中国科学院大学

期末大作业

所属学期：2020 学年度春季
课程编号：251M5019H
课程名称：机器学习导论
任课教师：向世明、孟高峰

姓名_____ 学号_____ 成绩_____

本次期末大作业由两部分构成。第一部分涉及对本课程基本知识的组织；第二部分是关于基本算法的实践。

- 根据第一部分和第二部分中所述具体任务和相应要求，需完成一份大作业报告。该报告需要以 PDF 文件形式呈现，并在国科大课程网站上进行提交。
- 每位同学需独立完成，不得抄袭。

第一部分：本课程基本知识组织

任务：根据本学期所学《机器学习导论》的内容，制作一棵知识树，它能够反映“机器学习”的重要任务和重要概念。具体要求如下：

具体要求 1：知识树的层次要清晰 （提示：可在 PPT 或绘图软件里制作，然后粘贴至 WORD 文件）。

具体要求 2：应能体现本学期课程内容。

第二部分：编程实践

编程实践所涉及的数据集如表 1 所示。

表 1： 编程实践所需的数据集

数据集	类别数	特征维数	训练样本数目	测试样本数目
MNIST	10	784	60,000	10,000
Letter	26	16	16000	4000

上述两个数据集的获取方式如下（请自行下载）：

- UCI Machine Learning Repository http://archive.ics.uci.edu/ml/ （步骤： 按前述网址打开页面，点击

页面右上角“View ALL Data Sets”，进入下一页，找到“Letter Recognition”，点击进入下载页面。下载完之后，请自行按各类样本数目的比例将 20000 个样本随机划分为 16000 个训练样本和 4000 测试样本。这里只要求大家做一次数据划分）；

- MNIST dataset http://yann.lecun.com/exdb/mnist/ （该数据集的训练数据和测试数据已事先分好，需要在该页面中分别进行下载。下载时需要同时下载样本特征文件和样本类别标签文件）。

请完成以下两个任务：

任务 1： 编程实现一个分类器算法，该分类器由主成分分析（Principal Component Analysis，PCA）、线性判别分析（Linear Discriminant Analysis，LDA）和 K 近邻（k-Nearest Neighbor，KNN）分类器共同完成，即 PCA+LDA+K-NN。具体过程如下：首先，对数据的原始特征采用 PCA 进行降维；然后，以 PCA 的降维结果作为输入，采用 LDA 提取鉴别特征；最后，以 LDA 提取的鉴别特征为输入，采用 K 近邻分类器完成最后的分类。

具体要求 1：写出 PCA 的基本原理、核心公式和主要计算过程。

具体要求 2：写出 LDA 的基本原理、核心公式和主要计算过程。

具体要求 3：在实验过程中，在采用 PCA 对原始数据进行降维时，需按表 2 所示将数据降低至不同的维度（维数）。进一步，针对 PCA 降维所获得的不同维度的数据，分别执行 “LDA+K-NN”，并报告所获得的识别精度。另外，在实验过程中，K-NN 分类器需要采用最近邻（1-NN）分类器和 3 近邻（3-NN）分类器来分别进行实验。最后，写出上述实验过程的主要步骤。

具体要求 4：对实验结果进行分析，并以附录 A 的形式提交源代码（编程语言：C、MATLAB 或 PYTHON）。

表 2： PCA 降维数

数据集	PCA 降维数
MNIST	50 维、100 维、200 维、300 维、400 维
Letter	3 维、5 维、7 维、9 维、11 维

任务 2： 请采用前向神经网络方法编程实现对上述两个数据集的分类。

具体要求 1：简述网络结构设计和网络训练步骤，给出误差反向传播算法的细节。

具体要求 2：报告在不同学习率、不同隐含层结点个数等情形下的分类精度。

具体要求 3：对实验结果进行分析，并以附录 B 的形式提交源代码（编程语言：C、MATLAB 或 PYTHON）。