兰州大学 2022 至 2023 学年第二学期
统计学习与数据挖掘

作业一

请根据老师提供的模板完成下列任务.

1. 对于感知机学习算法,

   (a) 如图所示的训练数据集, 其正实例点是 $x_1 = (3,3)^T$, $x_2 = (4,3)^T$, 负实例点是 $x_3 = (1,1)^T$, 感知机模型 $f(x) = sign(\omega \cdot x + b)$. 这里, $\omega = (\omega^{(1)}, \omega^{(2)})^T$, $x = (x^{(1)}, x^{(2)})^T$.

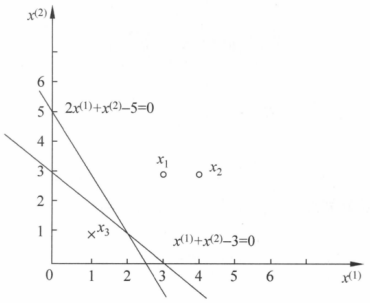

图 2.2　感知机示例

   试用 python 编程实现感知机原始形式和对偶形式两种算法.

   (b) 在手写数字识别中, 利用感知机学习算法的对偶形式来实现代码.

2. 在 k-近邻算法中: 对于新样本, 计算该样本与训练集中所有样本之间的距离, 选取训练集中距离新样本最近的 k 个样本中大多数样本的类别作为新的样本的类别.

   也就是说, 每次都要计算新的样本与训练集中全部样本的距离. 但是, 在实际应用中, 训练集的样本量和特征维度都是比较庞大的, 这就导致该算法不得不在计算距离上花费大量的时间, 那有没有什么方法可以在时间开销上对之前的 k-近邻算法进行优化呢?

采用以空间来换时间的思想, 就引出了今天的主角: kd 树, 试用 python 程序实现这个 kd 树的构造和搜索算法, 并对手写数字识别测试集的样例完全分类且计算准确率.

阅读报告模板请从下列网址下载:

实验报告模板

注意: 同学们请将实验报告上传到超星学习平台.