

# 机器学习实战（第四期）

笔记整理人：天国之影

## 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>

## 1 第 1 周

### 1.1 绪论与准备

#### 1.1.1 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击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。

### 1.1.2 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/1ekuaaYjUCINJemFnRDGOvw>

提取码：[einz](#)

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。

**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。

## 1.2 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

## 1.3 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/1kmiT0IeB71eKfP0xlg1NIlw>

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

## 1.4 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](#)

## 1.5 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）。

## 2. 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week1/MyHomeWork/homework_3.2.ipynb)

## 3. 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week1/MyHomeWork/homework_3.3.ipynb)

## 2 第 2 周

### 2.1 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：**

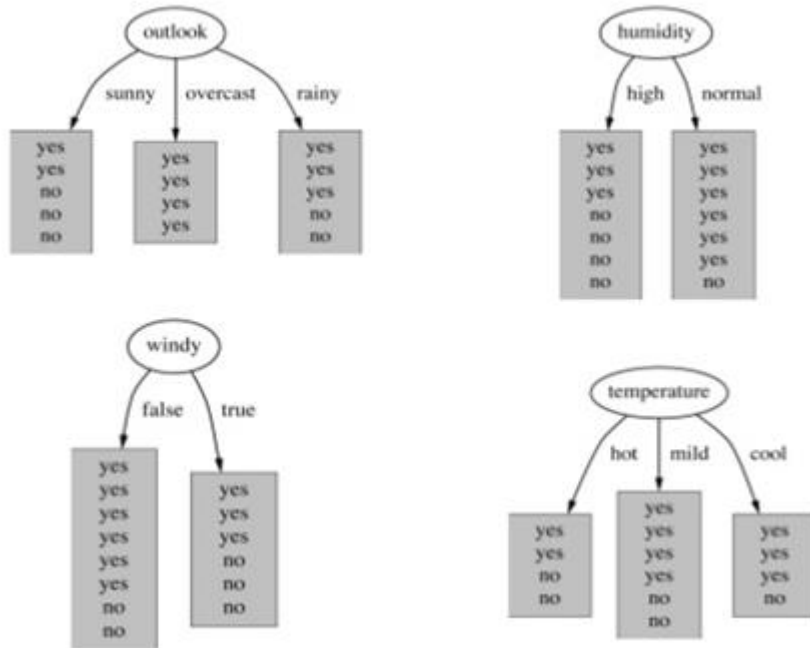
本节将通过算法一步步地构造决策树，并会涉及许多有趣的细节。首先我们讨论数学上如何使用信息论划分数据集，然后编写代码将理论应用到具体的数据集上，最后编写代码构建决策树。重点是掌握熵和信息增益的概念，根据信息增益的大小划分特征。难点是使用递归方式构建决策树，同学们需要重点攻克 3.1 中的代码。注意，3.2 节内容不重要，不需要看。

**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

## 2.2 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

## 2.3 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

## 2.4 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](https://pan.baidu.com/s/1KMWIUCS82W6_qV0Jld7P9g)

提取码: pqmy

打卡:

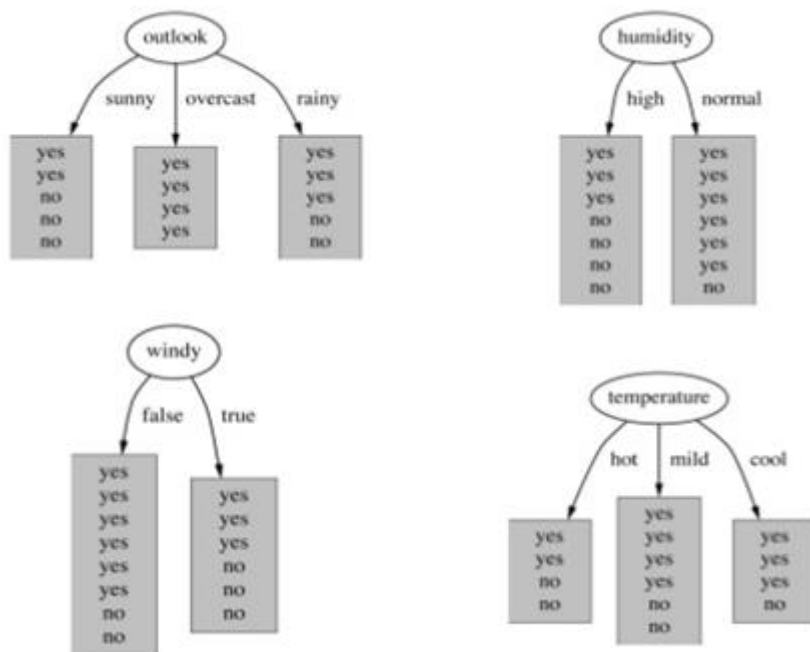
1) 内容: 运行程序, 上传结果, 查看成绩, 提交成绩截图。

2) 形式: 图片, 至少 1 张

打卡截至提交日期: 5/11

## 2.5 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week2/MyHomeWork/homework_1.2.md)

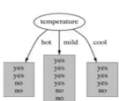
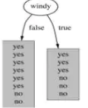
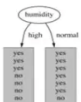

Branch: masterMachineLearningInAction-Camp / Week2 / MyHomeWork / homework\_1.2.mdFind fileCopy path

👤 Relph1119 修改第二周第一节作业第二题以适应github显示8890e3f 13 days ago

1 contributor

58 lines (57 sloc) | 2.68 KBRawBlameHistory📄🖋🗑

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？



解答：根据李航《统计学习方法》中，对信息增益有如下定义：特征A对训练数据集D的信息增益为 $g(D, A)$ ，定义为集合D的经验熵 $H(D)$ 与特征A给定条件下D的经验条件熵 $H(D|A)$ 之差，即

$$g(D, A) = H(D) - H(D|A)$$

从图中可以看到一共有14条数据，其中结果为yes的有9条，为no的有5条，故可以计算经验熵 $H(D)$ ：

$$\begin{aligned} H(D) &= -\frac{9}{14} \log\left(\frac{9}{14}\right) - \frac{5}{14} \log\left(\frac{5}{14}\right) \\ &= 0.6518 \end{aligned}$$

从左边到右，从上到下分别为图一、图二、图三、图四。

图一中，经验条件熵 $H(D|A)$ 为：

$$\begin{aligned} H(D|A) &= \frac{5}{14} \left( -\frac{2}{5} \log\left(\frac{2}{5}\right) - \frac{3}{5} \log\left(\frac{3}{5}\right) \right) \\ &\quad + \frac{4}{14} (-1 \cdot \log(1) - 0) \\ &\quad + \frac{5}{14} \left( -\frac{3}{5} \log\left(\frac{3}{5}\right) - \frac{2}{5} \log\left(\frac{2}{5}\right) \right) \\ &= 0.4807 \end{aligned}$$

图一中的信息增益为：

$$\begin{aligned} g(D, A) &= H(D) - H(D|A) = 0.6518 - 0.4807 \\ &= 0.1711 \end{aligned}$$

图二中，经验条件熵 $H(D|A)$ 为：

$$\begin{aligned} H(D|A) &= \frac{7}{14} \left( -\frac{3}{7} \log\left(\frac{3}{7}\right) - \frac{4}{7} \log\left(\frac{4}{7}\right) \right) \\ &\quad + \frac{7}{14} \left( -\frac{6}{7} \log\left(\frac{6}{7}\right) - \frac{1}{7} \log\left(\frac{1}{7}\right) \right) \\ &= 0.5465 \end{aligned}$$

图二中的信息增益为：

$$\begin{aligned} g(D, A) &= H(D) - H(D|A) = 0.6518 - 0.5465 \\ &= 0.1053 \end{aligned}$$

图三中，经验条件熵 $H(D|A)$ 为：

$$\begin{aligned} H(D|A) &= \frac{8}{14} \left( -\frac{6}{8} \log\left(\frac{6}{8}\right) - \frac{2}{8} \log\left(\frac{2}{8}\right) \right) \\ &\quad + \frac{6}{14} \left( -\frac{3}{6} \log\left(\frac{3}{6}\right) - \frac{3}{6} \log\left(\frac{3}{6}\right) \right) \\ &= 0.6184 \end{aligned}$$

图三中的信息增益为：

$$\begin{aligned} g(D, A) &= H(D) - H(D|A) = 0.6518 - 0.6184 \\ &= 0.0334 \end{aligned}$$

图四中，经验条件熵 $H(D|A)$ 为：

$$\begin{aligned} H(D|A) &= \frac{4}{14} \left( -\frac{2}{4} \log\left(\frac{2}{4}\right) - \frac{2}{4} \log\left(\frac{2}{4}\right) \right) \\ &\quad + \frac{6}{14} \left( -\frac{4}{6} \log\left(\frac{4}{6}\right) - \frac{2}{6} \log\left(\frac{2}{6}\right) \right) \\ &\quad + \frac{4}{14} \left( -\frac{3}{4} \log\left(\frac{3}{4}\right) - \frac{1}{4} \log\left(\frac{1}{4}\right) \right) \\ &= 0.6315 \end{aligned}$$

图四中的信息增益为：

$$\begin{aligned} g(D, A) &= H(D) - H(D|A) = 0.6518 - 0.6315 \\ &= 0.0203 \end{aligned}$$

可以得到，图一的信息增益最大

© 2019 GitHub, Inc. Terms Privacy Security Status Help

Contact GitHub Pricing API Training Blog About

## 2. 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week2/MyHomeWork/homework_1.3.ipynb)

## 3 第3周

### 3.1 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  $X_1$  和  $X_2$  为特征。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$X_1$	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3
$X_2$	S	L	M	M	S	L	S	S	L	L	M	M	L	S	M	M
$Y$	-1	1	1	-1	-1	1	1	-1	1	-1	1	1	1	1	-1	1

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

## 3.2 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

## 3.3 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

### 3.4 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](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

### 3.5 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  $X_1$  和  $X_2$  为特征。写下你的计算过程和结果

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3
X2	S	L	M	M	S	L	S	S	L	L	M	M	L	S	M	M
Y	-1	1	1	-1	-1	1	1	-1	1	-1	1	1	1	1	-1	1

[https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week3/MyHomeWork/homework\\_2.2.ipynb](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week3/MyHomeWork/homework_2.2.ipynb)

$$P(Y = 1) = \frac{10}{16}, P(Y = -1) = \frac{6}{16}$$

$$P(X1 = 1|Y = 1) = \frac{2}{9}, P(X1 = 2|Y = 1) = \frac{4}{9}, P(X1 = 3|Y = 1) = \frac{4}{9}$$

$$P(X2 = S|Y = 1) = \frac{2}{9}, P(X2 = M|Y = 1) = \frac{4}{9}, P(X2 = L|Y = 1) = \frac{4}{9}$$

$$P(X1 = 1|Y = -1) = \frac{3}{9}, P(X1 = 2|Y = -1) = \frac{2}{9}, P(X1 = 3|Y = -1) = \frac{1}{9}$$

$$P(X2 = S|Y = -1) = \frac{3}{9}, P(X2 = M|Y = -1) = \frac{2}{9}, P(X2 = L|Y = -1) = \frac{1}{9}$$

对于给定的 $x=(2,S)$ 计算：

$$P(Y = 1)P(X1 = 2|Y = 1)P(X2 = S|Y = 1) = \frac{10}{16} \cdot \frac{4}{9} \cdot \frac{2}{9} = \frac{5}{81}$$

$$P(Y = -1)P(X1 = 2|Y = -1)P(X2 = S|Y = -1) = \frac{6}{16} \cdot \frac{2}{9} \cdot \frac{3}{9} = \frac{1}{36}$$

因为： $\frac{5}{81} > \frac{1}{36}$ ，则预测类别 $y=1$ 。

## 2. 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week3/MyHomeWork/homework_2.3.ipynb)

## 4 第4周

### 4.1 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

## 4.2 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

## 4.3 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

## 4.4 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](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/Q5bFR3vDDXPhtzXIVAE3Rg>

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

## 4.5 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-](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week4/MyHomeWork/homework_1.1.md)

[Camp/blob/master/Week4/MyHomeWork/homework\\_1.1.md](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week4/MyHomeWork/homework_1.1.md)

Search or jump to...

Pull requests

Issues

Marketplace

Explore

Relph1119 / MachineLearningInAction-Camp

forked from RedstoneWill/MachineLearningInAction-Camp

Unwatch

1

★ Star

0

🍴 Fork

130

Code

Pull requests

Projects

Wiki

Insights

Settings

Branch: master

MachineLearningInAction-Camp / Week4 / MyHomeWork / homework\_1.1.md

Find file

Copy path

Relph1119 完成第三周第一节作业

2015ab10 days ago

1 contributor

43 lines (36 sloc) | 1.67 KB

Raw

Blame

History

## 写出并解释逻辑回归的损失函数，并推导参数 $w$ 的梯度下降公式

根据《统计学习方法》第6章中6.1节介绍，下面对损失函数以及参数 $w$ 的梯度下降公式的推导：  
*Sigmoid*函数为：

$$g(z) = \frac{1}{1 + e^{-z}}$$

给定一个样本 $x$ ，可以使用一个线性函数对自变量进行线性组合

$$z = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n = \sum_{i=0}^n w_i x_i = w^T X$$

根据*sigmoid*函数，预测函数表达式为：

$$h_w(x) = g(w^T X) = \frac{1}{1 + e^{-w^T X}}$$
$$P(Y = 1|X) = h_w(x)$$
$$P(Y = 0|X) = 1 - h_w(x)$$
$$P(Y|X) = h_w(x)^y (1 - h_w(x))^{1-y}$$

极大似然函数：

$$L(w) = \prod_{i=1}^m h_w(x_i)^{y_i} (1 - h_w(x_i))^{1-y_i}$$
$$\log L(w) = \sum_{i=1}^m \log[h_w(x_i)^{y_i} (1 - h_w(x_i))^{1-y_i}]$$
$$= \sum_{i=1}^m [y_i \log h_w(x_i) + (1 - y_i) \log(1 - h_w(x_i))]$$

损失函数：

$$J(w) = -\frac{1}{m} \sum_{i=1}^m [y_i \cdot \log h_w(x) + (1 - y_i) \log(1 - h_w(x_i))]$$
$$= -\frac{1}{m} \sum_{i=1}^m [y_i \cdot \ln \frac{1}{1 + e^{-w x_i}} + (1 - y_i) \cdot \ln \frac{e^{-w x_i}}{1 + e^{-w x_i}}]$$
$$= -\frac{1}{m} \sum_{i=1}^m [y_i \ln \frac{1}{1 + e^{-w x_i}} + (1 - y_i) \ln \frac{1}{e^{-w x_i} + 1}]$$
$$= \frac{1}{m} \sum_{i=1}^m [y_i \ln(1 + e^{-w x_i}) + (1 - y_i) \ln(1 + e^{w x_i})]$$

梯度下降 $w$ 参数的梯度为：

$$\frac{\partial J(w)}{\partial w_i} = \frac{1}{m} \sum_{i=1}^m [-x_{i,j} y_i + \frac{x_{i,j} \cdot e^{-w x_i}}{1 + e^{-w x_i}}]$$
$$= \frac{1}{m} \sum_{i=1}^m x_{i,j} (\frac{1}{1 + e^{-w x_i}} - y_i)$$
$$= \frac{1}{m} \sum_{i=1}^m [h_w(x_i) - y_i] x_{i,j}$$

所以最后的 $w$ 参数公式为：

$$w_{j+1} = w_j - \alpha \sum_{i=1}^m [h_w(x_i) - y_i] x_{i,j}$$

对于随机梯度下降的 $w$ 参数公式为：

$$w_{j+1} = w_j - \alpha [h_w(x) - y] x_j$$

© 2019 GitHub, Inc. Terms Privacy Security Status Help

Contact GitHub Pricing API Training Blog About

## 2. 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week4/MyHomeWork/homework_1.2.ipynb)

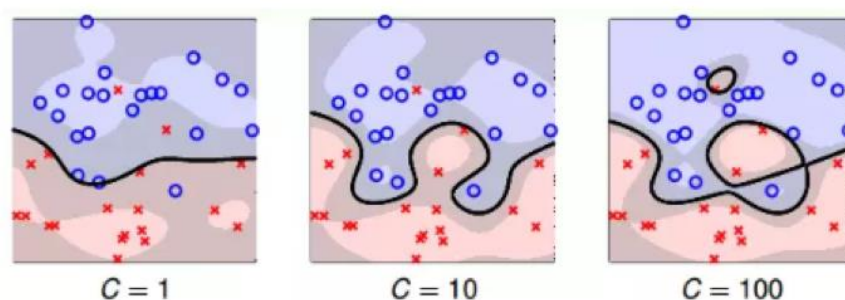
## 3. 为了防止 SVM 出现过拟合，应该对参数 C 进行如何设置？

解析：SVM模型出现欠拟合，表明模型过于简单，需要提高模型复杂度。

Soft-Margin SVM 的目标为：

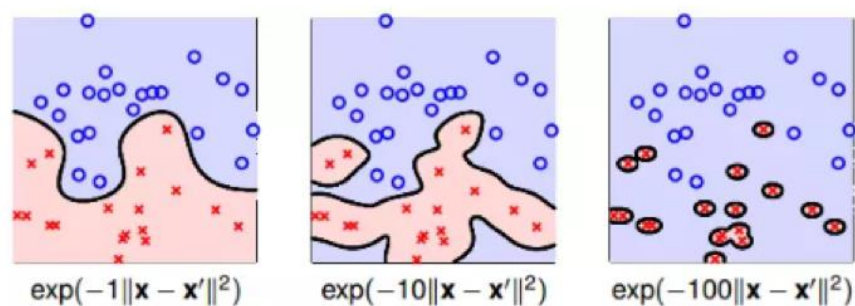
$$\min(b, w, \xi) \frac{1}{2} w^T w + C \cdot \sum_{n=1}^N \xi_n$$

C 值越大，相应的模型月复杂。接下来，我们看看 C 取不同的值时，模型的复杂程度。



从上图可以看出，C=1 时，模型比较简单，分类错误的点也比较多，发生欠拟合。当 C 越来越大的时候，模型越来越复杂，分类错误的点也在减少。但是，当 C 值很大的时候，虽然分类正确率提高，但很可能把 noise 也进行了处理，从而可能造成过拟合。

而对于 SVM 的核函数，同样，核系数越大，模型越复杂。举个例子，核系数分别取 1, 10, 100 时对应的分类效果如下：



从图中可以看出，当核系数比较小的时候，分类线比较光滑。当核系数越来越大的时候，分类线变得越来越复杂和扭曲，直到最后，分类线变成一个个独立的小区域。为什么会出现这种区别呢？这是因为核系数越大，其对应的核函数越尖瘦，那么有限个核函数的线性组合就比较离散，分类效果并不好。所以，SVM 也会出现过拟合现象，核系数的正确选择尤为重要，不能太小也不能太大。

## 5 第 5 周

### 5.1 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

## 5.2 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

## 5.3 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-DwseM7pqOROtGPIA>

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

## 5.4 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](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

## 5.5 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week6/MyHomeWork/homework_1.2.ipynb)

## 6 第 6 周

### 6.1 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

## 6.2 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:

本节将会建立一个单层决策树分类器。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

## 6.3 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

## 6.4 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

## 6.5 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week7/MyHomeWork/homework_1.2.ipynb)

## 7 第 7 周

### 7.1 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

## 7.2 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

## 7.3 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

## 7.4 CART 树

学习时长：1 天

任务简介：《机器学习实战》9.1-9.3

### 任务详解：

本节将介绍一个分类回归树 CART，该算法既可以用于分类还可以用于回归。本节将会利用 Python 来构建并显示 CART 树，代码会保持足够的灵活性以便能用于多个问题当中。重点内容是如何构建 CART 树，掌握 CART 树的切分函数。希望同学们认真读书上的代码，加深对理论的理解。

【视频】CART 树构建

### 打卡：

（1）内容：对比 CART 与 ID3 算法。

（2）形式：文字，至少 60 字

打卡截至日期：6/17

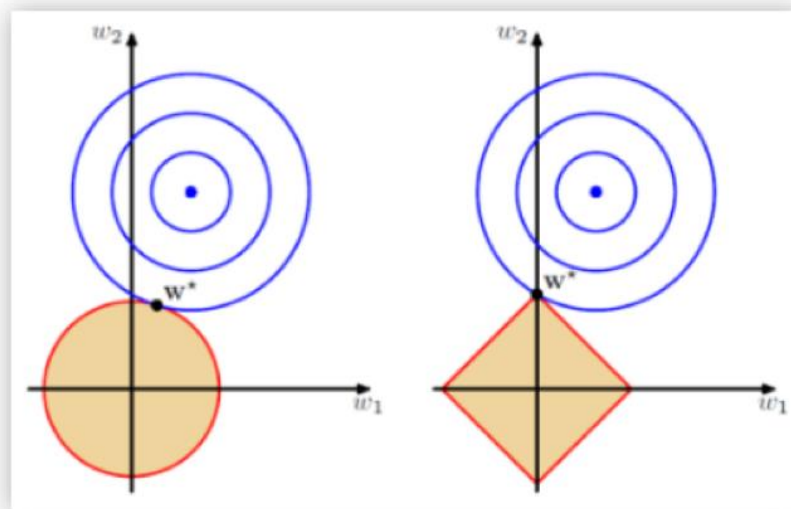
## 7.5 第 7 周作业参考答案

### 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](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week8/MyHomeWork/homework_2.2.ipynb)

## 2. 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。



以二维情况讨论，上图左边是 L2 正则化，右边是 L1 正则化。从另一个方面来看，满足正则化条件，实际上是求解蓝色区域与黄色区域的交点，即同时满足限定条件和  $E_{in}$  最小化。对于 L2 来说，限定区域是圆，这样，得到的解  $w_1$  或  $w_2$  为 0 的概率很小，很大概率是非零的。

对于 L1 来说，限定区域是正方形，方形与蓝色区域相交的交点是顶点的概率很大，这从视觉和常识上来看是很容易理解的。也就是说，方形的凸点会更接近  $E_{in}$  最优解对应的  $w_{lin}$  位置，而凸点处必有  $w_1$  或  $w_2$  为 0。这样，得到的解  $w_1$  或  $w_2$  为零的概率就很大了。所以，L1 正则化的解具有稀疏性。

扩展到高维，同样的道理，L2 的限定区域是平滑的，与中心点等距；而 L1 的限定区域是包含凸点的，尖锐的。这些凸点更接近  $E_{in}$  的最优解位置，而在这些凸点上，很多  $w_j$  为 0。

## 8 第 8 周

### 8.1 树剪枝

**学习时长：**1 天

**任务简介：**《机器学习实战》9.4

**详细说明：**

一棵树如果节点过多，表明该模型可能对数据进行了过拟合。通过降低决策树的复杂度来避免过拟合的过程称为剪枝。重点了解两种树剪枝的方法：预剪枝和后剪枝，掌握相应的 Python 代码。

**打卡：**

(1) 内容：对比预剪枝和后剪枝的区别。

(2) 形式：文字，至少 60 字

打卡截至日期：6/18

## 8.2 模型树

**学习时长：**1 天

**任务简介：**《机器学习实战》9.5-9.6

**详细说明：**

用树来对数据建模，除了把叶节点简单地设为常数值外，还有一种方法是把叶节点设定为分段线性函数，这就是所谓的分段线性是指模型由多个线性片段组成。重点内容是掌握模型树的构建方法以及回归树和模型树的区别。

**打卡：**

(1) 内容：如何比较模型树和回归树哪个更好

(2) 形式：文字，至少 60 字

打卡截至时间：6/19

## 8.3 项目作业打卡日

**学习时长：**1 天

**任务简介：**树回归算法项目打卡日，完成本周项目作业。

**任务详解：**

**Python 项目：**树回归与标准回归的比较（《机器学习实战》9.6）

链接: [https://pan.baidu.com/s/1a\\_ZZMm-NI7njjE50LdXS8w](https://pan.baidu.com/s/1a_ZZMm-NI7njjE50LdXS8w)

提取码: [yvlp](#)

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: 6/23

## 8.4 天池 o2o 优惠券使用预测比赛 (进阶)

学习时长: 1 天

**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进阶源代码, 运行程序, 提交结果, 查看成绩。

**任务详解:**

天池 o2o 优惠券使用预测比赛进阶源代码和数据集

链接: <https://pan.baidu.com/s/1tUWYgjMVFYQCfeI4c4VnoQ>

提取码: [0vu7](#)

天池 o2o 优惠券使用预测比赛解析 (进阶) 视频讲解

打卡:

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

(2) 形式: 图片, 至少 1 张

打卡截止日期: 6/22

## 8.5 第 8 周作业参考答案

1. Python 项目: 树回归与标准回归的比较 (《机器学习实战》9.6)

[https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week9/MyHomeWork/homework\\_1.1.ipynb](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week9/MyHomeWork/homework_1.1.ipynb)

## 9 第 9 周

### 9.1 k-means 聚类

学习时长：1 天

任务简介：《机器学习实战》10.1-10.2

任务详解：

聚类是一种无监督式学习，它将相似的对象归到同一个簇中。k-means 可以发现 k 个不同的簇，且每个簇得中心采用簇中所含值的均值计算而成。重点内容是掌握 k-means 算法是如何更新簇重心点的以及对应的 Python 代码。

打卡：

(1) 内容：请用文字描述 k-means 聚类算法的优点和缺点。

(2) 形式：文字，至少 50 字

打卡截至日期：6/25

### 9.2 二分 k-means 聚类

学习时长：1 天

任务简介：《机器学习实战》10.3/10.4.2

任务详解：

为克服 k-means 算法收敛于局部最小值的问题，有人提出了另一种称为二分 k-means 的算法。该算法首先将所有点作为一个簇，然后将该簇一分为二。之后选择其中一个簇继续进行划分，选择哪个簇进行划分取决于对其划分是否可以最大化程度降低 SSE 的值。上述基于 SSE 的划分过程不断重复，直到得到用户指定的簇数目为止。重点掌握二分 k-means 算法流程以及对应的 Python 代码。

【视频】二分 k-means 聚类

打卡：

- (1) 内容：二分 k-means 聚类相比较 k-means 聚类，有哪些优点？
- (2) 形式：文字，至少 50 字

## 9.3 项目作业打卡日

学习时长：1 天

任务简介：k-mean 算法项目打卡日，完成本周项目作业。

任务详解：

**Python 项目：**对地理坐标进行聚类（《机器学习实战》10.4.2）

链接：<https://pan.baidu.com/s/1OnyPV4G0WeOlP0EfOQ8V7g>

提取码：[7dcq](#)

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：6/30

## 9.4 第 9 周作业参考答案

**1. Python 项目：**对地理坐标进行聚类（《机器学习实战》10.4.2）

[https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week11/MyHomeWork/homework\\_2.1.ipynb](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week11/MyHomeWork/homework_2.1.ipynb)

## 10 第 10 周

### 10.1 降维 PCA

学习时长：1 天

任务简介：《机器学习实战》13.1-13.2

### 任务详解：

PCA 全称 Principal Component Analysis，即主成分分析，是一种常用的数据降维方法。它可以通过线性变换将原始数据变换为一组各维度线性无关的表示，以此来提取数据的主要线性分量。重难点是掌握 PCA 的数学推导过程，理解为什么第一个主成分就是特征方差最大的特征。

【视频】PCA 数学原理

### 打卡：

(1) 内容：PCA 算法的优点和缺点。

(2) 形式：文字，至少 50 字

打卡截至日期：7/1

## 10.2 项目作业打卡日 1

学习时长：1 天

任务简介：PCA 算法项目打卡日，完成本周项目作业。

### 任务详解：

**Python 项目：**利用 PCA 对半导体制造数据降维（《机器学习实战》13.3）

链接：<https://pan.baidu.com/s/1nC09zqcsMpgOwXyKyf8YGw>

提取码：[q43y](#)

### 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：7/2

## 10.3 奇异值分解 SVD

学习时长：1 天

任务简介：《机器学习实战》14.1-14.4

### 详细说明：

本节将介绍 SVD 的概念及其能够进行数据约简的原因。然后介绍基于



Python 的 SVD 实现以及将数据映射到低维空间的过程。重点是掌握如何使用 Python 实现 SVD，掌握在推荐系统中如何进行相似度计算。难点是 SVD 的数学推导过程，这需要较强的线性代数知识，同学们不必太拘泥此处的推导，可直接记住 SVD 的公式即可。

【视频】SVD 数学原理

打卡：

(1) 内容：SVD 的应用场景有哪些？

(2) 形式：文字，至少 50 字

打卡截至日期：7/3

## 10.4 项目作业打卡日 2

学习时长：1 天

任务简介：

SVD 算法项目打卡日，完成本周项目作业。

任务详解：

**Python 项目：**餐馆菜肴推荐引擎（《机器学习实战》14.5）

链接：[https://pan.baidu.com/s/1gT3E16nR2sz\\_BfGtRtv21g](https://pan.baidu.com/s/1gT3E16nR2sz_BfGtRtv21g)

提取码：[gkw9](#)

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：7/5

## 10.5 项目作业打卡日 3

学习时长：1 天

任务简介：SVD 算法项目打卡日，完成本周项目作业。

任务详解：

**Python 项目：**基于 SVD 的图像压缩（《机器学习实战》14.6）

链接: <https://pan.baidu.com/s/1eYc74UnXPtWIaCRSzzJPww>

提取码: 9vk9

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: 7/6

## 10.6 第 10 周作业参考答案

1. Python 项目: 利用 PCA 对半导体制造数据降维 (《机器学习实战》13.3)

[https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch13/homework\\_1.1.ipynb](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch13/homework_1.1.ipynb)

2. Python 项目: 餐馆菜肴推荐引擎 (《机器学习实战》14.5)

[https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch14/homework\\_2.1.ipynb](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch14/homework_2.1.ipynb)

3. Python 项目: 基于 SVD 的图像压缩 (《机器学习实战》14.6)

[https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch14/homework\\_2.2.ipynb](https://github.com/Relph1119/MachineLearningInAction-Camp/blob/master/Week14/MyHomeWork/ch14/homework_2.2.ipynb)