**机器学习实战（第四期）**

笔记整理人：天国之影（2019年6月15日）

# ****说明****

1. 本课程作业的所有代码都要基于Python3，在Jupyter Notebook上完成。
2. 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

## 绪论与准备

### 第四期开营仪式

**任务标题：**开营仪式

**任务简介：**参加今晚20:30的开营仪式，并根据讲解内容在训练营里完整操作一遍。

**任务详解：**熟悉一下我们的学习平台、每日打卡介绍和方法，导师见面会。

直播间地址：

<https://study.163.com/course/courselive/1279023535.htm?share=2&shareId=400000000445063>

**备注：**

1. 没有时间参加开营直播的同学可以看录播和回放，直播结束后的第二天点上方链接即可看回放。
2. 想在pc端看的，复制链接到网页打开即可。

### 绪论与准备

**任务简介：**

1. 学习绪论视频，了解预备知识，认识群内其他小伙伴。
2. 下载书籍的电子版，提前自己预习观看。
3. 注册天池账号，报名参加“天池新人实战赛o2o优惠券使用预测”比赛。提交给定的结果样例，查看成绩。目的是让大家走一遍比赛流程。

**任务详解：**

1. 观看绪论视频
2. 下载书籍电子版先预习：

英文链接：

<https://pan.baidu.com/s/1jdbnHKAkxqMRlzWoQYU9iw>

提取码：aurp

中文链接：

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

提取码：einz

1. 注册天池账号，报名参加“天池新人实战赛o2o优惠券使用预测”比赛。提交给定的结果样例，查看成绩。目的是让大家走一遍比赛流程。

**特别注意：**

今天的任务，不需要训练，只是让大家熟悉一下比赛流程。下面提供的百度云链接已经给大家结果文件了，大家只要按照指示流程注册账号，报名参赛，提交结果文件即可。无需下载数据集进行训练！

**结果样例：**

链接：

<https://pan.baidu.com/s/1TB1aHajcuJrExZ6ChSx0Rg>

提取码：6yg3

天池成绩每天12点和晚上八点更新，提交结果后请大家耐心等待成绩更新。

## ****学习k-邻近算法****

**学习时长：**4/30

**任务简介：**阅读《机器学习实战》2.1-2.3，学习k-近邻算法

**任务详解：**

今天学习任务比较简单，因此只有书籍阅读任务，无补充图文或视频教程。之后的难点任务和项目实战作业，均会有老师录制成讲解视频。

我们将介绍第一个机器学习算法：k-近邻算法，它非常有效而且易于掌握。主要内容包括k-近邻算法的基本原理，如何使用Python编写一个k-近邻算法，并将它应用在约会网站配对和手写识别系统中。本节的重点是掌握k-近邻的核心：基于距离的测量方式，例如欧式距离。难点是选取的k值不好确定。实际上，可以通过选择不同的k值比较分类效果来确定最佳k值。此外，需要注意的是，因为是基于距离比较，所以样本各特征之间的取值范围差别较大的时候，应该对特征进行归一化处理，提升分类效果。

**参考资料：**李航《统计学习方法》第3章

**打卡：**

（1）内容：简要概括k-近邻算法的原理，优缺点。

（2）形式：文字，至少50字

打卡截至日期：5/1

## ****项目作业打卡日****

**学习时长：**5/2

**任务简介：**k-近邻算法项目打卡日，完成本周项目作业。

**任务详解：**

本节我们将开始第一个Python实战代码项目，是不是很激动人心呢？主要包含两个项目，要求同学们使用Python一步一步搭建k-近邻算法，赶紧开始吧！同学们在编写代码的过程中，也可以尝试使用不同的距离测量函数，可以选择不同的k值，比较分类的准确率。

**Python项目：约会网站配对（《机器学习实战》2.2）**

链接：https://pan.baidu.com/s/1Jj2WwyD25yhgAaVJw5KSgg

提取码：eihp

**Python项目：手写识别系统（《机器学习实战》2.3）**

链接：https://pan.baidu.com/s/1kmiT0IeB71eKfP0xlg1Nlw

提取码：ab0a

**打卡：**

（1）内容：编写项目Python代码，运行正确，提交运行结果截图。

注意：项目的图可以不画！！

（2）形式：图片，至少2张

作业答案和讲解视频将在下周一公布

作业截至提交日期：本周日5/5

## 天池o2o优惠券使用预测比赛

**学习时长：**5/3

**任务简介：**

搭建Python开发环境，学习天池o2o优惠券使用预测比赛初级源代码，运行程序，提交结果，查看成绩。

**任务详解：**搭建Python开发环境

Python开发环境配置教程：

https://shimo.im/docs/W5pX5mENS20DCquh

Jupyter Notebook速成手册

上：https://mp.weixin.qq.com/s/O2nTGOtqGR-V33-YJgPgJQ

下：<https://mp.weixin.qq.com/s/AwSzkjlpwvdUzh6CmHq6AQ>

**打卡：**

（1）内容：运行天池o2o优惠券使用预测比赛初级源代码，上传结果，查看成绩，提交成绩截图。

（2）形式：图片，至少1张

打卡截至提交日期：5/5

**天池o2o优惠券使用预测比赛初级源代码和数据集**

链接：https://pan.baidu.com/s/1JkMCOmcmXIaOUoC9L6c3Vg

提取码：hhen

## ****第1周作业参考答案****

1. **简要概括k-近邻算法的原理，优缺点。**

**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）。

1. **Python项目：约会网站配对（《机器学习实战》2.2）**

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

1. **Python项目：手写识别系统（《机器学习实战》2.3）**

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

# 第2周

## 学习决策树的构造

**学习时长：**5/6

**任务简介：**阅读《机器学习实战》3.1，学习决策树的构造

**任务详解：**

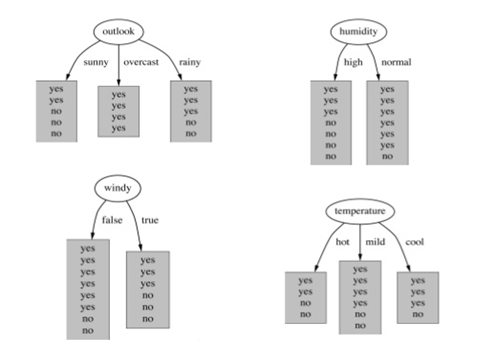
本节将通过算法一步步地构造决策树，并会涉及许多有趣的细节。首先我们讨论数学上如何使用信息论划分数据集，然后编写代码将理论应用到具体的数据集上，最后编写代码构建决策树。重点是掌握熵和信息增益的概念，根据信息增益的大小划分特征。难点是使用递归方式构建决策树，同学们需要重点攻克3.1中的代码。注意，3.2节内容不重要，不需要看。

**参考资料：**李航《统计学习方法》第5章中的5.1-5.3节

**打卡：**

（1）内容：

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？



写下你的计算过程和结果，拍照，上传图片。

（2）形式：图片，至少1张

打卡截止日期：5/7

## 测试和存储决策树

**学习时长：**5/7

**任务简介：**阅读《机器学习实战》3.3-3.4，学习测试和存储决策树。

**任务详解：**

本节将使用决策树构建分类器，并介绍实际应用中如何存储分类器。然后在真实数据上使用决策树分类算法，验证它是否可以正确预测出患者应该使用的隐形眼镜类型。重点是如何使用已经构建好的决策树来进行分类测试，如何利用Python模块pickle来存储决策树模型。

**打卡：**

（1）内容：请用文字描述，决策树模型如何存储。

（2）形式：文字，至少20字。

打卡截至日期：5/8

## 项目作业打卡日

**学习时长：**5/9

**任务简介：**决策树算法项目打卡日，完成本周项目作业。

**任务详解：**

**Python项目：**使用决策树预测隐形眼镜类型（《机器学习实战》3.4）

链接：https://pan.baidu.com/s/1HET4ogSZNnPrnIdcsg1hUw

提取码：ts4y

**打卡：**

（1）内容：编写项目Python代码，运行正确，提交运行结果截图。项目的图可以不画

（2）形式：图片，至少1张

作业答案和讲解视频将在下周一公布

作业截至提交日期：5/12

## 天池o2o比赛-决策树模型

**学习时长：**5/10

**任务简介：**修改“天池o2o优惠券使用预测比赛-初级”的代码，调用scikit-learn库，使用决策树算法来进行预测，运行程序，提交结果，查看成绩。

**任务详解：**

这部分代码给到大家，同学们也可以自行修改和优化。数据集之前给过了，记得把数据集放在代码所在的目录下

天池o2o优惠券使用预测比赛-决策树模型

链接：https://pan.baidu.com/s/1KMWIUCS82W6\_qV0Jld7P9g

提取码：pqmy

**打卡：**

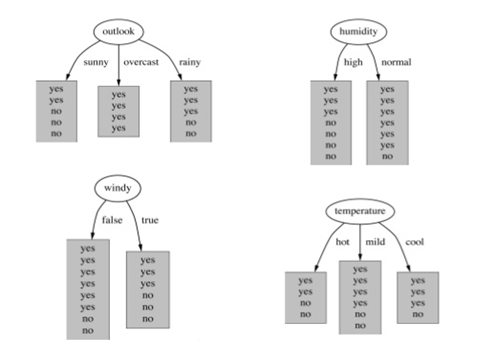
1）内容：运行程序，上传结果，查看成绩，提交成绩截图。

2）形式：图片，至少1张

打卡截至提交日期：5/11

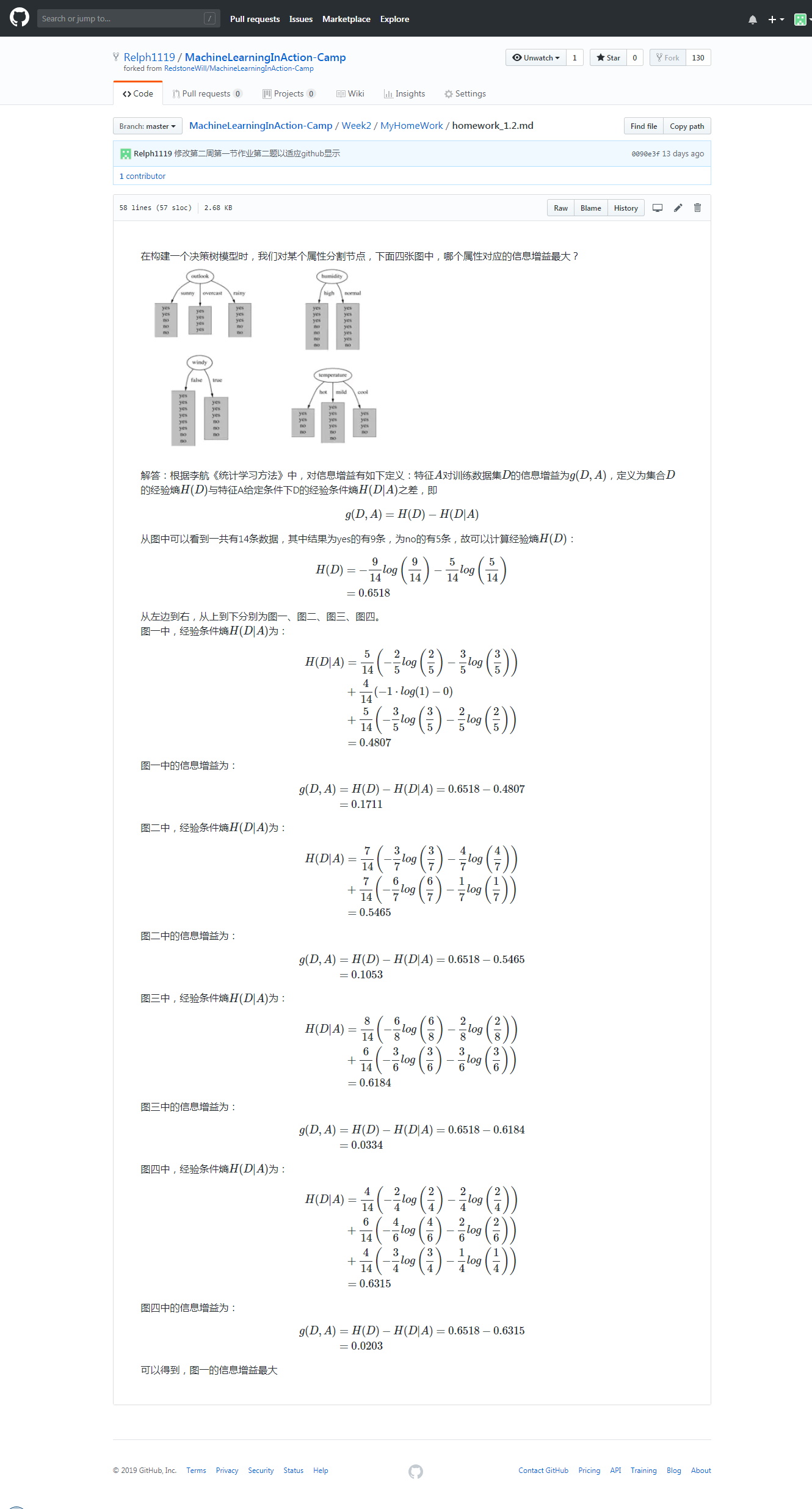
## 第2周作业参考答案

1. **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？**



写下你的计算过程和结果，拍照，上传图片。

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



1. **Python项目：使用决策树预测隐形眼镜类型（《机器学习实战》3.4）**

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

# 第3周

## 朴素贝叶斯

**学习时长：**1天

**任务简介：**学习《机器学习实战》4.1-4.4

**详细说明：**

本节将会给出一些使用概率论进行分类的方法。首先从一个最简单的概率分类器开始，然后给出一些假设来学习朴素贝叶斯分类器。我们之所以称之为“朴素”，是因为整个形式化过程中只做最原始、最简单的假设。重点理解贝叶斯公式和朴素二字的含义，学会如何利用朴素贝叶斯公式解决分类问题。难点是需要知道一些基本的统计学知识，包括条件概率、全概率公式等。

**参考资料：**

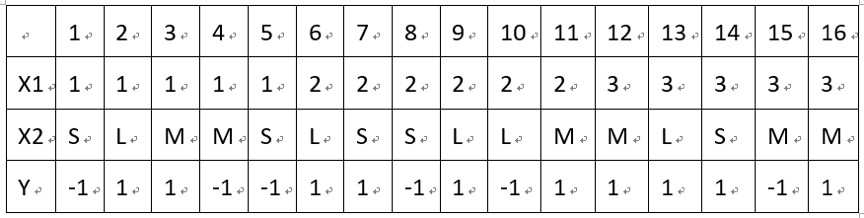
李航《统计学习方法》4.1-4.6

白话朴素贝叶斯：https://mp.weixin.qq.com/s/7xRyZJpXmeB77MZNLqVf3w

**打卡：**

（1）内容：

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。



写下你的计算过程和结果，拍照，上传图片。

（2）形式：图片，至少1张

打卡截至日期：5/14

## 文本分类与垃圾邮件过滤

**学习时长：**1天

**任务简介：**学习文本分类与垃圾邮件过滤，阅读《机器学习实战》4.5-4.6

**任务详解：**

本节将充分利用Python的文本处理能力将文档切分成词向量，然后利用词向量对文档进行分类。我们还将构建另一个分类器，观察其在真实的垃圾邮件数据集中的过滤效果。重点掌握文本的划分，以及朴素贝叶斯算法在训练函数中如何实现的。

**打卡：**

1）内容：什么是词集模型（set-of-words model），什么是词袋模型（bag-of-words model）？二者有何区别？

2）形式：文字，至少60字

打卡截止日期：5/15

## 项目作业打卡日

**学习时长：**1天

**任务简介：**朴素贝叶斯算法项目打卡日，完成本周项目作业。

**任务详解：**

**Python项目：**使用朴素贝叶斯过滤垃圾邮件（《机器学习实战》4.6）

链接：https://pan.baidu.com/s/1JX0Voc3bOgTSoD9PRKeKAQ

提取码：dpd5

**打卡：**

（1）内容：编写项目Python代码，运行正确，提交运行结果截图。

（2）形式：图片，至少1张

本周项目答案和讲解视频下周一公布

打卡截至日期：5/17

## 天池o2o比赛-朴素贝叶斯模型

**学习时长：**1天

**任务简介：**

修改“天池o2o优惠券使用预测比赛-初级”的代码，调用scikit-learn库，使用朴素贝叶斯算法来进行预测，运行程序，提交结果，查看成绩。

**任务详解：**

这部分代码给到大家，同学们也可以自行修改和优化。数据集之前给过了，记得把数据集放在代码所在的目录下。

天池o2o优惠券使用预测比赛-朴素贝叶斯模型

链接：<https://pan.baidu.com/s/1_7BLh1aT57sW-7d9tA34iQ>

提取码：c0ia

**打卡：**

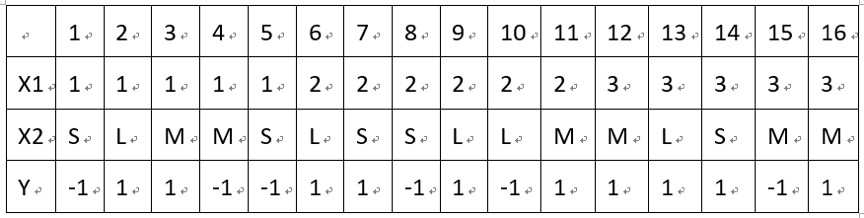
（1）内容：运行程序，上传结果，查看成绩，提交成绩截图。

（2）形式：图片，至少1张

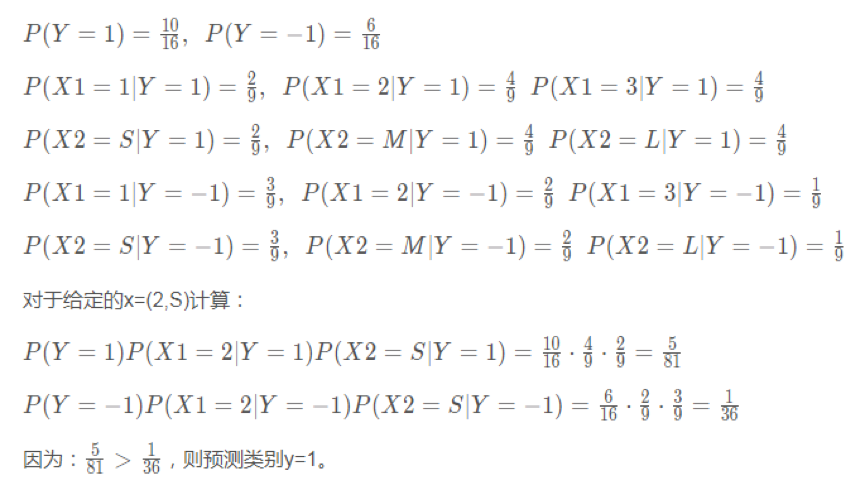
打卡截至日期：5/18

## 第3周作业参考答案

1. **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。写下你的计算过程和结果**



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



1. **Python项目：使用朴素贝叶斯过滤垃圾邮件（《机器学习实战》4.6）**

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

# ****第4周****

## 逻辑回归

**学习时长：**5/20—5/21

**任务简介：**学习逻辑回归，阅读《机器学习实战》5.1-5.3

**任务详解：**

逻辑回归（Logistic Regression）也是机器学习一个最基本也是最常用的算法模型。与线性回归不同的是，逻辑回归主要用于对样本进行分类。因此，逻辑回归的输出是离散值。对于二分类问题，通常我们令正类输出为1，负类输出为0。例如一个心脏病预测的问题：根据患者的年龄、血压、体重等信息，来预测患者是否会有心脏病，这就是典型的逻辑回归问题。重点是理解梯度上升算法。其实梯度上升与梯度下降原理是一样的，可以看作只是符号不同。

**打卡：**

（1）内容：尝试推导并解释逻辑回归的损失函数，推导参数w的b的梯度下降公式。写下你的推导过程，拍照，上传图片。

（2）形式：图片，至少1张

打卡截至提交日期：5/21

## 项目作业打卡日

**学习时长：**1天

**任务简介：**逻辑回归算法项目打卡日，完成本周项目作业。

**详细说明：**

**Python项目：**从疝气病症预测病马的死亡率（《机器学习实战》5.3）

链接：https://pan.baidu.com/s/1DvII-tFA-S0gPZXe6VtFzg

提取码：yb7u

**打卡：**

（1）内容：编写项目Python代码，运行正确，提交运行结果截图。

（2）形式：图片，至少1张

作业答案和讲解视频下周一公布

打卡截至提交日期：5/24

## 天池o2o比赛-逻辑回归模型

**学习时长：**1天

**任务简介：**

修改“天池o2o优惠券使用预测比赛-初级”的代码，调用scikit-learn库，使用逻辑回归算法来进行预测，运行程序，提交结果，查看成绩。

**任务详解：**

这部分代码给到大家，同学们也可以自行修改和优化。数据集之前给过了，记得把数据集放在代码所在的目录下。

天池o2o优惠券使用预测比赛-逻辑回归模型

链接：<https://pan.baidu.com/s/1wTxhmrDky3zlAIaM4kek8Q>

提取码：fzby

**打卡：**

（1）内容：运行程序，上传结果，查看成绩，提交成绩截图。

（2）形式：图片，至少1张

打卡截至提交日期：5/25

## 支持向量机基本原理

**学习时长：**1天

**任务简介：**学习支持向量机基本原理，阅读《机器学习实战》6.1-6.2

**任务详解：**

有些人认为，SVM是最好的现成的分类器，这里说的“现成”指的是分类器不加修饰即可直接使用。同时，这就意味着在数据上应用基本形式的SVM分类器就可以得到低错误率的结果。SVM能够对训练集之外的数据点做出很好的分类决策。重点内容了解是SVM的数学推导过程和软间隔SVM。难点是推导过程涉及大量的数学理论和公式，建议同学们感性理解为主，不要太拘泥于SVM的数学推导了，有大体的认识就好，关键是会熟练使用SVM。

**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

**打卡：**

（1）内容：为了防止SVM出现过拟合，应该对参数C进行如何设置？

（2）形式：文字，至少50字

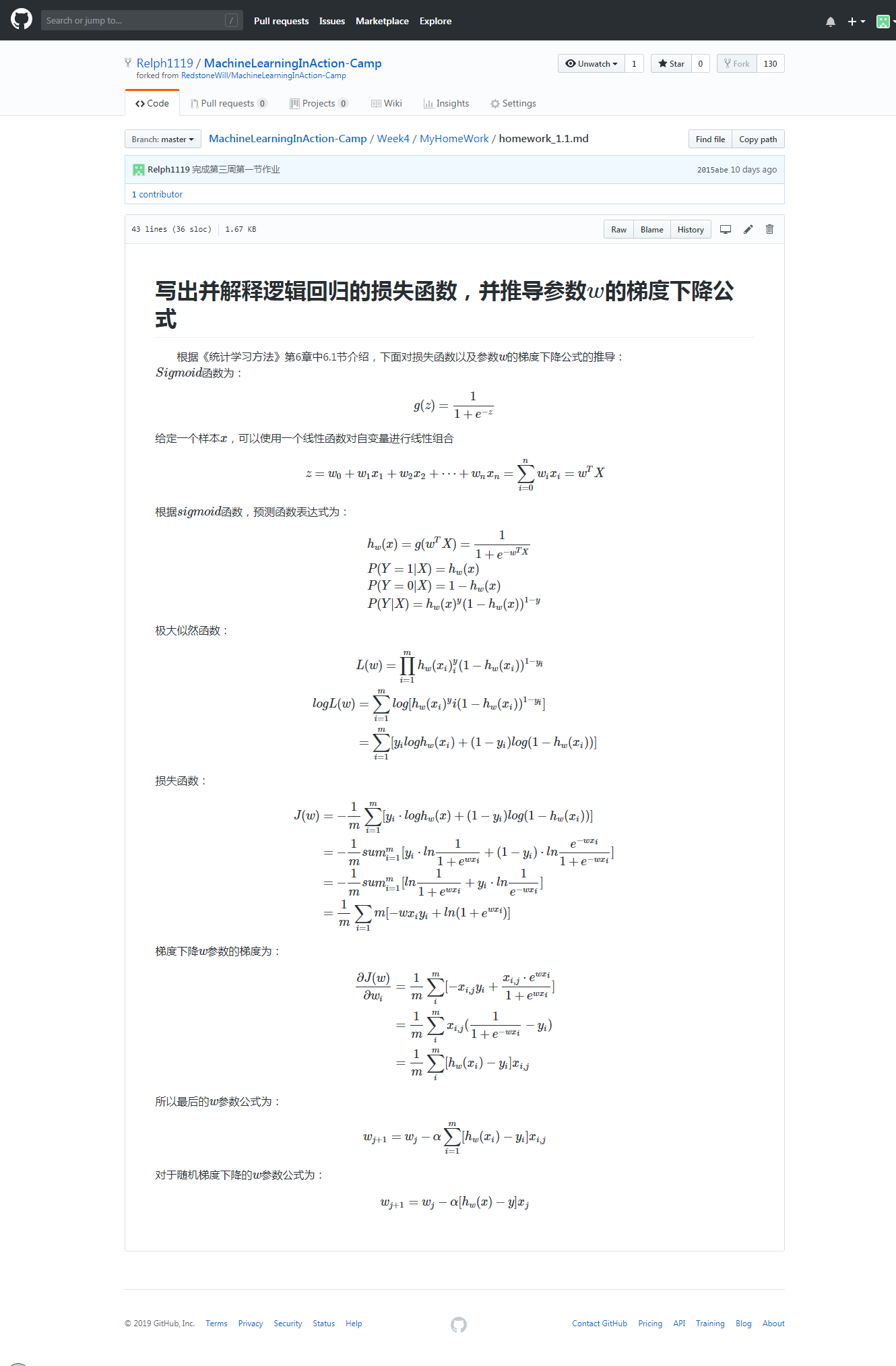
作业答案下周二发布

打卡截至提交日期：5/27

## ****第4周作业参考答案****

1. **尝试推导并解释逻辑回归的损失函数，推导参数w的b的梯度下降公式。写下你的推导过程。**

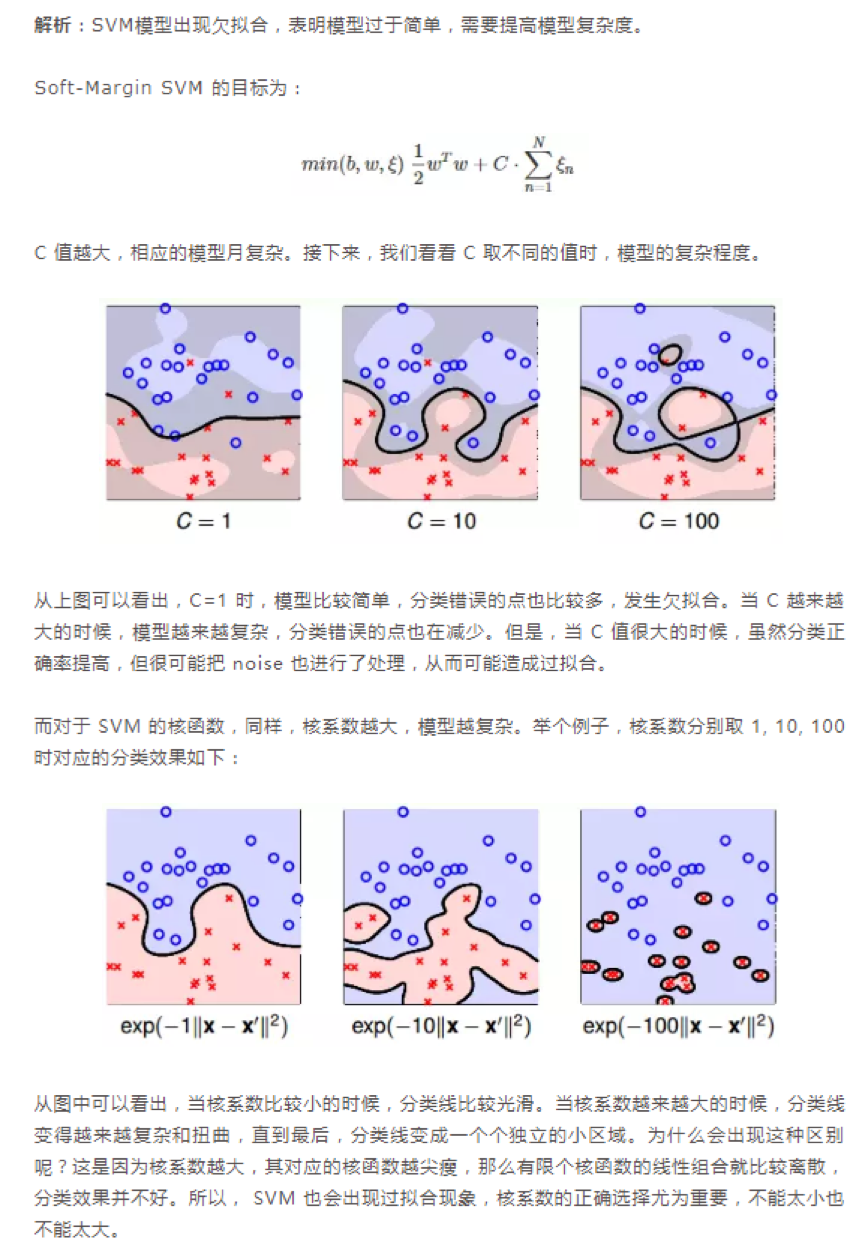
<https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week4/MyHomeWork/homework_1.1.md>



1. **Python项目：从疝气病症预测病马的死亡率（《机器学习实战》5.3）**

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

1. **为了防止SVM出现过拟合，应该对参数C进行如何设置？**



# 第5周

## SMO算法

**学习时长：**1天

**任务简介：**学习SMO算法，阅读《机器学习实战》6.3-6.4

**任务详解：**

SVM有很多实现，本节只会关注其中最流行的一种实现，即序列最小优化（SMO）算法。它是一种求解支持向量机的二次规划算法。重点是了解SMO算法的数学推导过程。SMO的推导比较复杂，同学们感性理解为主，不必太拘泥于纯数学理论中。掌握关键理论点，对应到SMO程序中，整体上理解程序。

**参考资料：**

李航《统计学习方法》第7章

**打卡：**

（1）内容：尝试推导SMO算法的过程。写下你的推导过程，拍照，上传图片。

（2）形式：图片，至少1张。

打卡截至提交日期：5/28

## 核函数

**学习时长：**1天

**任务简介：**学习核函数，阅读《机器学习实战》6.5-6.6

**任务详解：**

SVM优化中一个特别好的地方就是，所有的运算都可以写成内积的形式。向量的内积指的是两个向量相乘，之后得到单个标量或者数值。我们可以把内积运算替换成核函数，而不必做简化处理。将内积替换成核函数的方式被称为核技巧。重点了解典型的核函数：线性核函数和高斯和函数。这部分内容比较抽象，同学们不需要掌握核函数的推导过程，只要会使用核函数就好了。

**参考资料：**

李航《统计学习方法》第7章

https://mp.weixin.qq.com/s/cLovkwwgGJRgSSa1XWZ8eg

**打卡：**

（1）内容：SVM高斯核系数大小对模型复杂度有什么影响？

（2）形式：文字，至少60字

打卡截至提交日期：5/29

## 项目作业打卡日

**学习时长：**1天

**任务简介：**支持向量机算法项目打卡日，完成本周项目作业。

**任务详解：**

**Python项目：**手写识别问题回顾（《机器学习实战》6.6）

链接：https://pan.baidu.com/s/1A-ova-DwseM7pqOROtgPlA

提取码：0i60

**打卡：**

（1）内容：编写项目Python代码，运行正确，提交运行结果截图。

（2）形式：图片，至少1张

打卡截至提交日期：5/31

## 天池o2o比赛-支持向量机模型

**学习时长：**1天

**任务简介：**修改“天池o2o优惠券使用预测比赛-初级”的代码，调用scikit-learn库，使用支持向量机算法来进行预测，运行程序，提交结果，查看成绩。

**任务详解：**

这部分代码给大家，同学们也可以自行修改和优化。数据集之前给过了，记得把数据集放在代码所在的目录下。

天池o2o优惠券使用预测比赛-支持向量机模型

链接：https://pan.baidu.com/s/1BmQuInxFO1izM-NMXG4f\_g

提取码：y8t5

**打卡：**

（1）内容：运行程序，上传结果，查看成绩，提交成绩截图。

（2）形式：图片，至少1张

打卡截至提交日期：6/1

## 第5周作业参考答案

1. **Python项目：手写识别问题回顾（《机器学习实战》6.6）**

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

# 第6周

## Bagging、Boosting、AdaBoost

**学习时长：**1天

**任务简介：**《机器学习实战》7.1-7.2

**任务详解：**

当做重要的决定时，大家可能都会考虑吸取多个专家而不只是一个人的意见。机器学习处理问题时也是一样。这就是集成学习背后的思路。AdaBoost就是集成学习的典型代表，它是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。本节重点内容是理解AdaBoost的思想及其推导过程。

**参考资料：**

李航《统计学习方法》8.1-8.3

【视频】AdaBoost算法推导过程

**打卡：**

（1）内容：AdaBoost选择的子分类器是弱分类器还是强分类器？解释原因。

（2）形式：文字，至少60字

打卡截至提交日期：6/4

## AdaBoost实现、非均衡分

**学习时长：**1天

**任务简介：**《机器学习实战》7.3-7.7

**任务详解：**

本节将会建立一个但层决策树分类器。实际上，它是一个单节点的决策树。AdaBoost算法将应用在这个单层决策树之上。我们将在一个难数据集上应用AdaBoost分类器，以了解该算法是如何迅速超越其他分类器的。重点掌握如何使用Python构建单层决策树，如何构建完整的AdaBoost算法。

最后讨论非均衡分类问题的常用处理方法。掌握正确率、召回率的区别和ROC曲线。根据AUC如何判断模型的性能。

**打卡：**

（1）内容：数据不平衡时，分类性能度量指标哪些？

（2）形式：文字，至少60字

打卡截至日期：6/4

## 项目作业打卡日

**学习时长：**1天

**任务简介：**AdaBoost算法项目打卡日，完成本周项目作业。

**任务详解：**

**Python项目：**在一个较难数据集上应用AdaBoost（《机器学习实战》7.6）

链接：<https://pan.baidu.com/s/1q3hfCQ9XKo8dUTL5JVxUgg>

提取码：uqn5

**打卡：**

（1）内容：编写项目Python代码，运行正确，提交运行结果截图。

（2）形式：图片，至少1张

打卡截至日期：6/9

## 天池o2o比赛-AdaBoost

**学习时长：**1天

**任务简介：**

修改“天池o2o优惠券使用预测比赛-初级”的代码，调用scikit-learn库，使用AdaBoost算法来进行预测，运行程序，提交结果，查看成绩。

**任务详解：**

这部分代码给到大家，同学们也可以自行修改和优化。数据集之前给过了，记得把数据集放在代码所在的目录下。

天池o2o优惠券使用预测比赛-AdaBoost模型

链接：<https://pan.baidu.com/s/1YJiT55mDOEOA4dcu-r-aTA>

提取码：d4kw

**打卡：**

（1）内容：运行程序，上传结果，查看成绩，提交成绩截图。

（2）形式：图片，至少1张。

打卡截至日期：6/8

## 第6周作业参考答案

1. **Python项目：在一个较难数据集上应用AdaBoost**

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

# 第7周

## 线性回归

**学习时长：**1天

**任务简介：**《机器学习实战》8.1-8.3

**详细说明：**

在线性回归中，数据使用线性预测函数来建模，并且未知的模型参数也是通过数据来估计。这些模型被叫做线性模型。本节先介绍线性回归，然后引入了局部平滑技术，分析如何更好地拟合数据。重点掌握最小二乘法求解参数w的表达式。这里建议同学们使用梯度下降算法来求解线性回归问题。

**打卡：**

（1）内容：推导线性回归中最小二乘法公式，参数w的表达式。写下你的推导过程，拍照，上传图片。

（2）形式：图片，至少1张。

打卡截至日期：6/11

## 项目作业打卡日

**学习时长：**1天

**任务简介：**线性回归算法项目打卡日，完成本周项目作业。

**任务详解：**

**Python项目：**预测鲍鱼的年龄（《机器学习实战》8.3）

链接：<https://pan.baidu.com/s/1mh-iYJ-S0TM7hum5Mf5VBA>

提取码：i32p

**打卡：**

（1）内容：编写项目Python代码，运行正确，提交运行结果截图。

（2）形式：图片，至少1张

打卡截至时间：6/16

## 正则化、偏差与方差

**学习时长：**1 天

**任务简介：**《机器学习实战》8.4 - 8.5

**详细说明：**

正则化是线性回归中常用的防止过拟合技术，本节将介绍岭回归（ridge regression）、lasso法。lasso方法很好但是计算复杂。本节最后介绍了前向逐步回归，可以得到与lasso差不多的效果，且更容易实现。重点内容是理解不同的正则化技术以及偏差和方差的区别。

**打卡：**

（1）内容：岭回归和lasso有什么区别？

（2）形式：文字，至少50字

打卡截至日期：6/14

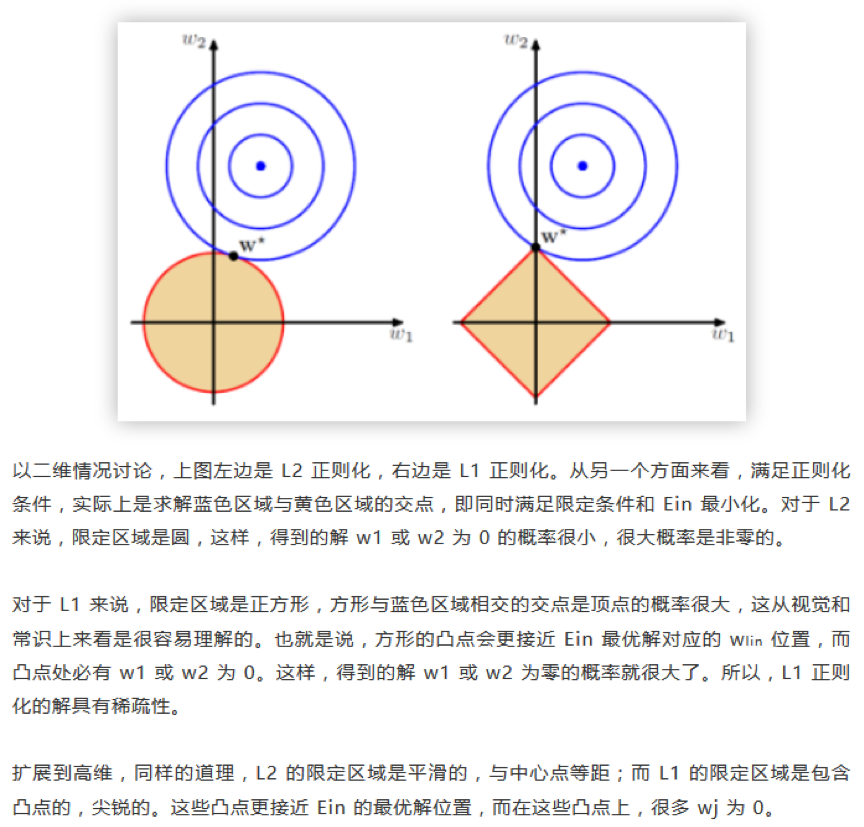
## 第7周作业参考答案

1. **Python项目：预测鲍鱼的年龄**

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

1. **岭回归和lasso有什么区别？**

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。



以下没有整理

# ****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>