哈尔滨工业大学(深圳) 2023 学年秋季学期

大数据导论课堂小测

注：

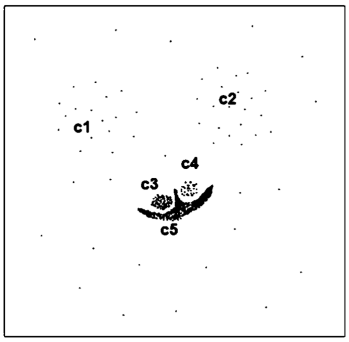
1. 可使用纸质版资料或离线的电子版资料，不可连接网络。
2. 老师会给每人发放白纸，题目在教室前面的投影上，同学在白纸上作答。
3. 本 PDF 为本人课后整理的，同学们在小测时看到的是老师放的 PPT 中的题目而不是本 PDF。

（一）请回答以下两个问题：（4 分）

（1）用于分类的硬间隔支持向量机中，需要满足什么条件的样本才能成为支持向量？支持向量一般最少有几个？

（2）梯度下降法解线性回归问题时，为什么学习率（η）一般不宜设太大？

（二）对于如右图所示的数据集（包括 C1、C2、C3、C4、C5 共五个簇），DBSCAN 算法是否能准确发现所有簇？为什么？请说明原理（3 分）



（三）在基于重构的离群点检测任务中，已知两个数据点$x_1=(2.3,0.5,0.2)^T$，$x_2=(0.8,0.5,0.9)^T$，它们经过重构后分别变换为：$\widehat{x_1}=(2.28,0.49,0.19)^T$，$\widehat{x_2}=(0.4,-0.21,0.52)^T$，请问这两个数据点哪一个更可能是离群点？为什么？请说明原理（3 分）