《数据科学基础》大作业要求

2019.4

1. 概述

分组进行，2 人一组，选定一个题目，题目分理论部分和应用与实验部分。对于理论部分需要调研文献，清楚阐述所包含模型或算法的原理；对于应用与实验部分，自行确定数据集，利用题目理论部分的模型/算法做实验，整理与说明实验结果。

2. 题目

参见表 1. 作业题目表。

3. 考核内容与形式

（1）课堂交流（Presentation）

以讲述 PPT 形式，对题目包含的理论与实验部分进行介绍，时间 10-15 分钟。

（2）大作业报告

包括理论部分与实验部分：理论部分包括题目中列出的模型/算法；实验包括：数据集介绍、评价指标、实验设置（实验内容）、实验结果及讨论等。

4. 时间安排与提交材料

11 周周五前上报题目；

16 周随堂进行大作业课堂交流；

18 周周五之前提交大作业材料，包括 PPT 与大作业报告（考核内容中的两项）。

表 1. 作业题目表

题目序号	主题	题目内容	理论、应用与实验	
1	模型参数估计方法	基于期望最大化 EM 算法估计混合高斯模型 GMM 参数	理论	极大似然估计法、EM 算法、GMM
			应用与实验	GMM 图像或文本聚类
2	降维方法	随机投影、基于 SVD 的数据降维方法 PCA 及其应用	理论	随机投影、SVD、PCA
			应用与实验	图像检索或分类
3	马尔可夫随机过程	基于随机游走的图像分割	理论	马尔科夫链/随机游走、图像分割
			应用与实验	基于随机游走的图像分割算法
4	随机模拟	MCMC-Gibbs 采样算法及其在文本主题模型 LDA 求解中的应用	理论	MCMC-Gibbs 采样算法、LDA
			应用与实验	文本分类或文本相似性计算
5	优化算法	随机梯度下降方法在深度学习中的应用	理论	随机梯度下降法 SGD、BP 算法、CNN
			应用与实验	图像分类
6	非监督机器学习-聚类	几种聚类算法及其应用	理论	聚类基本原理（包括算法性能度量指标、距离计算等）、两种聚类算法（例如基于中心的聚类、层次聚类，基于密度的聚类，谱聚类等）
			应用与实验	图像或文本聚类