《统计学习方法》第七章读书笔记

一、对自己提出问题的理解：

1. **提出的问题1：LR和SVM有什么异同？**

**讨论后的理解**：**相同点**1）都是有监督的分类算法；2）都是线性分类方法，两者都可以加核函数；3）都是判别模型。

**不同点：**

1. **损失函数**不同。LR的损失函数是**交叉熵**，SVM的损失函数是**最大化间隔距离**。
2. 不同的损失函数在实质上代表了**不同的假设前提**，也就代表了**不同的分类原理**。LR方法基于**概率理论**，假设样本为0或者1的概率可以用sigmoid函数来表示，然后通过极大似然估计的方法估计出参数的值。SVM基于**几何**间隔最大化原理，认为存在最大几何间隔的分类面为最优分类面，所以SVM只考虑分类面上的点，而LR考虑所有点，SVM中，在支持向量之外添加减少任何点都对结果没有影响，而LR则是每一个点都会影响决策。
3. **实际应用**的区别。根据经验，对于小规模数据集，SVM的效果要好于LR，但是大数据中，SVM的计算复杂度受到限制，而LR因为训练简单，可以在线训练，所以经常会被大量采用。
4. **提出的问题2： SVM的对偶形式为什么能够降低计算量？**

**讨论后的理解**：由于优化函数的变化。原问题的约束方程数对应于对偶问题的变量数, 而原问题的变量数对应于对偶问题的约束方程数, 而约束方程数目越少, 优化问题求解的复杂度越低；在线性SVM的原问题中，样本量为N；但在优化目标函数中只有1个w，有 N个限制条件；在线性SVM的对偶问题中，只有1个限制条件；当训练样本总数不大，特征空间的维度d>>n时，选择在对偶问题中求解将有效减少计算量。

1. **提出的问题3：SVM在大数据上应用时有哪些缺陷？**

**讨论后的理解**：SVM的空间消耗主要是在存储训练样本和核矩阵，由于SVM是借助二次规划来求解支持向量，而求解二次规划将涉及m阶矩阵的计算（m为样本的个数），当m数目很大时该矩阵的存储和计算将耗费大量的内存和运算时间。如果数据量很大，SVM的训练时间就会比较长，所以SVM在大数据的使用中比较受限。

二、别人提出的问题的理解

1. **问题1：SVM的优缺点？**

**个人的理解：**

优点：1）非线性映射是SVM方法的理论基础，SVM利用内积核函数代替向高维空间的非线性映射；

2）对特征空间划分的最优超平面是SVM的目标，最大化分类边际的思想是SVM方法的核心；

3）SVM 的最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾难”。

缺点：

1）SVM算法对大规模训练样本难以实施。由于SVM是借助二次规划来求解支持向量，而求解二次规划将涉及m阶矩阵的计算（m为样本的个数），当m数目很大时该矩阵的存储和计算将耗费大量的机器内存和运算时间（见“自己提出的问题3”）。

2）用SVM解决多分类问题存在困难。传统的SVM就是解决二分类问题的，虽然目前存在解决多分类问题的SVM技巧，不过各种方法都一定程度上的缺陷。

3）对缺失值敏感，核函数的选择以及调参比较复杂。

1. 问题2：不同的核函数各自有什么样的应用场景？

**个人的理解：**1）线性核函数，主要用于线性可分情况，一般情况下，对于特征数量相对于样本数量非常多时，适合线性核函数。

2）多项式核函数，低维空间线性不可分的模式通过非线性映射到高维特征空间则可能实现线性可分，而多项式核函数避免了高维展开的维度灾难。对于正交归一化后的数据，可以优先选择此函数。

3）高斯核函数：指数的计算量大，参数的挑选十分重要，在特征线性可分性较差的情况下，效果较好。

4）径向基核函数，灵活性强，在不确定使用什么核函数时，可以优选选择此核函数。

5）sigmoid核函数，SVM使用此核函数，相当于一个两层的感知机

三、读书计划

1. 本周完成的内容章节：第七章。
2. 下周计划：详细学习原问题与对偶问题，学习第八章。

四、3种算法的性能比较

在iris.arff数据集上，分别采用决策树、朴素贝叶斯、SVM这3种方法进行分类，分析各个分类器所得结果并进行比较。

1. 决策树分类方法

使用决策树分类得到的准确率为96%，其中150个实例中的144个被正确分类，6个被错误分类。若将Iris Setosa、Iris Versicolour和Iris Virginica三种原始类别分别记为a、b、c三类，根据混淆矩阵，被错误分类的实例如下：1个b类实例被错误分类到a，2个b类实例被错误分类到c，3个c类实例被错误分类到b，可视化结果如图1所示。该算法P= 0.96，R= 0.96，ROC面积为0.968。

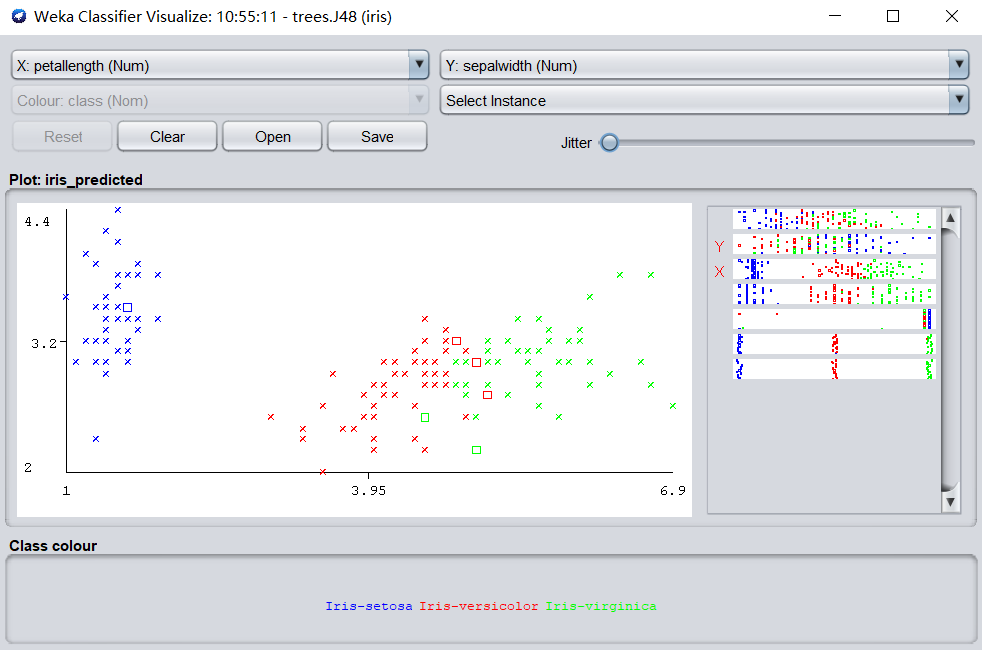


图1 决策树分类结果可视化图

1. 朴素贝叶斯分类方法

使用朴素贝叶斯分类得到的准确率为96%，其中150个实例中的144个被正确分类，6个被错误分类。若将Iris Setosa、Iris Versicolour和Iris Virginica三种原始类别分别记为a、b、c三类，根据混淆矩阵，被错误分类的实例如下：4个b类实例被错误分类到c；2个c类实例被错误分类到b。该算法P= 0.96，R= 0.96，ROC面积为0.994，Iris Versicolour的ROC曲线如图2所示。

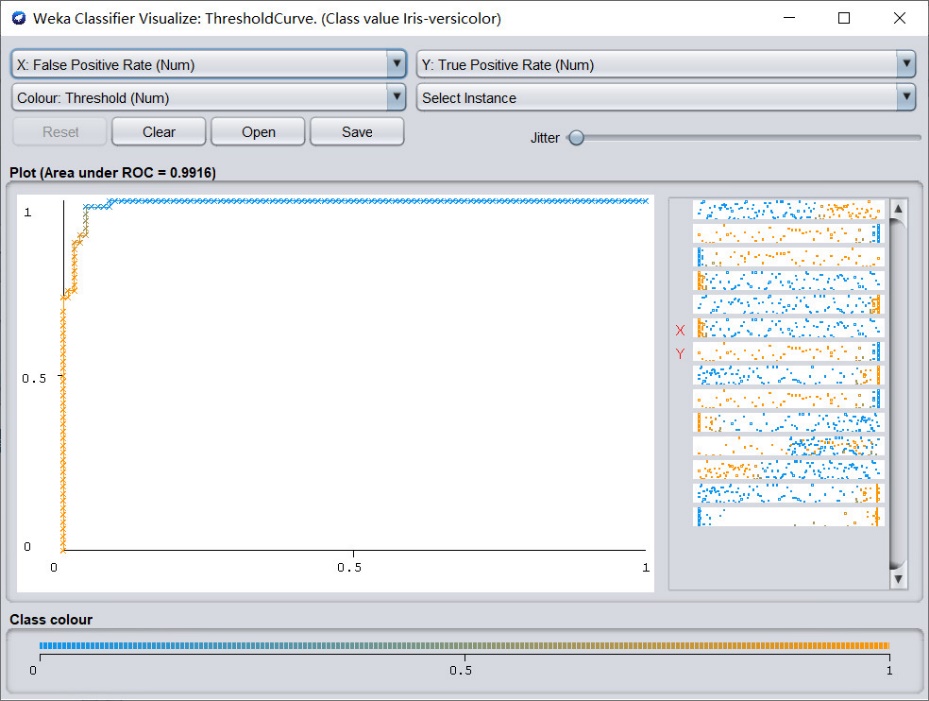


图2 朴素贝叶斯分类方法下Iris Versicolour的ROC曲线图

1. SVM分类方法

使用SVM分类得到的准确率为96.67%，其中150个实例中的145个被正确分类，5个被错误分类。若将Iris Setosa、Iris Versicolour和Iris Virginica三种原始类别分别记为a、b、c三类，根据混淆矩阵，被错误分类的实例如下：2个b类实例被错误分类到c；3个c类实例被错误分类到b，类似的，可以观察分类结果可视化图像和ROC曲线。该算法P= 0.967，R= 0.967，ROC面积为0.975。

1. 四种分类方法比较

一般对于多分类（大于两类）问题，由于不存在阳性和阴性，无法计算真阳性TP、真阴性TN、假阳性FP、假阴性FN等评估度量，通常使用准确率进行衡量，不考虑其它指标，计算公式如下：



其中，表示混淆矩阵中预测类别与实际类别一致的元组个数，表示参与分类的元组总个数。如果是类不平衡问题则需要重新设计准确率的公式，鸢尾花数据不属于类不平衡，可直接使用上述公式，三个分类器度量指标如表1所示。

表1 三个分类器的度量指标

|  |  |  |  |
| --- | --- | --- | --- |
| **分类器**  **度量指标** | **决策树** | **朴素贝叶斯** | **LibSVM** |
| **混淆矩阵** |  |  |  |
| **准确率** | 96% | 96% | 96.67% |

从表中数据可以看出，在本次实验的鸢尾花数据集上，SVM分类器有着更好的分类性能。