《数据科学基础》大作业要求

2020.3

1. 基本要求

（1）分组进行，1-3 人一组，从题目列表中选定一个主题和题目。

（2）题目分理论部分和应用与实验部分：

- 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，以及发展或应用现状；
- 应用与实验部分：自行确定应用对象与数据集，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。从表 1 的主题中选定一个主题和题目，应用和实验部分可自行确定，表的相应部分可做参考。注意，表 1 中带有“*”标记的主题，计算机学院同学不要选择，外专业同学可以选择。

3. 考核内容与形式

（1）中期进展

提交进展说明，包括已完成的工作，小组同学分工，以及后续计划。

占总成绩 20%。

（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。

占总成绩 40%。

（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；应用与实验包括：应用问题说明，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。

说明小组同学分工。

占总成绩 30%。

4. 时间安排与提交材料

第 6 周周日（4 月 5 日）前上报题目信息，包括主题、题目、理论部分内容、应用与实验设想，小组成员信息等；

第 10 周周日（5 月 3 日）前提交进展说明，包括已完成的工作，小组同学分工，以及后续计划；

第 15-16 周进行大作业交流；

第 18 周周日（6 月 28）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	参考题目		
		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	基于 GMM 的图像或文本聚类，或其他
2	* 数据压缩方法	基于 SVD 数据压缩方法	理论	SVD
			应用与实验	图像或其他类型数据的压缩
3	降维方法	随机投影或基于 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，或其他
4	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
			应用与实验	基于随机游走的图像分割算法，或其他
5	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，或其他
6	* 常用机器学习算法	SVM 或集成学习模型	理论	SVM 或 AdaBoost，或其他
			应用与实验	数据分类或回归
7	核方法及其的应用	核 SVM	理论	核方法，SVM
			应用与实验	数据分类或其他
8	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，或其他

9	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，或其他

注：带有“*”标记的主题，计算机学院同学不要选择，外专业同学可以选择