**基于SVM的鸢尾花分类**

18 人智 于松松

**一 数据集信息：**

··是一个很小的数据集，仅有150行，5列，每类有50个数据。该数据集的四个特征属性的取值都是数值型的，他们具有相同的量纲，不需要做任何标准化的处理，第五列为通过前面四列所确定的鸢尾花所属的类别名称。

·数据集处理：为了增加实验的科学性，随机将样本打乱了顺序，进行重新排列。

**二 2分类（基于前两个特征）：**

首先完成课程的任务，即根据前一百条数据的前两个特征完成学习，进行对鸢尾花的有效分类，并标注支持向量以及分离超平面效果如图2-1，图中红色与绿色代表正负样例，品红色为支持向量，直线为分离超平面。

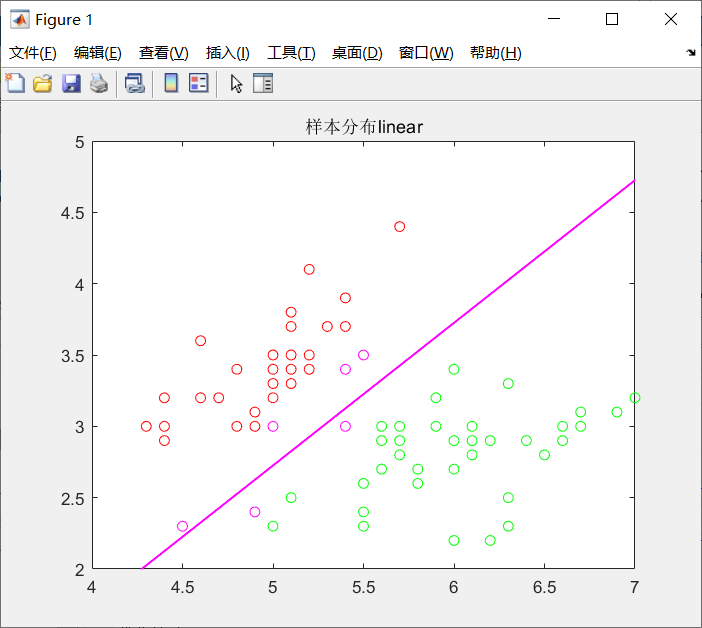


图2-1 基于前一百个样本两个特征的分类

为了演示方便（主要是为了在图中显示更多的样例），在这里我选取了0.8 : 0.2的样本比例分别用于训练和测试，采用核函数为’linear’ ( 公式见式2.1 ) 由于鸢尾花数据集为线性可分的，准确率达到了百分之百，降低训练测试比到 0.5 : 0.5之后，准确率如图2-2，仍为百分之百。

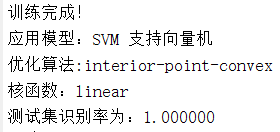


图2-2 识别准确率

 2.1

**三 2分类（基于四个特征）：**

对于依靠前两个特征，已经能够将 'setosa' 'versicolor'两类正确分类，现尝试利用四个特征对其进行分类，来验证在有冗余分类特征的前提下，支持向量机是否还能应对。采用核函数仍然为’linear’，结果如图3-1，准确率如图3-2，增加数据特征维数之后，表现仍然良好，这是容易解释的，因为鸢尾花数据集本身就是线性可分的。

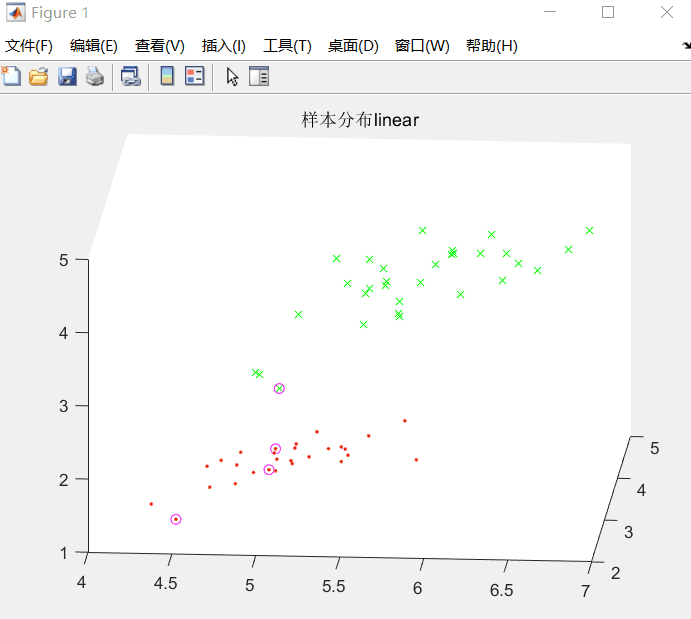


图 3-1 基于四特征的二分类

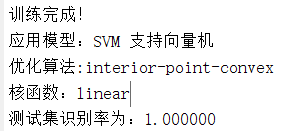


图 3-2 四特征的准确率

**参考：**

鸢尾花（iris）数据集

<https://www.gairuo.com/p/iris-dataset>