

实验三：支持向量机及其应用

一、 案例背景

支持向量机是一种知名的二元分类器，是一种基于统计学理论的算法。它作为一种监督式学习的算法，可以将数据从一个超平面将其分为两个类别以实现预测或者分类。在现实生活中，每个人的工资水平代表着每个人的购买力，为了探究工资水平和性别对商品购买欲望的影响，了解工资水平与职业的关系，需要我们结合数据进行一定的数据分析。因此，依靠工资水平和性别与购买欲望的数据信息，我们可以对用户进行相应的分类，挖掘购买欲背后的逻辑。

在双十一即将到来之际，我们不妨思考一下自己的生活水平对网购的影响，对于某件商品，我们收入达到多少的时候，购买起来才不会心疼。这样我们将数据挖掘运用于实际生活中，才能让学习变得更有乐趣。为此，我们需要通过现有数据建立职业工资影响购买欲的数学模型，实现以下目标：

- 1) 借助收入与购买欲数据进行分类，并做数据的可视化，绘制分界线。
- 2) 尝试采用非线性的支持向量机完成上述操作。
- 3) 对职业的收入水平进行回归，并预测 level 为 4.5 和 8.5 的员工薪水。

二、 实验目的

本上机实验有以下三个目的：

- 1) 了解模型优化的相应策略。
- 2) 了解核函数。
- 3) 探究线性支持向量机和非线性支持向量机的区别。

三、 实验内容

本上机实验的内容包括以下三个方面：

- 1) 根据用户的收入水平和性别作出散点图判断其是否为线性可分。
- 2) 通过可视化作图，探究线性与非线性之间的区别，以及分析哪个模型更好。
- 3) 通过回归将收入与职业进行一定的连续，再讨论背后的逻辑。

四、 实验方法与步骤

本上机实验的具体方法与步骤如下：

1. 实验一

使用支持向量机，根据收入及年龄分类，判断用户是否具有购买欲望

- a) 将 `Social_Network_Ads` 数据导入编程语言
- b) 根据性别和收入变量，对有购买欲的和无购买欲的用户使用不同的颜色画出对应的散点图，观察分类情况。
- c) 使用线性支持向量机进行分类并尝试评价模型准确率
- d) 尝试使用非线性支持向量机完成上述实验

2. 实验二

- a) 将 `Position_Salaries` 数据导入编程语言
- b) 根据线性回归与职业等级作线性回归，并求得线性回归的 `R-Squared`
- c) 预测 `level` 为 4.5 和 8.5 的员工薪水

3. 实验三

- a) 将实验二的线性回归方程利用在 `Social_Network_Ads` 数据数据中，得到每个用户的职业等级
- b) 尝试使用支持向量机对三维的数据进行分类。

五、思考与实验总结

通过本上机实验，我们可以对以下问题进行思考与总结：

- 1) 什么情况下可用线性支持向量机，什么情况下需要用到非线性的支持向量机，以及两者有什么区别？
- 2) 针对支持向量机而言，对于预测结果的评价标准有哪些？