SVM 作业报告

201250182 郑义

SVM 算法原理

1 线性可分

首先介绍一下线性可分的概念，线性可分在直观的概念理解上指的就是：在二维空间上，两类点被一条直线完全分开就叫做线性可分。

2 支持向量

支持向量和超平面有关，超平面指的是将两类点完全划分开的 wx + b = 0 就成为了一个超平面。

为了使这个超平面更具有鲁棒性，我们会去找最佳超平面，以最大间隔把两类样本分开的超平面，也称之为最大间隔超平面。

- 两类样本分别分割在该超平面的两侧；
- 两侧距离超平面最近的样本点到超平面的距离被最大化了

在样本中距离超平面最近的一些点，这些点叫做支持向量，SVM 要做的就是尝试寻找一个最优的决策边界，让两个类别最近的样本相距最远。即要最大化支持向量。

之后就是通过一系列的优化方法想办法通过数学方法来最大化这个支持向量距离。

作业步骤说明

使用的是 Breast Cancer Wisconsin (Diagnostic) Data Set 这个数据集，对癌症进行预测。该数据集在 sklearn 有直接提供

1. 获取数据，划分训练集和测试集
2. 通过标准化转化训练集和测试集（数据预处理）
3. 通过 sklearn 提供的 SVM API 构建支持向量机，通过训练集进行训练，然后对测试集进行预测
4. 输出预测结果、准确率和相关的评价指标

运行结果和截图

详情可以看 `/src/svm.ipynb` 中的结果