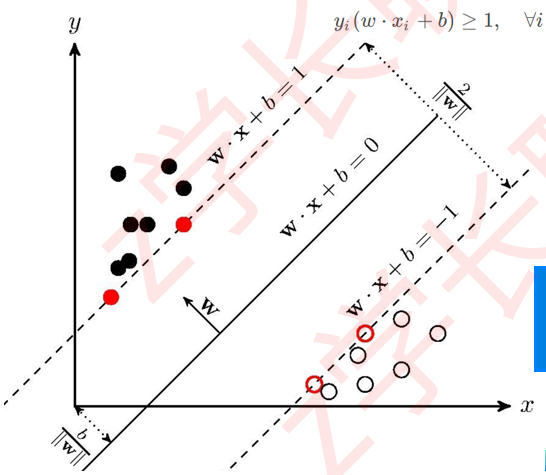
01

支持向量机分类回归与实践

一、支持向量机分类回归与实践

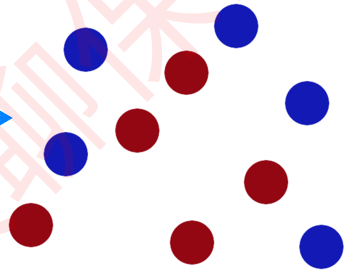
1. 理论介绍-实际案例



支持向量机是一种用于分类和回归的**监督学习**算法，其核心思想是**找到一个最优超平面，以最大化数据分类的间隔。**

为了找到这个最佳的超平面，SVM使用支持向量，即离超平面最近的一些数据点。这些支持向量决定了超平面的位置和方向。

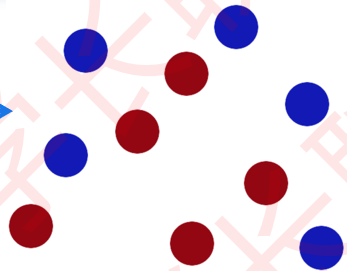
若是线性不可分的?



一、支持向量机分类回归与实践

1. 理论介绍-实际案例

若是线性不可分的?



这时，就引入**松弛变量**，允许一定的分类错误

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i$$

$$\min_{w,b} \frac{1}{2}||w||^2$$

$$min_{w,b,\xi} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

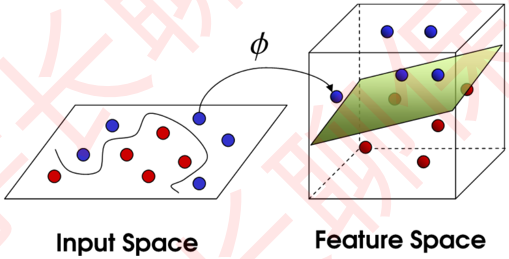
SVM 的优势在于它**可以处理高维数据集，同时具有较强的泛化能力**。它适用于线性和非线性分类问题，可以通过使用不同的核函数来处理非线性关系。

常见的核函数包括线性核、多项式核、径向基函数（RBF）核等。

一、支持向量机分类回归与实践

1. 理论介绍-实际案例

若是复杂且线性不可分的呢?



对于复杂数据，SVM 采用**核方法（引入核函数）**，将低维数据映射到高维空间，使其在高维空间线性可分。

常见核函数包括：

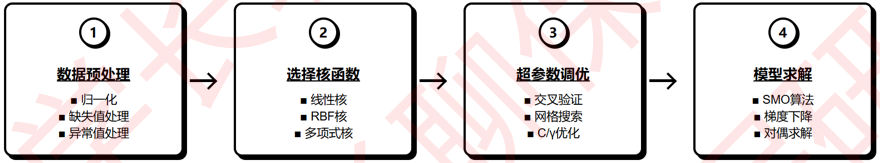
- **线性核**：$K(x_i, x_j) = x_i \cdot x_j$
- **多项式核**：$K(x_i, x_j) = (x_i \cdot x_j + c)^d$
- **高斯核（RBF 核）**：$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$

通过核函数，SVM 在高维特征空间找到最优超平面，实现非线性分类。

一、支持向量机分类回归与实践

1. 理论介绍-实际案例

训练SVM的具体步骤（评估指标与其他分类任务一致）：

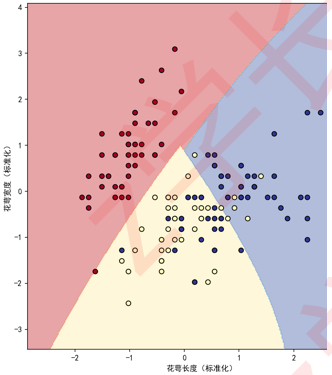


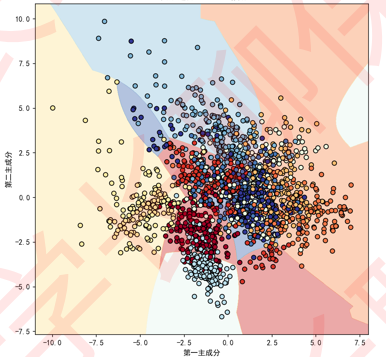
推荐选择：

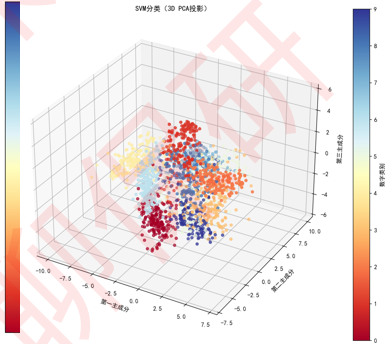
- **小规模数据集（<10,000样本）：** 优先使用交叉验证 + 网格搜索 + SMO + 对偶求解，这是最经典且可靠的组合。
- **中大规模数据集（10,000-100,000样本）：** 继续使用SMO作为核心求解器，结合交叉验证 + 随机搜索（代替网格搜索以减少计算量）。
- **超大规模数据集（>100,000样本）：** 考虑梯度下降（尤其是SGD）或数据采样，结合C优化和降维技术。
- **非线性问题:** 必须使用对偶求解 + 核函数，SMO是最佳实现方式。

一、支持向量机分类回归与实践

2. 实战代码







03

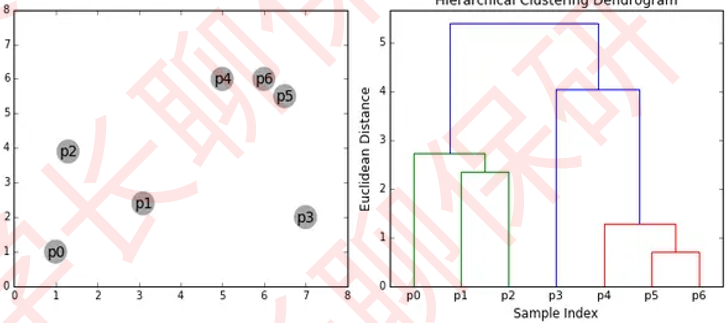
层次聚类算法与实践

三、层次聚类算法与实践

1. 理论介绍-实际案例

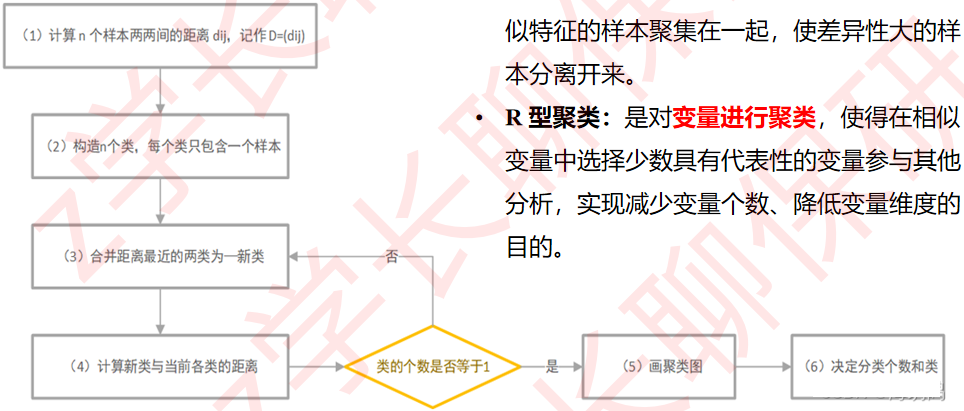
首先**将一定数量的样本或指标各自看成一类**，然后根据样本的亲疏程度，将亲疏程度**最高的两类进行合并**，然后考虑合并后的类与其他类之间的亲疏程度，再进行合并。重复这一过程，直到将所有的样本(或指标)合并为一类。

核心：合并相似项



三、层次聚类算法与实践

1. 理论介绍-实际案例



- **Q 型聚类**：是对**样本进行聚类**，它使具有相似特征的样本聚集在一起，使差异性大的样本分离开来。
- **R 型聚类**：是对**变量进行聚类**，使得在相似变量中选择少数具有代表性的变量参与其他分析，实现减少变量个数、降低变量维度的目的。

三、层次聚类算法与实践

1. 理论介绍-实际案例

距离度量的选择：

距离度量是**层次聚类的关键**，决定了数据点之间的相似性。常用的距离度量包括：

- **欧几里得距离：**适用于连续数值型数据，计算公式为：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

- **曼哈顿距离：**适用于离散数据，计算公式为：

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

- **余弦相似度：**适用于文本数据，计算两个向量之间的夹角，公式为：

$$\text{cosine}(x,y)=\frac{x\cdot y}{\|x\|\|y\|}$$

三、层次聚类算法与实践

1. 理论介绍-实际案例

聚类合并策略：

在自底向上的层次聚类中，合并策略决定了如何计算聚类之间的距离。常见的合并策略包括：

- **单链接**：聚类之间的距离为**两个聚类中最近的两个点之间的距离**。此策略可能导致“链式效应”，使得聚类结果呈现出长条形状。
- **全链接**：聚类之间的距离为两个聚类中**最远的两个点之间的距离**。此策略倾向于形成紧凑的聚类。
- **平均链接**：聚类之间的距离为**所有点对的平均距离**，综合考虑了聚类内部的所有点。
- **Ward法**：通过**最小化聚类内的方差**来选择合并的聚类，通常能够产生更均匀的聚类结果。

三、层次聚类算法与实践

2. 代码实战

```
数据无缺失值，跳过清洗步骤
链接方法：single，调整后的Rand指数：0.5584
链接方法：complete，调整后的Rand指数：0.5726
链接方法：average，调整后的Rand指数：0.5621
链接方法：ward，调整后的Rand指数：0.6153

最佳链接方法：ward，最佳调整Rand指数：0.6153
轮廓系数（Silhouette Score）：0.4467
调整后的Rand指数（Adjusted Rand Score）：0.6153
```