模式识别 支持向量机

22009200601汤栋文

2024年12月08日

目录

[摘要 2](#_Toc184583060)

[1. 方法 2](#_Toc184583061)

[1.1 数据集 2](#_Toc184583062)

[1.2 线性可分支持向量机 v.s. 线性支持向量机 2](#_Toc184583063)

[1.3 实验设置 2](#_Toc184583064)

[2. 实验 3](#_Toc184583065)

[1.1 SVM算法验证与核函数比较 3](#_Toc184583066)

[1.2 支持向量可视化 3](#_Toc184583067)

[3. 结论 4](#_Toc184583068)

# 摘要

本实验旨在验证支持向量机（SVM）算法在Sonar数据集和Iris数据集上的分类性能。我们采用了线性核函数、高斯核函数和多项式核函数来比较不同核函数对分类结果的影响。数据集通过标准化处理后，划分为训练集和测试集，并利用SVM进行分类实验，最后利用Iris类别合并和降维构造线性可分数据，对支持向量进行可视化。实验结果显示，不同核函数在Iris数据集和Sonar数据集上的表现各有差异，但总体好于线性核。

# 方法

## 数据集

本研究使用了Sonar数据集和Iris数据集。其中Sonar数据集包含了60个特征（声纳信号返回的强度），并且每个样本都标记为岩石或矿石。Iris数据集包含了鸢尾花属植物三个种类共150个样本的测量数据，每种类型的鸢尾花各有50个样本，每个样本记录了四个特征：萼片长度、萼片宽度、花瓣长度、花瓣宽度，单位均为厘米。

为了消除特征间的数据分布影响，我们对原始特征进行了标准化处理，即减去平均值并除以标准差，使得每个特征具有零均值和单位方差。这也使得各种方法之间的对比完全公平。

## 线性可分支持向量机 v.s. 线性支持向量机

**线性可分SVM:** 适用于数据点之间线性可分的情况，即可以找到一个超平面将不同类别的数据点完全分开。在这种情况下，不需要引入松弛变量，因为数据是完全可分的。**线性SVM:** 适用于数据点之间线性不可分的情况，即数据点不能用一条直线完全分开。为了处理这种情况，引入了一个松弛变量和惩罚因子，以允许一些错误分类并找到最优的超平面。**相互转换：**将惩罚因子设置为一个非常大的值，这意味着模型对误分类的容忍度非常低，因此会倾向于找到一个可以完全分开训练数据的超平面，这时如果数据确实满足线性可分，那么此时的线性SVM等价于线性可分SVM。

## 实验设置

1. **在Sonar数据集和Iris数据集上验证SVM算法：**使用线性核函数验证Sonar数据集和Iris数据集上分类的准确率。我将数据集按70%训练30%测试的方式划分为测试集和训练集合，并以准确率作为评价指标。
2. **使用三种不同核函数进行比较：**我分别使用了线性核函数，高斯核函数，以及多项式核函数来验证不同核函数下分类准确率会有什么变化。
3. **支持向量的可视化：**我将Iris数据集进行类别合并从而保证了线性可分，并使用线性可分SVM进行二分类，最后对支持向量进行了可视化。

# 实验

## SVM算法验证与核函数比较

这个表格展示了不同方法在Iris和Sonar数据集上的分类准确率。我分别使用了线性核函数，高斯核函数，以及多项式核函数来验证不同核函数下分类准确率会有什么变化。他们的数学表达形式如下：

Linear:

Gaussian:

Polynomial:

线性方法和Gaussian核函数在Iris数据集上表现相似，但在Sonar数据集上，Gaussian核函数的表现随着gamma值的增加而提升。Polynomial核函数在不同degree值下表现各异，总体上高于线性方法。实验结果如下：

|  |  |  |
| --- | --- | --- |
| **Method** | **Iris** | **Sonar** |
| Linear | 97.8% | 81.0% |
| Gaussian (gamma=0.1) | 97.8% | 74.6% |
| Gaussian (gamma=0.5) | 97.8% | 82.5% |
| Gaussian (gamma=1.0) | 97.8% | 85.7% |
| Gaussian (gamma=2.0) | 97.8% | 87.3% |
| Polynomial (degree=1.0) | 97.8% | 81.0% |
| Polynomial (degree=2.0) | 95.6% | 87.3% |
| Polynomial (degree=3.0) | 97.8% | 82.5% |

## 支持向量可视化

首先，我将Iris数据集中的类别进行了合并，以确保数据集可以通过线性方式分开，我将这三种类别合并成两种类别。接下来，我使用主成分分析（PCA）对数据集进行了降维处理。在Iris数据集中，我将数据从四维降到了二维，以便可视化结果与SVM算法看到的结果完全对齐。然后，我在降维后的二维数据上应用了线性可分支持向量机。支持向量是那些在决策边界上或者靠近决策边界的数据点，它们对确定决策边界起着关键作用。通过可视化支持向量，我们可以更直观地理解SVM的工作原理和分类效果。

# 结论

在简单数据集上，线性核函数、高斯核函数和多项式核函数均能取得较高的分类准确率，差异不显著。在稍微复杂的数据集上，高斯核函数和多项式核函数表现优于线性核函数，但需要少许的调参。综上所述，选择合适的核函数有助于提升SVM的分类准确率。