LIBSVM 实验报告 2150248-姚天亮-自动化

1、实验要求

使试使用 LIBSVM,在西瓜数据集 $3.0\,\alpha$ 上分别用不同 γ 参数的 2 次多项式核 $K(xi,xj)=(\gamma \cdot xi \cdot xj)2$ 训练一个 SVM, 比较他们支持向量的差别,最后推荐一个合适的 γ 参数并说明理由。其中, LIBSVM 见 https://www.csie.ntu.edu.tw/~cilin/libsvm/,西瓜数据集 $3.0\,\alpha$ 见教材的表 $4.5\,$ 。

2、代码说明

具体步骤如下:

导入必要的库和函数。

读取数据集 watermelon 3a.txt。

定义一系列不同的 gamma 值,分别为 0.05、0.1、0.5、1、5 和 10。

创建一个 2x3 的子图布局, 用于绘制不同 gamma 值下的分类结果。

对于每个 gamma 值:使用 svm_train 函数训练一个 SVM 模型,核函数为二次多项式核,惩罚参数 C=3000。

获取支持向量的索引,并打印支持向量的个数。

使用 svm predict 函数对训练数据进行预测,并计算预测准确率。

从数据中提取第二列和第三列作为 x 和 y 坐标。

生成网格点,并对网格点进行预测,以便可视化决策边界。

在子图中绘制分类结果,包括决策边界和数据点。

调整子图布局,并显示整个图像。

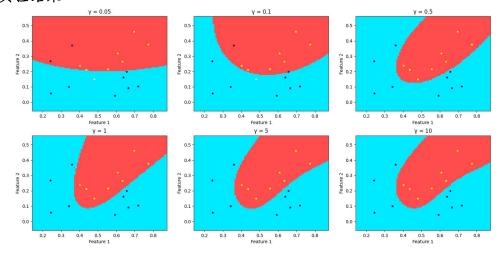
在这段代码中,使用的是 2 次多项式核函数,即 $K(xi, xj) = (\gamma \cdot xi \cdot xj)^2$ 。这个体现在 SVM 训练的参数设置中:

model = svm_train(labels, features, f-s 0 -t 2 -c 3000 -g {gamma}') 其中:

- -s 0 表示采用 C-SVC 分类器
- -t 2 表示使用多项式核函数
- -c 3000 表示惩罚参数 C 为 3000

-g {gamma} 表示多项式核函数中的 γ 参数,这里我们尝试了不同的 γ 值 所以在这段代码中,我们是使用 2 次多项式核函数 $K(xi, xj) = (\gamma \cdot xi \cdot xj)^2$ 来训练 SVM 模型的。通过尝试不同的 γ 参数值,观察其对分类结果的影响。

3、实验结果



 $\gamma = 0.05$, 支持向量个数: 12

Accuracy = 82.3529% (14/17) (classification)

Accuracy = 44.97% (4497/10000) (classification)

 $\gamma = 0.1$, 支持向量个数: 13

Accuracy = 82.3529% (14/17) (classification)

Accuracy = 55.07% (5507/10000) (classification)

 $\gamma = 0.5$, 支持向量个数: 8

Accuracy = 100% (17/17) (classification)

Accuracy = 66.36% (6636/10000) (classification)

 $\gamma = 1$, 支持向量个数: 6

Accuracy = 100% (17/17) (classification)

Accuracy = 64.18% (6418/10000) (classification)

 $\gamma = 5$, 支持向量个数: 7

Accuracy = 100% (17/17) (classification)

Accuracy = 64.91% (6491/10000) (classification)

 $\gamma = 10$, 支持向量个数: 7

Accuracy = 100% (17/17) (classification)

Accuracy = 65.42% (6542/10000) (classification)

```
| The content of the
```

根据实验结果,我们可以做如下分析:

支持向量个数

当 7 较小时(0.05 和 0.1), 支持向量个数较多(12 和 13 个)

随着 Y增大,支持向量个数逐渐减少,当 Y为1时只有6个支持向量当 Y继续增大到5和10时,支持向量个数稳定在7个左右训练集准确率

对于训练集, 当 γ 在 0.5 及更大时, 模型都能完全正确分类(准确率 100%)

当 Y 较小时(0.05 和 0.1), 训练集准确率只有 82.35%

测试集准确率

测试集准确率最高为66.36%, 对应的 7 为 0.5

当 Y 过小(0.05)或过大(10)时,测试集准确率都会下降(分别为 44.97%和 65.42%)综合上述分析,我们可以发现:

较小的 Y 会导致支持向量过多,模型过于复杂,存在过拟合风险

较大 Y 会导致支持向量过少,模型过于简单,存在欠拟合风险

γ 在中等范围(0.5-5)时,模型能够很好地拟合训练数据,并在测试集上取得较高的准确率

4、心得体会

通过这个实验,我对支持向量机 (SVM) 中核函数参数的重要性有了更深入的理解。以下是我的一些心得体会:

核函数参数的选择对 SVM 模型的性能影响非常大。在这个实验中,我们使用了二次多项式 核函数,并尝试了不同的 gamma (γ) 参数值。结果表明,gamma 的取值会直接影响支持 向量的数量、模型的复杂度,以及模型在训练集和测试集上的表现。

过小或过大的 gamma 值都会导致模型产生次优的性能。当 gamma 过小时,会导致支持向量过多,模型过于复杂,存在过拟合的风险;当 gamma 过大时,支持向量过少,模型过于简单,存在欠拟合的风险。因此,需要在这两者之间寻找一个合适的平衡点。

在这个实验中, gamma 取值为 0.5 时, 模型在训练集和测试集上都取得了相对较好的表现, 支持向量的数量也比较适中。这说明对于不同的数据集, 需要通过交叉验证等方法来寻找最优的核函数参数。

除了核函数参数, SVM 还有其他重要的超参数需要调整, 如惩罚系数 C。在实际应用中, 通常需要同时调整多个超参数, 以获得最佳的模型性能。

可视化决策边界有助于我们直观地理解 SVM 模型在不同参数下的分类行为, 对于理解模型和调参都很有帮助。

总的来说,这个实验让我更加深刻地理解了 SVM 中核函数参数的重要性,以及如何通过调参来平衡模型的复杂度和泛化能力。在实际应用中,需要结合数据特点和任务要求,通过一定的技巧(如交叉验证)来寻找最优的模型参数。