**《机器学习》训练营作业（第十期）**

笔记整理人：天国之影

# ****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. 本课程作业的所有代码基于Python3.7.2，在Jupyter Notebook上完成。
3. 课程老师：Jian老师
4. 电子版：周志华《机器学习》西瓜书链接：<https://pan.baidu.com/s/1Kr-8ayfVv2fjrfz9nfaR3A>

密码：d5zm

1. 附赠：李航《统计学习方法》链接：<https://pan.baidu.com/s/1artFMZGKJ_UMd0sI488fkQ>

密码：mr41

**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/MachineLearning-WatermelonBook>

# 第1周

## 学习机器学习绪论

**学习时长：**1天

**任务标题：**绪论

**任务简介：**

学习机器学习绪论，了解预备知识，认识群内其他小伙伴，下载书籍电子版，提前自己预习观看。

**任务详解：**

本部分是机器学习的绪论部分，对于这部分的学习，我并未让大家去阅读西瓜书的第一章的绪论部分，因为那部分对于初学者来讲太难了，应该就是天书，越看越蒙，这部分内容希望大家在学习完整个机器学习课程后，再回头看就好。为了让大家更轻松更好的去理解机器学习，我们特此做了这个视频，主要讲述了两个方面内容，一个是机器学习的应用案例，这部分是核心内容，另外一个是机器学习理论的一个大概的分类，主要目的是想告诉同学们机器学习是什么？机器学习有哪些应用？你们应该重点关注哪些应用？希望大家在本节内容学习后，心里已经有了这三个问题的答案。

**书籍下载：**到公众号深度之眼后台回复“西瓜书” ，即可领取电子版

**打卡：**

（1） 内容：请给自己立一个Flag，告诉自己未来2个月一定能把西瓜书啃完！打完这个卡，你吹的牛逼，就会被所有人看到。所以自己吹的牛逼，含着泪也要坚持下去。

（2）形式：文字，最少10字；

打卡截止日期：6/19

## 达观杯NLP算法大赛

**学习时长：**6/20-6/21

**任务标题：**打达观杯NLP算法大赛

**任务简介：**请按照教程讲解的内容报名，打比赛、提交比赛成绩、查看名次。 并在学习群内，积极参与比赛战队的组建

**任务详解：**

**特别注释：**

第一周就打比赛，很多人会感觉懵逼，说什么都不会怎么打比赛。对的，这就是我们的目的，你先不要尝试理解，先按照老师教程，完整跑一遍比赛流程，体验一下算法的实际应用。之后在学习理论的过程中，把理论应用于比赛中，看看不同算法对比赛成绩和结果的影响，对比其中的不同，尝试提高比赛成绩。

本部分主要是给大家讲一个实际比赛的内容，这部分的主要目的是带着大家走完一个参加机器学习比赛的完整过程。对于没有参加过机器学习比赛的同学来讲，可能觉得参加比赛会很难，甚至不知如何参加，但是我想告诉你们其实很简单，希望能够带大家进入机器学习比赛这个道路上来，让大家更好的进行机器学习的实践，这也是我们的初衷，希望大家可以开启比赛之门， 同时，我们提供了一个傻瓜式的完成这个比赛过程的操作文档pdf，请大家自行下载，然后按着操作流程走完这个过程，其中会涉及比赛的报名/开发软件的安装/代码的编，写和执行/结果的提交，请大家务必完成。

PDF下载链接：

<https://pan.baidu.com/s/1TZepfl6DXItn1cwxbf-rMQ>

提取码：zc1e

不想下载也可以直接看图文版

**打卡：**

（1） 内容：提交报名成功的截图，提交比赛成绩和名次的截图

（2） 形式：图片，最少2张；

打卡截止日期：6/21

## 本周学习任务简单总结

**学习时长：**1天

**任务简介：**简单回顾本周学到几个重要知识

**任务详解：**

每一周的学习任务都比较重，第一次学过之后特别容易忘，所以在周日及时做一个要点回顾，会让学习效率大大的提升，不会的知识也会越来越少

**打卡：**

（1）内容：请用文字描述，本周所学知识的重点，也可以思维导图、手写、电子版截图或者拍照均可，格式不限

（2） 形式：文字，最少20字；图片，0-9张；

打卡截止时间：6/23

## 参考答案

1. **绪论笔记**

<https://github.com/Relph1119/MachineLearning-WatermelonBook/blob/master/Week1/MyHomeWork/homework_1.md>

1. **1小时打完一场比赛代码，详见github**

<https://github.com/Relph1119/MachineLearning-WatermelonBook/blob/master/Week1/MyHomeWork/homework_2.py>

# 第2周

## 学习线性模型

**学习时长：**6/24——6/25

**任务标题：**线性模型

**任务简介：**

学习西瓜书3.1/3.2/3.3，观看西瓜书公式推导学习指南和线性模型公式推导视频。

**任务详解：**

这部分就是我们要学习的第一个机器学习模型，线性模型，一个简单而重要的模型，前两节主要讲线性回归，这个就是我们过去学过的最小二乘法的内容，相信大家都不陌生的。而第三小节讲的对数几率回归才是我们正章的核心，这个算法也是目前应用最广的算法之一，希望大家能够完全理清这个算法的工作流程，这也是对大家在本周最为核心的一个要求。

视频：

（1）西瓜书公式推导——学习指南

（2）一元线性回归公式推导

（3）多元线性回归公式推导

（4）对数几率回归公式推导

**打卡：**

（1）内容：logistic回归解决的是什么问题？它的表达式是什么，求导公式是什么？

（2）形式：打卡提交总结的文字和公式推导的图片（20字以上，1张图片）

打卡截止日期：6/25

## 学习sklearn包中逻辑回归算法的使用

**学习时长：**6/27—6/28

**任务标题：**学习sklearn包中逻辑回归算法的使用

**任务简介：**实验-sklearn-user guide 1.1.11

**任务详解：**

本节的主要任务是，了解sklearn包中逻辑回归算法的使用。逻辑回归算法大家基本了解了，简单使用大家也用过了。

但是这次实验的主要目的，是让大家了解一下sklearn包的官网，这个官网提供了详细的教程和案例，对于机器学习来讲是一种最为宝贵的资源，我们应该怎么使用这个官网来进行机器学习的学习，和实际任务的开发，这个资源也是我们日后学习的一个重点内容。

资料链接：

<https://scikit-learn.org/dev/modules/linear_model.html#logistic-regression>

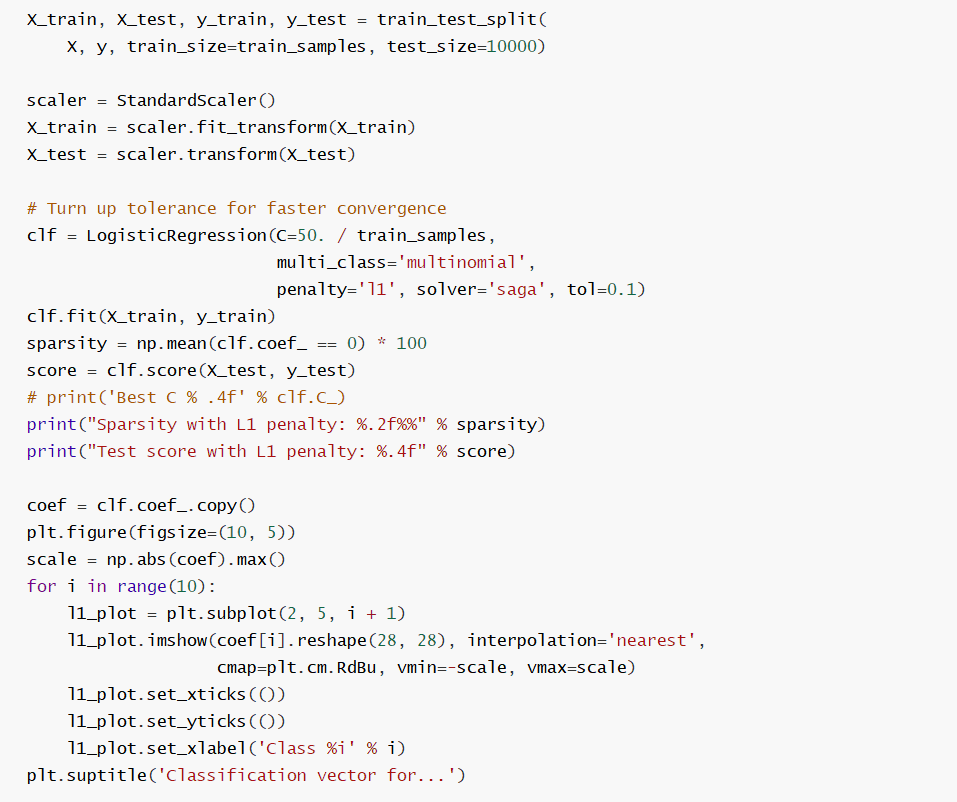
**打卡：**

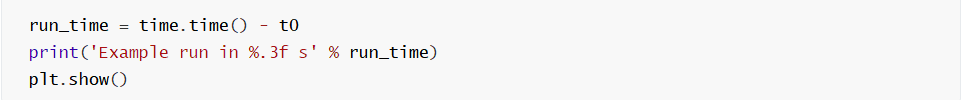
（1）内容：参照链接给的代码，实际动手敲一遍，使用sklearn的逻辑回归接口实现MNIST 分类，将运行结果截图。

<https://scikit-learn.org/stable/autoexamples/linearmodel/plotsparselogisticregressionmnist.html#sphx-glr-auto-examples-linear-model-plot-sparse-logistic-regression-mnist-py>

如果链接打不开，可直接看下方代码：







（2） 形式：打卡提交图片，最少1张；

打卡截止时间：6/28

## 本周学习任务简单总结

**学习时长：**1天

**任务简介：**简单回顾本周学到几个重要知识

**任务详解**：

每一周的学习任务都比较重，第一次学过之后特别容易忘，所以在周日及时做一个要点回顾，会让学习效率大大的提升，不会的知识也会越来越少

**打卡：**

（1） 内容：请用文字描述，本周所学知识的重点，也可以思维导图、手写、电子版截图或者拍照均可，格式不限

（2） 形式：文字，最少20字；图片，0-9张；

打卡截止时间：6/30

## 参考答案

1. **logistic回归解决的是什么问题？它的表达式是什么，求导公式是什么？**

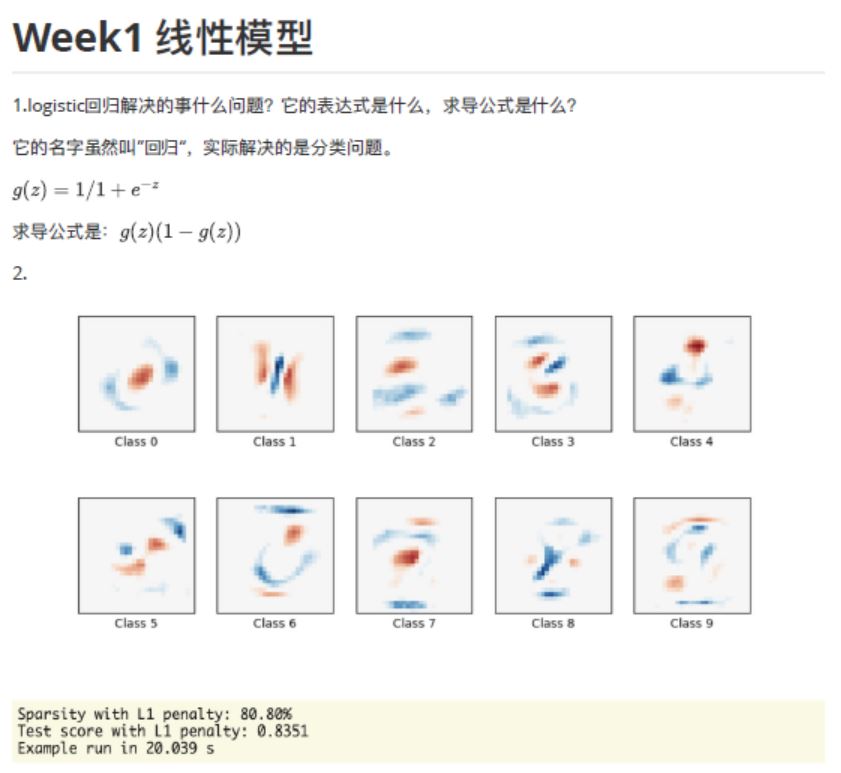
**解答：**

它的名字虽然是“回归”,实际解决的是分类问题。

表达式：

求导公式：

1. **使用sklearn的逻辑回归接口实现MNIST 分类，运行结果截图**



# 第3周

**本周学习内容介绍：**

本周我们要学习的主要内容为西瓜书的第四章 决策树，在如今的机器学习领域中，树模型可以说是最为重要的模型，在各种数据挖掘的竞赛中和生产环境中，威力巨大的大杀器XGboost和lightGBM 两个工具都是基于树模型构建的，所以说大家要认真掌握本章树模型的每一个基本概念。我们需要重点把握住的一个问题就是，一棵树是怎么构建起来的？大家完成本章学习后，要做到心理能够默念构建一颗树的整个过程。

## 决策树的分裂准则

**学习时长：**7/1—7/3

**任务标题：**决策树的分裂准则

**任务简介：**学习西瓜书4.1/4.2，观看西瓜书决策树公式推导视频

**任务详解：**

这部分主要讲两方面的内容：第一，面对一个实际的数据集，应该如何构建出一颗树；第二，在构建树的过程中，树分裂节点时，如何选择出最优的属性作为分裂节点。这两方面是树模型的最为关键的部分，请大家认真搞清楚其中的每一个概念。希望大家能够掌握构建一个棵树的基本流程。

**视频：**

决策树

**打卡：**

（1）内容：常见决策树算法有哪些？它们的划分准则分别是什么，是否有缺陷？

（2） 形式：打卡提交总结的文字和公式推导的图片（20字以上，1张图片）

打卡截止日期：7/3

## 决策树的剪枝和连续值处理

**学习时长：**2天

**任务标题：**决策树的剪枝和连续值处理

**任务简介：**阅读西瓜书4.3/4.4

**任务详解：**

经过前两节的学习后，我们知道了，如何根据一个实际的数据集去构建出一颗树。但是，此时我们构建出的树可能存在一个问题，就是过拟合的问题，而解决这个问题就是4.3节所讲述的问题，请大家认真学习。在4.4节我们所接触的数据类型都是离散值，但是我们实际生产环境中，很多属性都是连续值，此时我们应该怎么处理连续数据呢，使其能够使用我们的树模型来解决问题，这就是4.4节要讲述的问题之一。同时4.4节还讲了一个重要问题，是我们生产环境经常面临的问题，就是当我们采集一些数据的时候，经常有一些数据的某些值是空的，这就严重影响树的构建，那么应该怎么解决呢？这就是4.4节需要讲述的另一个问题。简而言之，请大家掌握两个问题：第一，如何进行树的剪枝来防止过拟合，第二，对于含有空值的数据，此时应该怎么构建树。

**打卡：**

（1） 内容：决策树为什么要剪枝？有几种方法？简述一下，并分析其优缺点

（2） 形式：文字，最少50字；

打卡截止日期：7/5

## 学习sklearn包中决策树算法的使用

**学习时长：**1天

**任务标题：**学习sklearn包中决策树算法的使用

**任务简介：**实验-sklearn-user guide 1.10

**任务详解：**

了解sklearn包中的决策树算法是如何使用的，请大家认真看完1.10.1的内容，重点关注两部分内容：第一，怎么使用“DecisionTreeClassifier”这个api，第二，这个api中的参数是什么意思。其中在最末端有一个案例的链接，大家可选择看，不作为任务要求。同学们，对这部分学习的深入程度，依据个人代码能力而定。

**资料链接：**<https://scikit-learn.org/dev/modules/tree.html>

**打卡：**

（1） 内容：参照链接给的代码或者下图代码，实际动手敲一遍，使用sklearn的决策树接口实现鸢尾花分类，将运行结果截图。

<https://scikit-learn.org/dev/auto_examples/tree/plot_iris_dtc.html#sphx-glr-auto-examples-tree-plot-iris-dtc-py>

如果链接打不开，可直接看下方代码：





（2） 形式：打卡提交图片，最少1张；

打卡截止时间：7/7

## 本周学习任务简单总结

**学习时长：**1天

**任务简介：**简单回顾本周学到几个重要知识

**任务详解：**

每一周的学习任务都比较重，第一次学过之后特别容易忘，所以在周日及时做一个要点回顾，会让学习效率大大的提升，不会的知识也会越来越少

**打卡：**

（1）内容：请用文字描述，本周所学知识的重点，也可以思维导图、手写、电子版截图或者拍照均可，格式不限

（2） 形式：文字，最少20字；图片，0-9张；

打卡截止时间：7/7

## 参考答案

1. **常见决策树算法有哪些？它们的划分准则分别是什么，是否有缺陷？**

**解答：**

常见决策树算法有：ID3、C4.5、CART

划分准则分别是：信息增益、信息增益率、基尼指数

信息增益缺点：对可取值较多的属性有偏好，比如编号、日期

信息增益率缺点：对可取值较少的属性有偏好

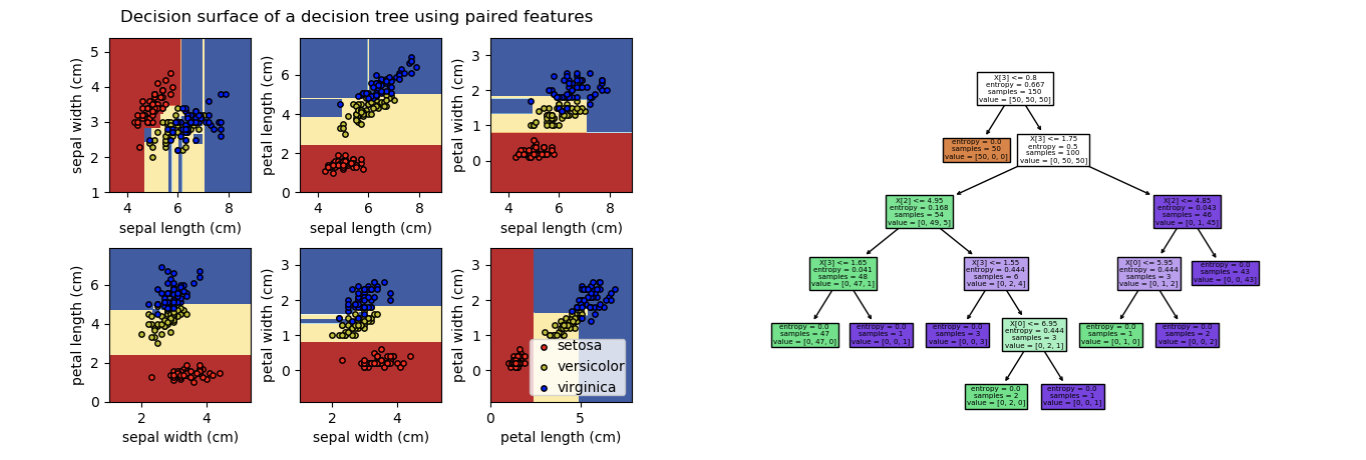
1. **决策树为什么要剪枝？有几种方法？简述一下，并分析其优缺点**

**解答：**

为了防止过拟合。有两种方法：预剪枝、后剪枝。

* 预剪枝：预剪枝的核心思想是在树中进行扩展之前，先计算当前的划分是否能带来模型泛化能力的提升，如果不能则不再生长子树。此时可能存在不同类别的样本同时存在结点中，按照多数投票原则判断该结点所属类别。优点：思想直接、算法简单、效率高等特点，适合解决大规模问题。缺点：有欠拟合风险。
* 后剪枝：后剪枝的核心思想是让算法生成一棵完全生长的决策树，然后从最底层向上计算是否剪枝。剪枝过程将子树删除，用一个叶子结点代替，该结点的类别同样按照多数投票的原则进行判断。同样，后剪枝也可以通过在测试集的准确率进行判断，如果剪枝后准确率有提升则进行剪枝。优点：相比于预剪枝，泛化能力强 。缺点：时间开销大

1. **使用sklearn的决策树接口实现鸢尾花分类，运行结果截图**



# 第4周

**本周学习内容介绍：**

本周我们要学习的主要内容为西瓜书的第6章支持向量机，在深度学习未火之前，支持向量机一直都是学术界和工业界的热点，这得益其优秀的性能，特别是面对数据规模比较小的时候，不同简单的逻辑回归，支持向量机能够构建出数据之间的非线性关系。同时，在机器学习的面试过程中，支持向量机一直都是热点问题，因为这部分能够对同学的数学功底进行考察，所以请大家认真对待这章的关键公式。学习完本章后，希望大家能够理清两个逻辑思维：一个是，支持向量机的原始公式是怎么由实际问题产生，这涉及到的是算法灵魂部分，希望大家认真思考；第二个就是数学推导，原始公式如何推导至其对偶形式的。

## 支持向量机原始模型的建立和求解

**学习时长：**7/8——7/9

**任务标题：**支持向量机原始模型的建立和求解

**任务简介：**学习西瓜书6.1/6.2，观看西瓜书的支持向量机公式推导视频

**任务详解：**

这部分主要讲两方面的内容：第一，支持向量机的原始公式是怎么由实际问题产生，这涉及的是算法的思想，也就是数学模型的建立；第二，当将数学模型建立后，接下来要做的就是对数学模型进行求解，这是这部分的主要内容，请大家掌握原始问题到对偶问题的数学推导过程。这部分的学习推导，大家可重点学习下李航老师的《统计学习方法》的内容，如果大家觉得公式推导吃力，推荐大家学习的《统计学习方法》训练营。

**视频：**

SVM-1

SVM-2

SVM-3

**打卡：**

（1） 内容：谈谈SVM原理是什么？

（2） 形式：打卡提交总结的文字和公式推导的图片（20字以上，1张图片）

打卡截止日期：7/9