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）

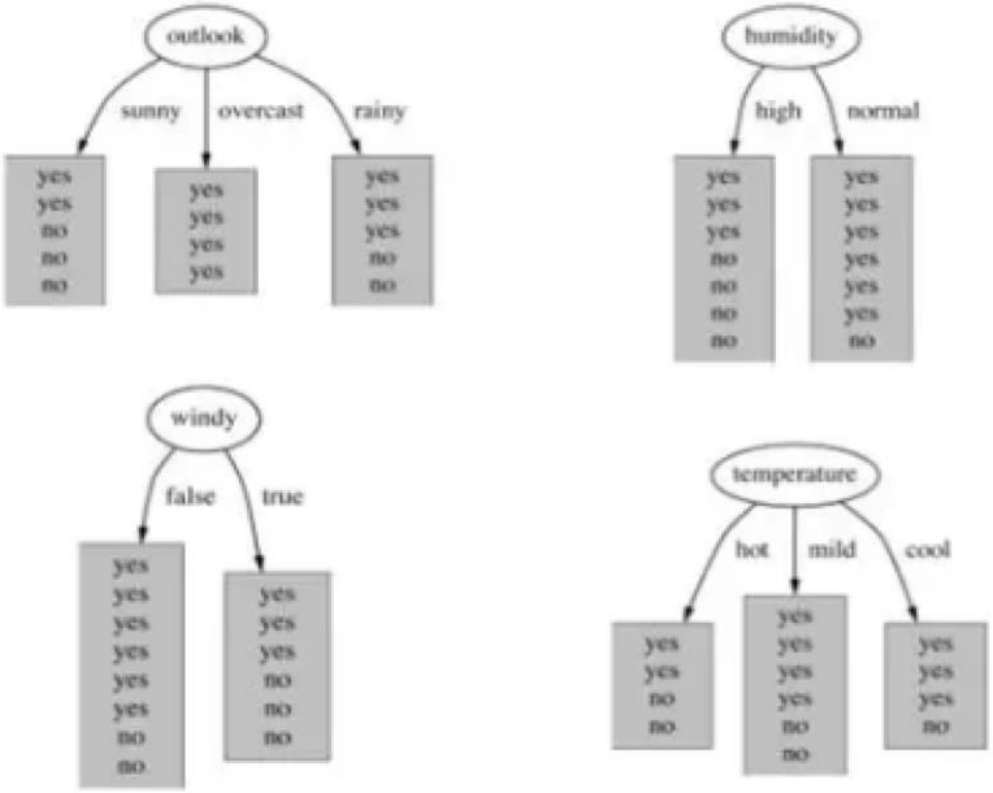
参考文献：李航《统计学习方法》第 5 章中的 5.1-5.3 节

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：

	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

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

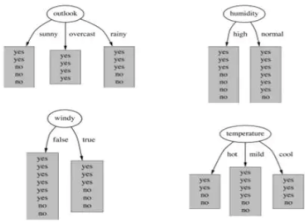
1.1 见 GitHub 链接：

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

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：

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

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：

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

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：

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

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：

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

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

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

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

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

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

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

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

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。

2.3 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

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 0 Projects 0 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 2015abe 10 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

$$z=w_0+w_1x_1+w_2x_2+\cdots+w_nx_n=\sum_{i=0}^{n}w_ix_i=w^TX$$

根据$sigmoid$函数，预测函数表达式为：

$$\begin{aligned}h_w(x)&=g(w^TX)=\frac{1}{1+e^{-w^TX}}\\P(Y=1|X)&=h_w(x)\\P(Y=0|X)&=1-h_w(x)\\P(Y|X)&=h_w(x)^y(1-h_w(x))^{1-y}\end{aligned}$$

极大似然函数：

$$\begin{aligned}L(w)&=\prod_{i=1}^{m}h_w(x_i)_i^y(1-h_w(x_i))^{1-y_i}\\logL(w)&=\sum_{i=1}^{m}log[h_w(x_i)^yi(1-h_w(x_i))^{1-y_i}]\\&=\sum_{i=1}^{m}[y_iloghw(x_i)+(1-y_i)log(1-h_w(x_i))]\end{aligned}$$

损失函数：

$$\begin{aligned}J(w)&=-\frac{1}{m}\sum_{i=1}^{m}[y_i\cdot loghw(x)+(1-y_i)log(1-h_w(x_i))]\\&=-\frac{1}{m}sum_{i=1}^{m}[y_i\cdot ln\frac{1}{1+e^{wx_i}}+(1-y_i)\cdot ln\frac{e^{-wx_i}}{1+e^{-wx_i}}]\\&=-\frac{1}{m}sum_{i=1}^{m}[ln\frac{1}{1+e^{wx_i}}+y_i\cdot ln\frac{1}{e^{-wx_i}}]\\&=\frac{1}{m}\sum_{i=1}^{m}[-wx_iy_i+ln(1+e^{wx_i})]\end{aligned}$$

梯度下降w参数的梯度为：

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

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

2.2 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

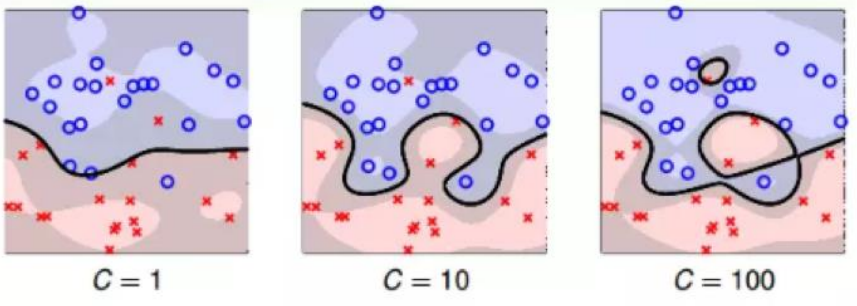
1.1

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：

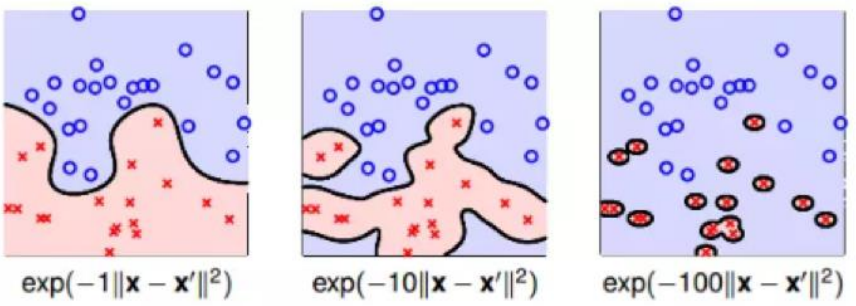
$$min(b,w,\xi)\ \frac{1}{2}w^Tw+C\cdot\sum_{n=1}^{N}\xi_n$$

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。

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

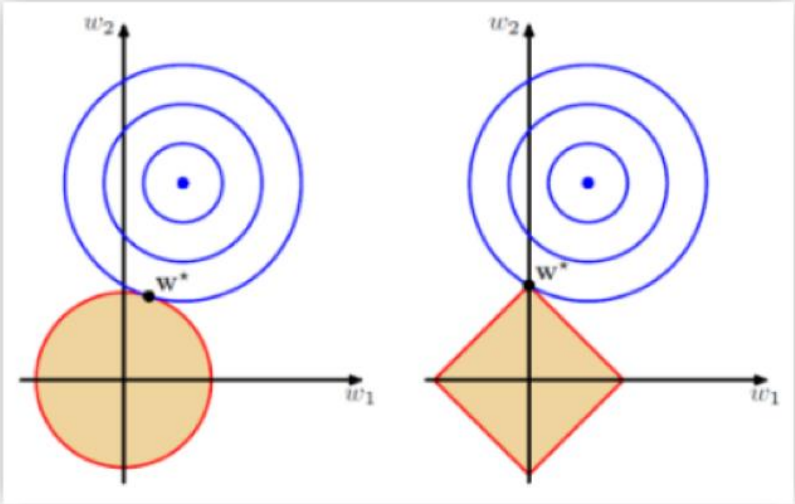
第 5 周作业参考答案

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。



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 Ein 最小化。对于 L2 来说，限定区域是圆，这样，得到的解 w1 或 w2 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 Ein 最优解对应的 w_{lin} 位置，而凸点处必有 w1 或 w2 为 0。这样，得到的解 w1 或 w2 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 Ein 的最优解位置，而在这些凸点上，很多 wj 为 0。

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