**2020年机器学习大作业题目**

不限于以下题目，可以自行拟定题目，例如参考国外机器学习案例思路（参考思路，不可照搬数据和源代码），一定要有自己的工作量。最终每个题目都要提交工程报告、训练数据、工程源代码（github）、打包程序。以个人或小组为单位，如果以组为单位，每个人都要提交工程报告，围绕自己的工作写工程报告，需要由**组长分配好每人的工作量（按照百分比）**，并在提交时向老师说明，不可均分。最终大作业分数会根据题目本身难度、题目完成程度、工作量、工程报告结构和内容、界面展示度等决定。

（如果以小组为单位的话，最多人数上限为5人。原则上不要跨专业组队。）

# 第一大类：华为实验可选题目

1. **有监督学习**：APP评分预测
2. **有监督学习**：信用违约预测
3. **无监督学习**：电信用户分析
4. **无监督学习**：购物篮分析
5. **无监督学习**：消费者聚类
6. **推荐系统**：电影推荐实验
7. **集成学习**:贷款产品预测
8. **集成学习**:房价预测

第一大类题目是来自华为实验，共计8个可选实验。这一大类的工作包含两部分：一是**复现**：根据华为实验指导书的复现。二是**改进**：更换数据集和添加更新算法的改进工作。第二部分是主要展现和重点考察的部分。

提交：**总的工程报告**（包括华为实验复现和改进，复现放在附件）、**华为实验部分的**训练数据+工程源代码+打包程序、**改进算法的**训练数据+工程源代码+打包程序。

# 第二大类：本地可选题目

**1.收集一定数量的学生成绩数据（例如10个男生，10个女生，10门课），根据成绩数据，预测该学生是男生还是女生。（分类问题，可以使用libsvm或者神经网络软件包等）**

提交：工程报告+训练数据+工程源代码+打包程序

**2.在libsvm的数据集合中，选取一个多分类数据集合，利用不同的libsvm核函数、和参数进行训练libsvm，并比较实验结果。至少要验证3个以上的核函数，并尝试5组以上的参数（实验）。https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/。（libsvm工具包的使用）**

提交：工程报告+训练数据+工程源代码+打包程序

**3.基于scikit-learn软件包，对mnist数据集合进行分类，并可视化。至少要尝试5中不同的分类方法，并比较分类结果。**

提交：工程报告+训练数据+工程源代码+打包程序

**4. 从http://archive.ics.uci.edu/ml/上下载任意一个数据集，然后根据其问题类型，用相应的软件包，构建相应的分类器或者回归模型。**

提交：工程报告+训练数据+工程源代码+打包程序

**5.收集合肥地区过去一段时间（例如过去一年每个月的平均值）的空气质量（例如pm2.5值），然后构建回归模型，能够预测今年某个月的空气质量值。**

提交：工程报告+训练数据+工程源代码+打包程序

**6.收集一定数量的男和女的照片，训练一个分类器，能够直接根据照片分类出男女。**

提交：工程报告+训练数据+工程源代码+打包程序

**7.收集一定数量的男生的手写字（图片）和一定数量的女生手写字（图片），然后训练一个分类器，能够直接根据这个手写字的照片分类出男女。**

提交：工程报告+训练数据+工程源代码+打包程序

**8.基于深度学习的面部表情分类**

自行查找关于深度学习的资料，基于tensorflow等深度学习系统，构建一个面部表情分类，可以检测静态图片中的人脸面部表情，或者面部表情实时监测系统。

提交：工程报告+训练数据+工程源代码+打包程序

**9.新闻文本分类**

从新闻网站上，收集各类新闻文本（5-10种），根据课堂上讲解的one-hot特征提取方式，或者自行查找文本特征提取方式，对文本进行提取特征并向量化，利用机器学习模型进行分类，构建一个新闻文本分类系统。

提交：工程报告+训练数据+工程源代码+打包程序

10**.垃圾邮件分类系统**

从邮件系统中，收集垃圾邮件和非垃圾邮件每类50-100组，根据课堂上讲解的垃圾邮件检测技术，进行垃圾邮件分类，构建垃圾邮件分类器。

提交：工程报告+训练数据+工程源代码+打包程序

**11.饭卡消费预测系统**

收集自己或同学的饭卡消费数据，从校园网账户中，对每月（每天）的消费数据进行建模，预测下月（或者下一天）的饭卡消费。考虑饭卡消费和月份之间的关联。

提交：工程报告+训练数据+工程源代码+打包程序

**12.水果检测**

设计一个水果检测系统，可以对3-4种水果进行分类，收集3-4种水果的照片，每种水果收集20-50张照片，用机器学习模型进行学习训练，构建桌面应用或者移动端应用，推荐开发移动端应用，实现水果的实时检测与分类。

提交：工程报告+训练数据+工程源代码+打包程序

**13.国籍检测**

设计一个小型的国籍分类系统，能够分类出中国人和日本人。通过摄像头，或者搜索引擎，搜索50-100张日本人和中国人头像照片，训练神经网络（可以直接用课程中讲的神经网络模型），或者通过SIFT特征提取+训练SVM，或者其他任何机器学习模型，完成桌面应用或者移动应用开发。课时实现对一张静态照片中的人脸进行国际检测，或者进行实时的国籍检测。

提交：工程报告+训练数据+工程源代码+打包程序

**14. 基于机器学习的图片自动拼接**

自行查找相关模型算法，构建一个能够将两个图片进行自动拼接为一个完整图片的系统，类似于全景照片的拼接过程。

提交：工程报告+训练数据+工程源代码+打包程序

15.自行设计的解决实际问题的题目，或者参加国际国内比赛的机器学习相关项目。

提交：工程报告+训练数据+工程源代码+打包程序



**××××工程报告**



**课 程：** 机器学习基础

**姓 名：**

**学 号：**

**完成时间：**

一． 工程摘要与每人贡献

摘要：总体介绍本工程的工作。

每人分工：

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 角色  （组长，组员，独自完成） | 工作量  比例 | 负责内容 |
| 张三 | 组长 | 60% | 收集\*数据集；\*算法代码实现；报告的第三/第五部分撰写..... |
|  |  |  |  |
|  |  |  |  |

二、研究背景与意义

说明本工程所涉及的研究背景研究的意义，包括主要涉及领域，主要研究方法，主要存在问题，现有解决方案等等，可以自行补充其他内容。

三、模型方法

**详细说明**本工程所使用的模型方法和理论等

四、系统设计

系统的**详细**设计，系统流程，系统的每一步的具体流程，例如，如何处理语料，如何训练模型，如何测试模型，最后对模型进行评估等。

五．实验结果分析、对比和讨论

对实验结果进行分析说明，测试一些课程中演示的样例，根据结果说明为什么对或者为什么错等具体分析；对不同模型和参数的对比进行分析；最后对系统提出改进方案等。

六．对本门课的感想、意见和建议

希望听到你的心声。

1. 华为实验指导书相关实验的复现报告书

（如果选择第一大类的题目）