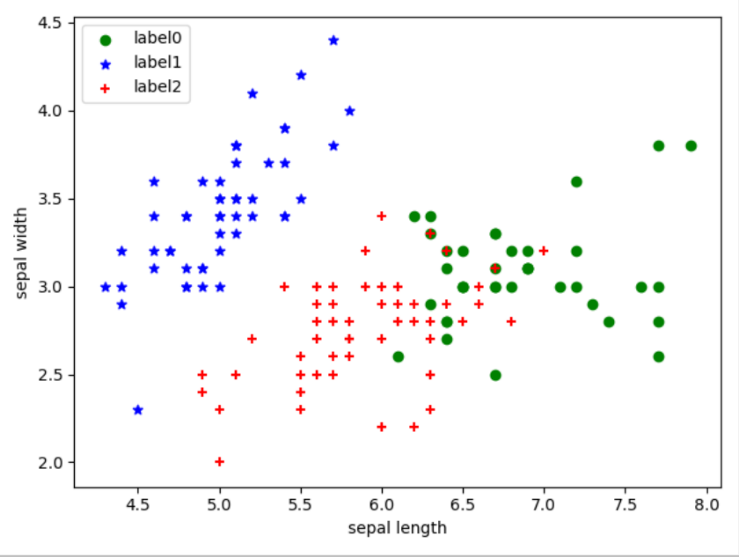
原数据集:



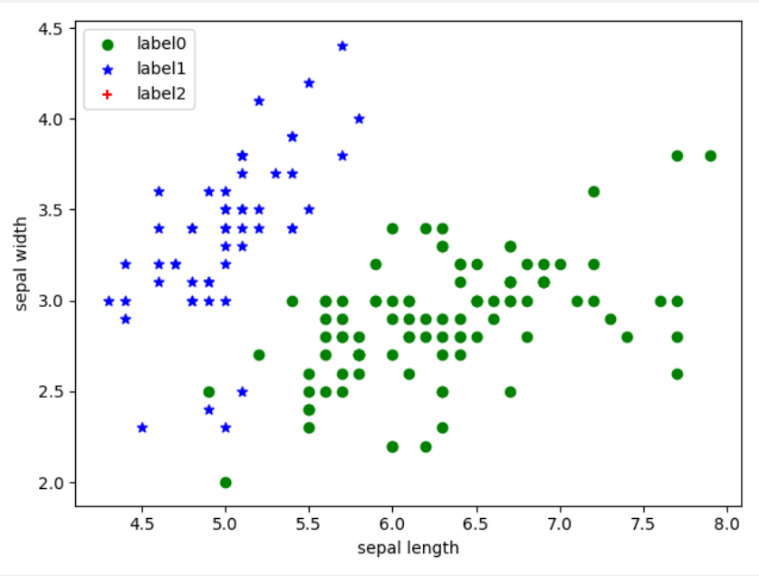
初始聚类3，特征个数2：



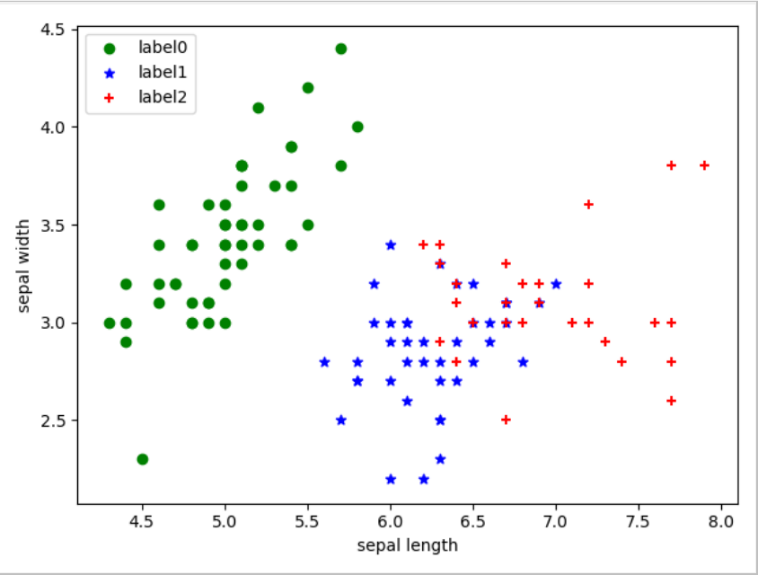
初始聚类3，特征个数4：



初始聚类2，特征个数4：



初始聚类4，特征个数4：



**四、实验总结**

1、输入特征个数越多，机器能够学习到的越多，从而能够更精确的判断，得到的准确性就越高。

2、当观测样本较多时，逻辑回归的效率较高，但当特征个数较多时SVM的性能较好

3、n：样本 m：特征  
①相对于n，m很大。比如m≥n, m=10000, n=10~1000,即(m/n)>10。  
考虑“linear”  
②m很小，n一般大小。比如m=1-1000, n=10~10000,即(m/n)在[0.0001,100].  
考虑“rbf”  
③m很小，n很大。比如n=1-1000，m=50000+，即(m/n)在[~,0.02].  
增加m的量，考虑“linear”

④算法的超参数较高时考虑“linear”

4、特征个数相同时，初始聚类数越多，正确性就越低

5、心得体会：这实验太难搞懂了

**逻辑回归（LR）与支持向量机（SVM）的区别和联系**

**相同点：**

1. LR和SVM都是分类算法。
2. 如果不考虑核函数，LR和SVM都是线性分类算法，也就是说他们的分类决策面都是线性的。
3. LR和SVM都是监督学习算法。
4. LR和SVM在学术界和工业界都广为人知并且应用广泛。

**不同点：**

1. 损失函数是二者最为本质的区别。
2. 两个模型对数据和参数的敏感程度不同 。
3. SVM 基于距离分类，LR 基于概率分类。
4. 在解决非线性问题时，支持向量机采用核函数的机制，而LR通常不采用核函数的方法。
5. 在小规模数据集上，Linear SVM要略好于LR，但差别也不是特别大，而且Linear SVM的计算复杂度受数据量限制，对海量数据LR使用更加广泛。