**机器学习实战**

笔记整理人：天国之影（2019年1月29日）

# ****说明****

1. 每周三、周六为休息日，当天无须打卡，不会安排任何作业和任务。若学习时长中包含周三或周六，则默认忽略当天计划。
2. 本课程提供的所有资料将汇总在 GitHub 上，包括作业。参考答案会在第二天由助教同步到 GitHub 上，并从学员提交的答案中选出最佳答案同步在 GitHub 上。
3. 本课程作业的所有代码都要基于Python3，在Jupyter Notebook上完成。
4. 知识星球具有代表性的问题由导师红色石头或助教同步到 GitHub 上，旨在给所有学员建立一个完备的机器学习实战资料库。

**原始作业GitHub地址：**

<https://github.com/RedstoneWill/MachineLearningInAction-Camp>

**我的作业GitHub地址（在每一个Week中均有一个MyHomeWork文件夹，用于记录我的作业完成情况，所有ipynb文件均带注释）：**

https://github.com/Relph1119/MachineLearningInAction-Camp

# 第1周学习计划

## 第一节学习内容

**学习时长：**12/2

**任务1题目：**观看机器学习实战绪论视频**+**天池o2o比赛完全流程解析PPT

**任务详解：**第一次视频课主要以《机器学习实战》第一章为基础，主要介绍机器学习的基本概念、算法类型、推荐学习路线和一些预备知识，包括Numpy、Pandas、Matplotlib 等Python 基本库。还有天池o2o比赛完全流程解析。

**作业：**每个学员注册天池账号，报名参加比赛。提交结果，查看成绩。（结果 submit1.csv文件提供，学员只需按照直播视频讲述的方法提交查看成绩就好。submit1.csv 文件已放置在GitHub上）

**作业提交形式：**比赛上传结果界面排名截图打卡上传

## ****第二节学习内容****

**学习时长：**12/3

**任务1题目：**配置开发环境，熟悉 Jupyter Notebook

**任务详解：**以Python3为开发语言，安装软件Anaconda。Anaconda自带Jupyter Notebook，熟悉Jupyter Notebook的基本用法。

**参考资料：**

**[Jupyter Notebook入门教程（上）]**

**(**<https://mp.weixin.qq.com/s/O2nTGOtqGR-V33-YJgPgJQ>**)**

**[Jupyter Notebook入门教程（下）]**

**(**<https://mp.weixin.qq.com/s/AwSzkjlpwvdUzh6CmHq6AQ>**)**

**作业：使用Jupyter Nootbook，对Numpy、Pandas、Matplotlib各写一个小的demo程序。要求是解释性说明和代码相结合的形式。**

**作业提交形式：代码截图打卡提交**

## ****第三节学习内容****

**学习时长：**12/4—12/7

**任务1题目：**书籍阅读

**任务详解：**阅读《机器学习实战》书籍第二章2.1、2.2、2.3章节

**参考资料：**李航《统计学习方法》第3章

**作业1：**简要概括 k-近邻算法的原理，优缺点。

**提交日期：**12/5

**提交形式：**文字打卡提交或者上交.md文件的链接

**作业2：**将本章中“使用 k­近邻算法改进网站的配对效果”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

**提交日期：**12/7

**提交形式：**代码截图打卡提交或git链接提交

**作业3：**将本章中“手写识别系统”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

**提交日期：**12/7

**提交形式：**代码截图打卡提交或git链接提交

# ****第1周作业参考答案****

（说明：当天作业参考答案隔天发布）

1.1 略

2.1 略

3.1 原理：存在一个样本数据集，也称作训练样本集，并且样本中每个数据都存在标签，即我们知道样本集中每一数据与所属分类的对应关系，输入没有标签的新数据后，将新数据的每个特征与样本集中的数据对应的特征进行比较，然后算法提取样本集中特征最相似的数据（最近邻）的分类标签。一般来说，我们只选择样本集中前k个最相似的数据，这就是k-近邻算法中k的出处，通常k是不大于20的整数，最后，选择k个最相似的数据中出现次数最多的分类，作为新数据的分类。

优点：精度高，对异常数据不敏感（你的类别是由邻居中的大多数决定的，一个异常邻居并不能影响太大），无数据输入假定；算法简单，容易理解，无复杂机器学习算法。

缺点：计算发杂度高（需要计算新的数据点与样本集中每个数据的“距离”，以判断是否是前k个邻居），空间复杂度高（巨大的矩阵）。

3.2 见GitHub链接：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week1/MyHomeWork/homework_3.2.ipynb>

3.3 见GitHub链接：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week1/MyHomeWork/homework_3.3.ipynb>

# 第2周学习计划

一、学习总周期

2018/12/09 – 2018/12/15

二、分节学习内容

## 第一节学习内容

学习时长：12/09-12/10

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第三章3.1、3.3、3.4节（3.2节选做）

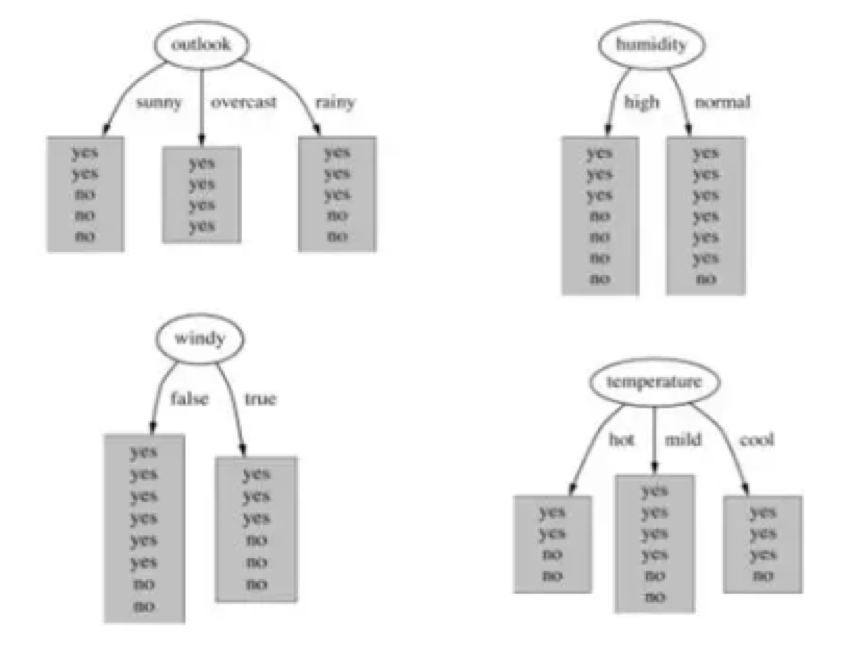
参考文献：李航《统计学习方法》第5章中的5.1-5.3节

作业1：概括决策树分类算法的原理。

提交日期：12/09

提交形式：文字打卡提交或者上交.md文件的链接

作业2：在构建一个决策树模型时，我们对某个属性分割节点，下面四张图中，哪个属性对应的信息增益最大？



提交日期：12/09

提交形式：文字或者截图打卡提交

作业3：将本章中“使用决策树预测隐形眼镜类型”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：12/10

提交形式：代码截图打卡或git链接提交

## 第二节学习内容

学习时长：12/11-12/14

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第四章

参考文献：李航《统计学习方法》第4章

参考资料：[通俗易懂！白话朴素贝叶斯]

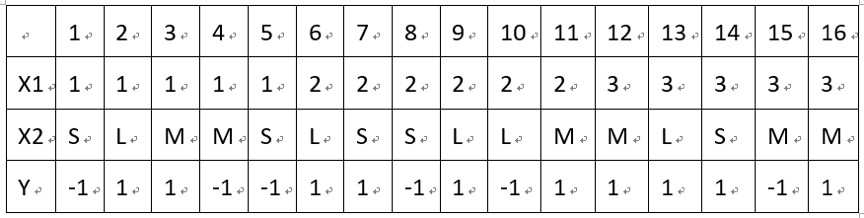
(<https://mp.weixin.qq.com/s/7xRyZJpXmeB77MZNLqVf3w>)

作业1：概括朴素贝叶斯分类算法的原理，为什么称之为“朴素”？

提交日期：12/11

提交形式：文字打卡提交或者上交.md文件的链接

作业2：



试由下表的训练数据学习一个朴素贝叶斯分类器并确定x=（2,S）的类标记y。表中X1和X2为特征。

提交日期：12/11

提交形式：文字或者截图打卡提交

作业3：将本章中“使用朴素贝叶斯过滤垃圾邮件”完整代码键入Jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：12/12

作业提交形式：代码截图打卡或git链接提交

作业4：将本章中“使用朴素贝叶斯分类器从个人广告中获取区域倾向”完整代码键入Jupyter Notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：12/14

提交形式：代码截图打卡或git链接提交

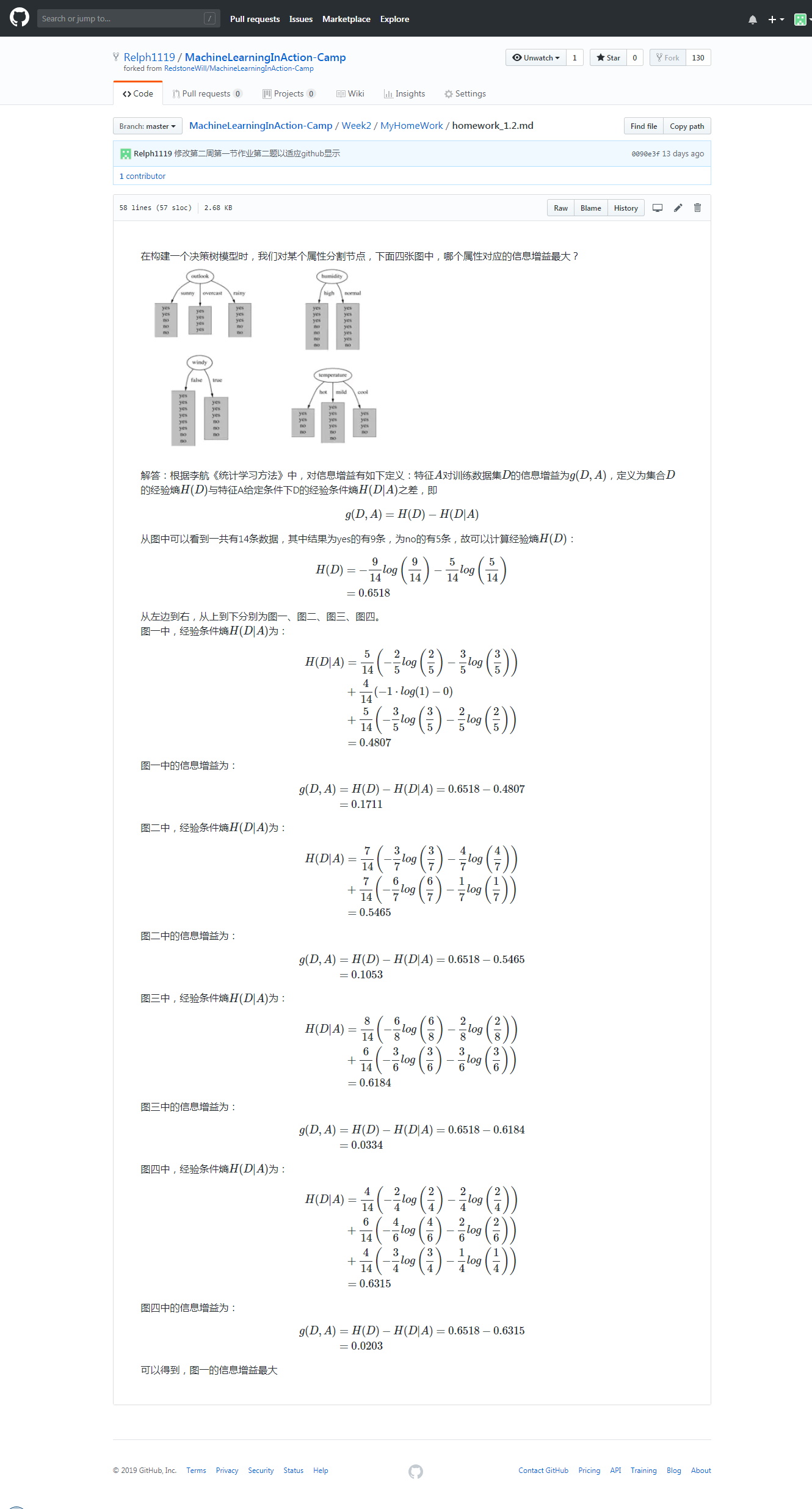
# 第2周作业参考答案

1.1见GitHub链接：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week2/MyHomeWork/homework_1.1.md>

1.2见GitHub链接：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week2/MyHomeWork/homework_1.2.md>



1.3 Jupyter Notebook 见GitHub：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week2/MyHomeWork/homework_1.3.ipynb>

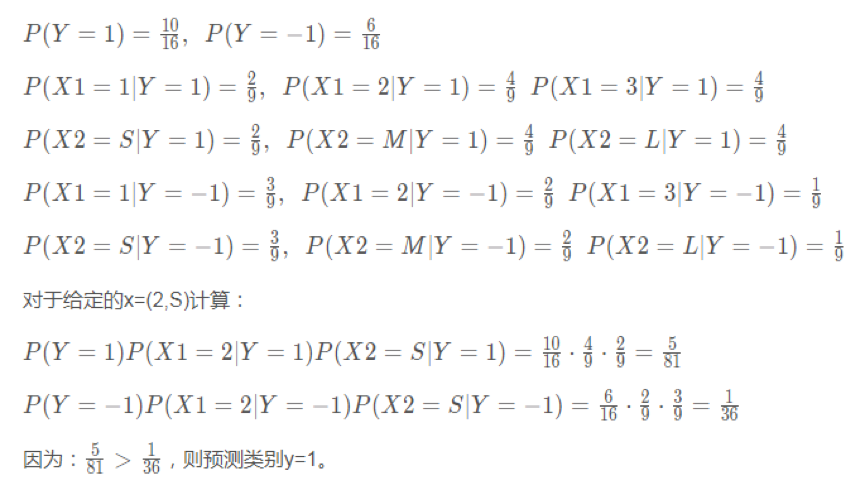
2.1 见GitHub链接：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week3/MyHomeWork/homework_2.1.md>

朴素贝叶斯法对条件概率分布作了条件独立性的假设。由于这是一个较强的假设，朴素贝叶斯法也因此得名。

2.2 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week3/MyHomeWork/homework_2.2.ipynb>



2.3 Jupyter Notebook 见GitHub：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week3/MyHomeWork/homework_2.3.ipynb>

2.4 Jupyter Notebook 见GitHub：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week3/MyHomeWork/homework_2.4.ipynb>

# ****第3周学习计划****

一、学习总周期

2018/12/16– 2018/12/22

二、分节学习内容

## 第一节学习内容

学习时长：12/16-12/17

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第5章

参考文献：李航《统计学习方法》第6章中的6.1节

作业1：写出并解释逻辑回归的损失函数，推导参数w的梯度下降公式。

提交日期：12/16

提交形式：文字或者截图打卡

作业2：将本章中“从疝气病症预测病马的死亡率”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：12/17

提交形式：代码截图打卡或git链接提交

## ****第二节学习内容****

学习时长：12/18-12/21

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第6章6.1/6.2/6.3节

参考资料：

李航《统计学习方法》第7章

[深入浅出机器学习技法（一）：线性支持向量机（LSVM）]

(<https://mp.weixin.qq.com/s/Ahvp0IAdgK9OVHFXigBk_Q>)

[深入浅出机器学习技法（二）：对偶支持向量机（DSVM）]

(<https://mp.weixin.qq.com/s/Q5bFR3vDDXPhtzXlVAE3Rg>)

作业1：推导SMO算法

提交日期：12/19

提交形式：文字或者截图打卡

作业2：理解书中程序清单6-2的简化SMO算法程序，对程序中详细注释。

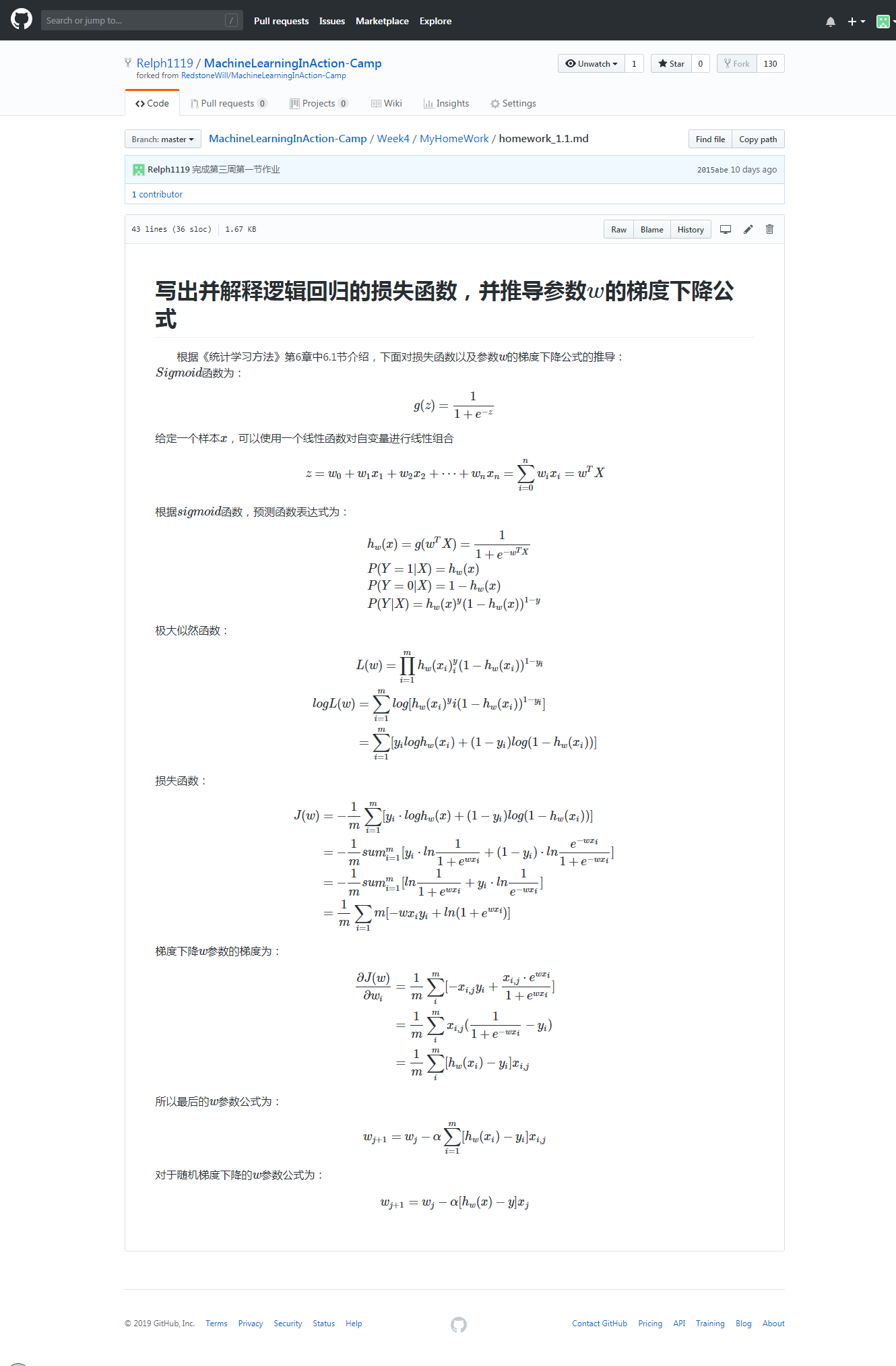
提交日期：12/21

提交形式：文字或者截图打卡

# ****第3周作业参考答案****

1.1见GitHub链接：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week4/MyHomeWork/homework_1.1.md>



2.2 Jupyter Notebook 见GitHub：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week4/MyHomeWork/homework_1.2.ipynb>

2.1 李航《统计学习方法》第7章7.4.1小节

2.2 Jupyter Notebook 见GitHub：

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week6/MyHomeWork/homework_2.2.ipynb>

# ****第4周学习计划****

一、学习总周期

2018/12/23 – 2018/12/29

二、分节学习内容

## 第一节学习内容

学习时长：12/23-12/24

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第6章6.4/6.5/6.6节

参考文献：

李航《统计学习方法》第7章

[深入浅出机器学习技法（一）：线性支持向量机（LSVM）]

(<https://mp.weixin.qq.com/s/Ahvp0IAdgK9OVHFXigBk_Q>)

[深入浅出机器学习技法（二）：对偶支持向量机（DSVM）]

(<https://mp.weixin.qq.com/s/Q5bFR3vDDXPhtzXlVAE3Rg>)

[深入浅出机器学习技法（三）：核支持向量机（KSVM）]

(<https://mp.weixin.qq.com/s/cLovkwwgGJRgSSa1XWZ8eg>)

作业1：为了防止SVM出现过拟合，应该对参数C进行如何设置？

提交日期：12/23

提交形式：文字或者截图打卡

作业2：将本章中“手写识别问题”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：12/24

提交形式：代码截图打卡或git链接提交

## ****第二节学习内容****

学习时长：12/25-12/28

任务1题目：天池o2o预测赛（初级）

任务详解：建立一个简单的线性模型，在线提交预测结果，查看成绩

视频不清晰也可以去荔枝微课看，地址：

https://m.lizhiweike.com/lecture2/10234967（观看密码：011220）

源码文件：链接：

<https://pan.baidu.com/s/1FwCcG0Pk1V_0mK1MCbkPlg>

提取码：y5z6

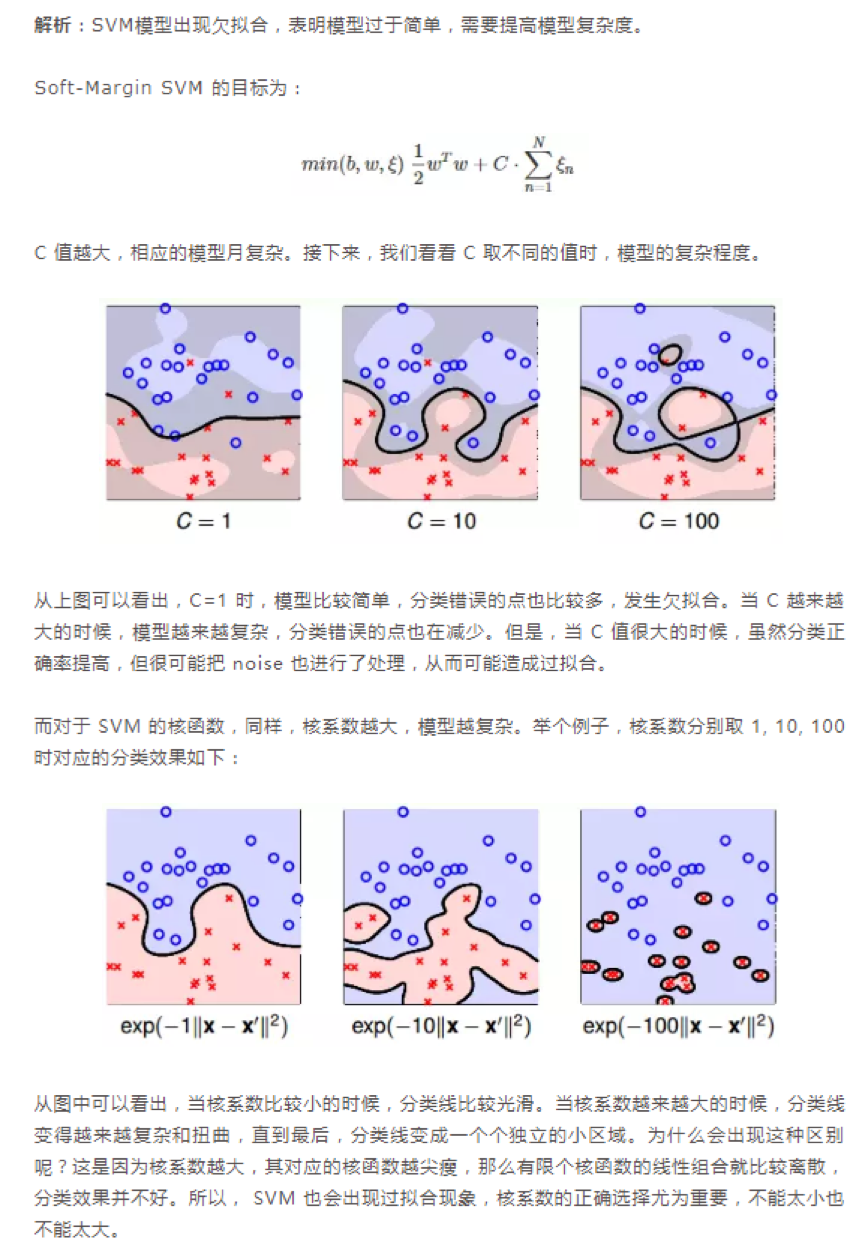
作业1：使用简单模型，在线提交预测结果，查看成绩

提交日期：12/28

提交形式：代码截图打卡或git链接提交，比赛上传结果界面排名截图打卡上传

# 第4周作业参考答案

1.1



1.2 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week6/MyHomeWork/homework_1.2.ipynb>

2.1 Jupyter Notebook 见GitHub，带注释

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week5/MyHomeWork/homework.ipynb>

# 第5周学习计划

一、学习总周期

2018/12/30 – 2018/01/05

二、分节学习内容

## 第一节学习内容

学习时长：12/30-01/02

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第7章

参考文献：

李航《统计学习方法》第8章8.1/8.2/8.3节

作业1：AdaBoost选择分类器是弱分类器还是强分类器？解释原因。

提交日期：12/30

提交形式：文字或者截图打卡

作业2：将本章中“在一个难数据集上应用AdaBoost”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：01/02

提交形式：代码截图打卡或git链接提交

## 第二节学习内容

学习时长：1/03-1/04

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第8章

作业1：岭回归和Lasso回归有什么区别？

提交日期：1/03

提交形式：代码截图打卡或git链接提交

作业2：将本章中“预测鲍鱼的年龄”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/04

提交形式：代码截图打卡或git链接提交

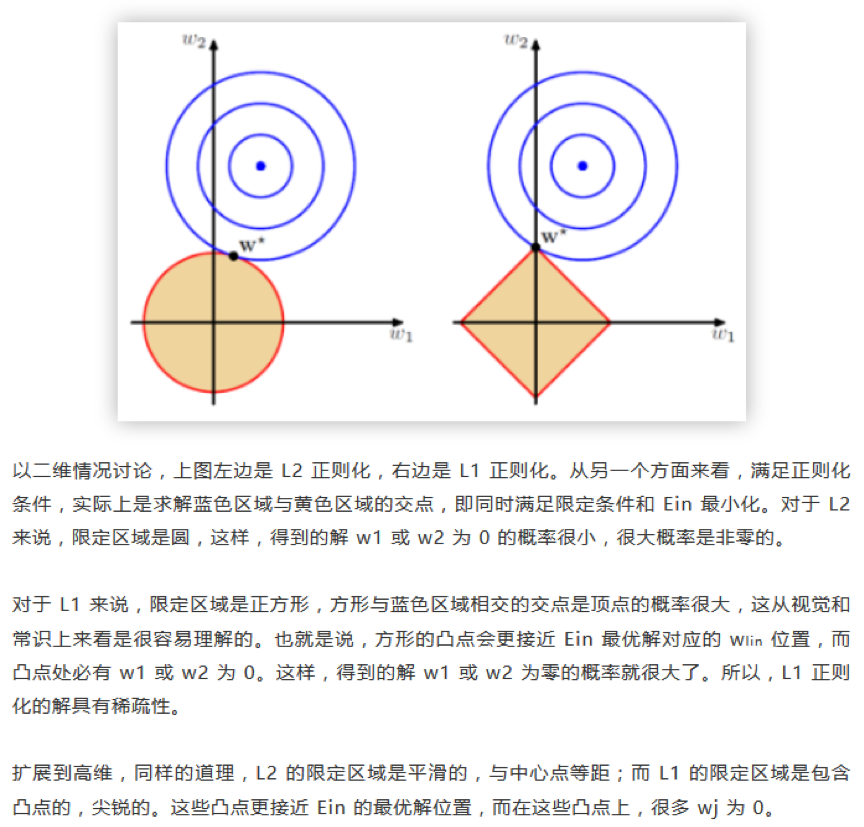
# 第5周作业参考答案

1.1   弱分类器。若是强分类器，那么该分类器占的权重alpha会很大，相当于其它分类器不起作用了。所以，多个弱分类器起到“三个臭皮匠，赛过诸葛亮”的效果。

1.2 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week7/MyHomeWork/homework_1.2.ipynb>

2.1 使用的正则化不同，岭回归使用L2正则化，Lasso使用L1正则化。L2正则化优点是易于求导，简化计算，更加常用一些。L1正则化优点是能得到较稀疏的解，但缺点是不易求导。



2.2 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week8/MyHomeWork/homework_2.2.ipynb>

# ****第6周学习计划****

一、学习总周期

2018/1/06– 2018/1/12

二、分节学习内容

## 第一节学习内容

学习时长：1/06-1/09

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第9章

作业1：将本章中“树回归与标准回归的比较”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/09

提交形式：代码截图打卡或git链接提交

作业2（选做）：将本章中“使用Python的Tkinter库创建GUI”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/09

提交形式：代码截图打卡或git链接提交

补充作业！！！！！！！

天池O2O优惠券使用预测分析比赛开始啦！

学习时长：1/6-1/11

任务1题目：阿里云天池o2o优惠券使用预测分析比赛（进阶）

任务详解：建立一个简单的线性模型，在线提交预测结果，查看成绩

视频地址：<https://m.lizhiweike.com/lecture2/11570830>（观看密码：031220）

源码文件：

<https://pan.baidu.com/s/11H41u4Y7iBkvl4fgTQCeoA>（提取码：n5mo）

作业名称：使用简单模型，在线提交预测结果，查看成绩

作业提交日期：1/11

任务提交形式：代码截图打卡或git链接提交，比赛上传结果界面排名截图打卡上传

## 第二节学习内容

学习时长：1/10-1/12

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第10章

作业1：将本章10.4.2中“对地理坐标进行聚类”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/11

提交形式：代码截图打卡或git链接提交

# 第6周作业参考答案

1.1 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week9/MyHomeWork/homework_1.1.ipynb>

1.2 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week9/MyHomeWork/homework_1.2.ipynb>

2.1 Jupyter Notebook 见GitHub

https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week11/MyHomeWork/homework\_2.1.ipynb

# 第7周学习计划

一、学习总周期

2018/1/13– 2018/1/19

二、分节学习内容

## 第一节学习内容

学习时长：1/13-1/14

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第11章11.1/11.2/11.3节

作业1：使用Apriori算法进行关联分析的目标主要包含哪两个方面？Apriori的原理是什么？

提交日期：1/14

提交形式：文字或者截图打卡

## 第二节学习内容

学习时长：1/15-1/18

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第11章11.4/11.6节

作业1：将本章11.6中“发现毒蘑菇的相似特征”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/18

提交形式：代码截图打卡或git链接提交

# 第7周作业参考答案

1.1 Apriori算法关联分析的目标主要包括两项：发现频繁项集和发现关联规则。Apriori原理是说如果某个项集是频繁的，那么它的所有子集也是频繁的。反过来说，如果一个项集是非频繁集，那么它的所有超集也是非频繁的。

2.1 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week12/MyHomeWork/homework_2.1.ipynb>

# 第8周学习计划

一、学习总周期

2018/1/20– 2018/1/26

二、分节学习内容

## 第一节学习内容

学习时长：1/20-1/21

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第12章12.1/12.2节

作业1：FP-growth算法的基本工作流程是什么？其相比Apriori算法优点是什么？

提交日期：1/21

提交形式：文字或者截图打卡

作业2：理解带头指针表的FP树（图12.2），理解FP树生成代码。

提交日期：1/21

提交形式：文字或者截图打卡

## 第二节学习内容

学习时长：1/22-1/25

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第12章12.3/12.5/12.6节

作业1：将本章12.5中“从新闻网站点击流中挖掘”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/25

提交形式：代码截图打卡或git链接提交

# 第8周作业参考答案

1.1  FP-growth算法的基本工作流程分为两步。一、首先构建FP树。需要对原始数据集扫描两遍，第一遍对所有元素项的出现次数进行计数，第二遍只考虑那些频繁元素。二、挖掘频繁项集。

FP-growth算法只需要对数据库进行两次扫描，而Apriori算法对于每个潜在的频繁项集都会扫描数据集判定给定模式是否频繁，因此FP-growth算法的速度要比Apriori算法更快。

1.2 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week13/MyHomeWork/homework_1.2.ipynb>

2.1 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week13/MyHomeWork/homework_2.1.ipynb>

# 第9周学习计划

一、学习总周期

2018/1/27– 2018/2/1

二、分节学习内容

## 第一节学习内容

学习时长：1/27-1/28

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第13章13.1/13.2/13.3节

作业1：将本章13.3中“利用PCA对半导体制造数据降维”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/28

提交形式：代码截图打卡或git链接提交

## 第二节学习内容

学习时长：1/29-2/1

任务1题目：书籍阅读

任务详解：阅读《机器学习实战》书籍第14章14.1-14.6节

作业1：将本章14.5中“餐馆菜肴推荐引擎”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：1/30

提交形式：代码截图打卡或git链接提交

作业2：将本章14.6中“基于SVD的图像压缩”完整代码键入jupyter notebook，并添加详细注释。若有可能，自己可以优化该代码。

提交日期：2/1

提交形式：代码截图打卡或git链接提交

# 第9周作业参考答案

1.1 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch13/homework_1.1.ipynb>

2.1 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch14/homework_2.1.ipynb>

2.2 Jupyter Notebook 见GitHub

<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch14/homework_2.2.ipynb>