**第5章 支持向量机**

**一、简述数据预处理相关典型技术**（逻辑要通，**要有小标题**，要排版清晰，请不要网络大幅摘抄）。

1. **数据清洗**
   1. **将部分数据突变值，超出范围的值进行删除；**
   2. **对于缺失数据使用中位数、均值众数等数字特征进行填充缺失数据；**
   3. **去除重复值；**
2. **数据集成**
   1. **合并来自于不同数据源的数据**
   2. **处理不一致的数据格式等**
3. **数据变换**
   1. **归一化/标准化：将数据缩放到特定范围（如0到1）或转换为均值为0、标准差为1的分布。**
   2. **离散化：将连续数据分段为离散区间。**

**二、简述支持向量机**的思路、关键操作（不超过300字，逻辑要通，**要有小标题**，要排版清晰，请不要网络大幅摘抄）

**支持向量机（Support Vector Machine, SVM）是一种监督学习算法，广泛应用于分类和回归问题。其核心思想是通过找到一个最优超平面，将不同类别的数据点分开，并且在分类问题中最大化类别之间的间隔。SVM的目标是找到一个超平面，使得不同类别的数据点能够被正确分类，并且距离超平面最近的点的距离（即间隔）最大化。这些距离超平面最近的点被称为支持向量。**

**概念介绍：**

**超平面：指将两类样本点划分的直线或者平面称为超平面。**

**线性可分 ：指简单样本集合中可以将两类不同的样本，通过一个超平面进行分离，即正例和负例分别分布在超平面的两侧。**

**感知机模型：即对于样本点特征值，根据所给出的函数将该样本点的特征值输入后，通过函数输出其分类（或称之为正例或者反例）的模型。**

**三、利用SVM解决问题（回归或分类）。**

附：Sklearn常用数据集

|  |  |  |
| --- | --- | --- |
| 序号 | 数据集 | 学号尾号 |
| 1 | 鸢尾花数据集：load\_iris（） | 全做 调参 |
| 2 | 手写数字数据集：load\_digits（） | 0 1 2 |
| 3 | 乳腺癌数据集load-barest-cancer（） | 6 7 |
| 4 | 糖尿病数据集：load-diabetes（） | 8 9 3 |
| 5 | 波士顿房价数据集：load-boston（） | 4 5 |

利用SVM对以上数据集（学号尾号不同采用不同数据集）进行分析，找出最佳参数下的SVM更适合本数据集（准确率更高，不同回归、不同参数下的不同准确率要列出）。

要求：

1）针对当前数据、模型，**详叙数据预处理、模型参数调整过程及效果比对**；

2）代码要列出，重点代码加注释说明，特别是自己调试过程中的自我理解；

3）运行结果要截图，结果要文字说明；

4）注意排版

**四、[附加题].**利用支持向量机解决**自己遇到的问题**。

[解题要求同题2]