**统计学习方法作业（第四期）**

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都要基于Python3，在Jupyter Notebook上完成。
3. 课程老师：Eddy老师

**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/StatisticalLearningMethod-Camp>

# 前言

教学内容完全依据《统计学习方法》一书，每一章的学习计划会详细列出本章需要学习的内容和不做学习要求的内容。

每章会有一个导读视频和两个重难点讲解视频（第三章和第五章内容比较容易理解，第三章没有重难点讲解视频，第五章只有一个重难点讲解视频），重难点讲解视频涉及模型理解、理论推导、算法实现等内容。书中共有10个算法，每个算法会在学习当天提供一个示例代码，同学们先自行理解，助教会在每周周末讲解作业和代码。

资料领取：

《统计学习方法》电子书请在公众号深度之眼后台直接回复关键词【统计】，即可领取。

# 第1周

## 学习第1章统计学习方法概论

**任务简介：**

学习第1章统计学习方法概论，理解统计学习方法的一些基本概念。

**详细说明：**

第1章是对统计学习中基本思想、基本概念以及常见问题类型的介绍。其中涉及的一些特定的技术不用深究。需要重点理解的内容是模型过拟合的含义、模型泛华能力两部分，另外通过习题熟悉一下极大似然估计和贝叶斯估计两个估计方法。

**学习目标：**

1. 观看绪论视频，了解训练营学习计划。
2. 观看导读视频
3. 理解“本章概要”的5点内容。
4. 理解模型过拟合产生的原因以及造成的影响（对应书籍中的第一章第4节）
5. 理解机器学习的评价标准：模型的泛化能力（对应书籍中的第一章第6节）
6. 熟悉极大似然估计和贝叶斯估计基本思想和求解方法（对应书籍中的习题1、习题2）。

**作业内容：**

作业1：

推导下述正态分布均值的极大似然估计和贝叶斯估计。数据来自正态分布，其中已知。

（1）根据样本写出的极大似然估计。

（2）假设的先验分别是正态分布，根据样本写出的贝叶斯估计。

作业答案在本周日公布，助教会进行视频讲解。

**打卡要求：**提交图片一张。

## 学习第2章感知机

**任务简介：**

学习第2章感知机，理解感知机模型解决的问题，模型形式、学习策略和求解算法。

**详细说明：**

第2章讲了在数据线性可分的情况下的感知机模型。通过阅读第1节，理解感知机模型的基本思想和模型形式；通过阅读第2节了解感知机模型采用的损失函数的形式及含义；第3节描述了感知机模型对应的优化问题的原始形式和对偶形式，请大家学习原始形式对应的随机梯度算法及算法的收敛性，对偶形式不做要求。

**学习目标：**

1. 导读视频
2. 掌握感知机的模型形式、损失函数及对应的优化问题
3. 掌握随机梯度下降算法原理
4. 理解感知机模型中随机梯度算法的收敛性。

**作业内容：**

作业2：

1. 思考感知机模型假设空间是什么？模型复杂度体现在哪里？打卡进行文字说明。
2. 已知训练数据集D，其正实例点是，负实例点是：

（a）用python自编程实现感知机模型，对训练数据集进行分类，并对比误分类点选择次序不同对最终结果的影响。可采用函数式编程或面向对象的编程。

（b）试调用sklearn.linear\_model的Perceptron模块，对训练数据集进行分类，并对比不同学习率h对模型学习速度及结果的影响。

（c）附加题：对比传统感知机算法及其对偶形式的运行速度。

打卡代码运行结果的截图

作业答案及代码讲解在本周日公布，助教会进行视频讲解。

**打卡要求：**文字20字，图片1张

打卡截止提交日期：2019/5/24

## 学习第3章k近邻

**任务简介：**

学习第3章k近邻，学习k近邻算法在分类问题中的应用，理解k近邻法的三要素及模型对应的损失函数。

**详细说明：**

第3章讲了如何用k近邻算法进行分类。同学们通过学习第1节，理解k近邻算法的基本思想；通过学习第2节，掌握k近邻算法如何实现；第3节是一个搜索技术，关于如何对某一训练数据快速找到相邻的k个示例，这里不做学习要求。

**学习目标：**

1. 导读视频
2. 掌握k近邻算法的原理
3. 理解k近邻算法三要素及模型对应的损失函数
4. 掌握k近邻算法在分类问题上的求解过程

**作业内容：**

作业3：

1. 思考k近邻算法的模型复杂度体现在哪里？什么情况下会造成过拟合？打卡进行文字说明。
2. 给定一个二维空间的数据集，试基于欧氏距离，找到数据点的最近邻，并对S点进行分类预测。

（a）用“线性扫描”算法自编程实现。

（b）试调用sklearn.neighbors的KNeighborsClassifier模块，对S点进行分类预测，并对比近邻数k取值不同，对分类预测结果的影响。

（c）思考题：思考“线性扫描”算法和“kd树”算法的时间复杂度。

打卡代码运行结果的截图。

作业答案及代码讲解在下周日公布，助教会进行视频讲解。

**打卡要求：**文字20字，图片1张

打卡截止时间：2019/5/25

## 第1周作业讲解及代码公布

看助教第1章作业讲解视频，输出笔记。

看助教第2章作业讲解视频，输出笔记。

第2章附代码下载

链接：<https://pan.baidu.com/s/16RB9Z6V3sJiqZQvoSFxh1g>

提取码：xjua

（第二章作业讲解视频前加一句：自编程视频中y\_train写法没有问题，当时搞错了）

**任务简介：**回看助教录制的前3章作业讲解视频

代码链接：

（1）Chap3—k近邻暴力算法优化&kd树实现代码

链接：<https://pan.baidu.com/s/1a5ysKpip8YOe__LGTpywDw>

提取码：a90s

（2）Chap3—KNN自编程KNN-Sklearn代码

链接：<https://pan.baidu.com/s/1Susz2S33FNUKvVQ3dIg_dg>

提取码：4ibz

**作业：**总结第1至3章内容，形成笔记。

**打卡要求：**图片1张

作业截止提交日期：2019/5/26

# 第2周

## 学习第4章朴素贝叶斯法

**任务简介：**

学习第4章朴素贝叶斯法，理解朴素贝叶斯法解决的问题，模型假设、损失函数和估计方法。

**学习时长：**2019/5/27—2019/5/28

**详细说明：**

第4章介绍的朴素贝叶斯法依然适用于分类问题。通过学习第1节，需要了解朴素贝叶斯模型的基本思想和模型假设，了解后验概率最大化对应的损失函数；通过学习第2节，需要掌握朴素贝叶斯中，极大似然估计的求解方法及对应的算法，理解引入贝叶斯估计的原因及贝叶斯估计的求解方法。

**学习目标：**

1. 导读视频。
2. 理解朴素贝叶斯模型的模型假设。
3. 理解后验概率最大化与期望损失最小化的关系。
4. 掌握极大似然估计的求解过程。
5. 掌握贝叶斯估计的求解过程。

**作业内容：**

作业4

1. 取，试由下表的训练数据，利用先验概率的贝叶斯估计确定的类标记y。表中为特征，取值的集合分别为，Y为类标记，。

训练数据

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| X(1) | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| X(2) | S | M | M | S | S | S | M | M | L | L | L | M | M | L | L |
| Y | -1 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 |

（2）（a）自编程实现朴素贝叶斯算法，对上述表格中的训练数据进行分类。

（b）试分别调用sklearn.naive\_bayes的GaussianNB、BernoulliNB、MultinomialNB模块，对上述表格中训练数据进行分类。

打卡代码运行结果的截图。

作业答案及代码讲解在本周日公布，助教会进行视频讲解。

**打卡要求：**图片2张

## 学习第5章决策树

**任务简介：**掌握决策树模型的思想和算法。

**详细说明：**

第5章介绍的决策树模型即可以解决分类问题也可以解决回归问题，本书重点介绍分类问题。通过学习第1节，了解决策树模型的基本思想；通过学习第2节，了解选择分类特征的两个准则：信息增益和信息增益比；通过学习第3节，理解两个准则下对应的两种决策树算法；通过学习第4节，了解决策树模型中如何通过剪枝控制模型复杂度；通过学习第5节，掌握基尼系数的计算方法和CART算法，其中回归问题不做学习要求。本章的难点是如何理解信息增益和基尼系数以及它们的计算。

**学习目标：**

1. 导读视频
2. 理解信息增益、基尼系数的含义和计算方法
3. 掌握ID3算法
4. 了解决策树剪枝的目的和剪枝准则
5. 掌握CART生成算法。

**作业内容：**

作业5：

（1）证明CART剪枝算法中，当确定时，存在唯一的最小子树使损失函数最小。

（2）尝试调用sklearn.tree.DecisionTreeClassifier模块，训练数据集采用课本例题5.1的数据，判断是否应该批准下列人员的贷款申请，打卡代码运行结果的截图。

A={青年，否，是，一般}

B={中年，是，否，好}

C={老年，否，是，一般}

作业答案及代码讲解在本周日公布，助教会进行视频讲解。

**打卡要求：**图片2张

作业截止提交日期：2019/5/31

## 第2周作业讲解及代码公布

**任务简介：**回看助教录制的前3章作业讲解视频，以及下方的第4章作业讲解视频，对这两周的任务做一个阶段性总结，参加今晚的直播答疑。

代码链接：

（1）Chap4朴素贝叶斯自编程实现代码

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

提取码：3gim

学习第5章作业讲解视频

代码链接

（2）Chap5决策树自编程实现代码

链接：<https://pan.baidu.com/s/1QVI_mUIJ1DL1541MWrGm3w>

提取码：fmfr

作业：总结第4至5章内容，形成笔记。

打卡要求：（图片1张）

作业截止提交日期：2019/6/02

# 第3周

## 学习第6章Logistic回归与最大熵模

**任务简介：**

学习第6章Logistic回归与最大熵模型，理解Logistic回归的模型形式和求解方法，了解最大熵模型的思想和求解方法。

**详细说明：**

第6章介绍的Logistic回归与最大熵模型都属于对数线性模型，都用来解决分类问题。通过学习第1节，需要掌握Logistic回归的模型形式和似然函数；通过学习第2节，需要理解最大熵模型的思想和求解方法；第3节不做学习要求。

**学习目标：**

1. 导读视频
2. 掌握二项Logistic和多项Logistic模型的模型形式和似然函数
3. 掌握二项Logistic求解中的梯度下降法。
4. 理解最大熵模型的思想，了解拉格朗日对偶性。
5. 理解最大熵模型中的改进的迭代尺度算法。

**作业内容：**

作业6：

1. 已知训练数据集D，其正实例点是，负实例点是。

（a）用python自编程实现逻辑斯缔回归模型，并对点进行分类。

（b）试调用sklearn.linear\_model的LogisticRegression模块，对点进行分类，尝试改变参数，选择不同算法，如梯度下降法和拟牛顿法。

打卡代码运行结果的截图。

作业答案及代码讲解在本周日公布，助教会进行视频讲解。

**打卡要求：**图片2张

作业截止提交日期：2019/6/04

## 学习第7章支持向量机

**任务简介：**

理解线性可分支持向量机、线性支持向量机和非线性支持向量机。

**详细说明：**

第7章介绍了支持向量机如何用于二分类问题。通过学习第1节，掌握线性可分支持向量机与感知机的区别，了解对应的凸优化问题、对偶问题及相应的算法；通过学习第2节，掌握软间隔最大化对应的优化问题、对偶问题及相应的算法，其中合页损失函数不做学习要求；通过学习第3节，了解核函数在非线性支持向量机中的应用，其中7.3.2和7.3.3不做学习要求；第4节介绍了训练样本较大时的序列最小最优化算法，了解即可，不做学习要求。

**学习目标：**

1. 导读视频
2. 理解线性可分支持向量机硬间隔最大化的思想、对应的优化问题、对偶问题和相应算法
3. 理解硬间隔最大化解的存在唯一性
4. 理解线性支持向量机软间隔最大化的思想、对应的优化问题、对偶问题和相应算法
5. 了解核函数在非线性支持向量机中的应用和对应的算法
6. 了解序列最小最优化算法的基本思想。

**作业内容：**

作业7：

1. 完成习题7.2：已知正例点，负例点，试求最大间隔分离超平面和分类决策函数，并在图上画出分离超平面、间隔边界及支持向量。
2. 完成习题7.3：线性支持向量机还可以定义成以下形式：

试求其对偶形式。

1. 试调用sklearn.svm中的SVC模块求解习题7.2，尝试改变参数，如C，kernel，比较结果。

作业答案及代码讲解在本周日公布，助教会进行视频讲解

**打卡要求：**图片2张

作业截止提交日期：2019/06/07

## 第3周作业讲解及代码公布

任务简介：学习第6章作业讲解视频，对这周的任务做一个简单总结。

代码链接：

链接：<https://pan.baidu.com/s/1di7xWX4Rmq2jw-RnKcee2g>

提取码：yph7

任务简介：学习第7章作业讲解视频，对这周的任务做一个简单总结。

代码链接：

链接：<https://pan.baidu.com/s/1vwGHz_7RsiJ6u9bnHhOj-g>

提取码：tpj7

作业：总结第6至7章内容，形成笔记。

**打卡要求：**图片1张

作业截止提交日期：2019/6/09

# 第4周

## 学习第8章提升方法

**任务简介：**

学习第8章提升方法，学习AdaBoost算法和提升树。

**详细说明：**

第8章介绍了AdaBoost算法和提升树。在第1节的学习中，通过例题8.1掌握AdaBoost算法的求解过程；通过学习第2节，了解AdaBoost算法训练误差的性质；通过学习第3节，理解AdaBoost算法与前向分步算法的关系；在第4节中，通过例题8.2掌握回归问题的提升树算法。

**学习目标：**

1. 导读视频。
2. 掌握AdaBoost算法的求解过程。
3. 理解AdaBoost算法的训练误差。
4. 理解AdaBoost算法是前向分步算法的一个特例。
5. 掌握回归问题的提升树算法。

**作业内容：**

作业8：

（1）自编程实现课本例题8.1

（2）调用sklearn.ensemble.AdaBoostClassifier对例题8.1进行实现

打卡代码运行结果的截图

作业答案及代码讲解在下周公布，助教会进行视频讲解。

**打卡要求：**图片1张

打卡截止提交日期：2019/6/11

## 学习第9章EM算法及推广

**任务简介：**

学习第9章EM算法及推广，理解EM算法的思想和E步、M步的求解过程。

**详细说明：**

第9章介绍了EM算法，EM算法用于含有隐变量的概率模型的参数估计。EM算法不是一个具体的分类或回归算法，而是广泛用于含有隐变量的模型的求解问题。通过学习第1节，掌握EM算法E步和M步的求解过程，了解EM算法和极大似然估计的关系；第2节讨论了EM算法的收敛性，需要了解定理的内容，证明过程不做学习要求；通过学习第3节，需要掌握在高斯混合模型中如何用EM算法估计参数；第4节不做学习要求。

**学习目标：**

1. 导读视频。
2. 通过例题9.1掌握EM算法E步和M步的求解过程。
3. 了解EM算法求解如何用从最大化观测数据似然函数导出。
4. 掌握高斯混合模型如何用EM算法估计参数。

**作业内容：**

作业9：

1. 完成习题9.3：已知观测数据

-67，-48，6，8，14，16，23，24，28，29，41，49，56，60，75

试估计两个分量的高斯混合模型的5个参数。

1. 试用自编程的方式求解习题9.3，打卡代码运行结果的截图。

作业答案及代码讲解在下周公布，助教会进行视频讲解。

**打卡要求：**图片2张

打卡截止提交日期：2019/6/14

## 第4周作业讲解及代码公布

**任务简介：**观看助教录制的Chap8作业讲解视频。

代码附件：

链接：<https://pan.baidu.com/s/1PA7Zfgfi8AElDqTjsW3v6w>

提取码：uq0k

**任务简介：**观看助教录制的Chap9作业讲解视频。

代码附件：

链接：<https://pan.baidu.com/s/1_-D83t-7uT7MKNpTdu6tFg>

提取码：qjt5

**作业内容：**总结第8至9章内容，形成笔记。

**打卡要求：**（图片1张）

作业截止提交日期：2019/6/16

# 第5周

## 学习第10章隐马尔科夫模型

**任务简介：**

学习隐马尔科夫模型中的三个基本问题。

**详细说明：**

第10章介绍了隐马尔科夫模型。通过学习第1节，了解隐马尔科夫模型的三要素和三个基本问题；通过学习第2节，理解隐马尔可夫模型概率计算中的前向算法和后向算法；通过学习第3节，理解隐马尔科夫模型参数的监督学习算法和非监督学习算法；在第4节中，通过例题10.3掌握维特比算法。

**学习目标：**

1. 导读视频
2. 了解隐马尔科夫模型的三要素和三个基本问题。
3. 通过例题10.2，掌握隐马尔可夫模型概率计算中的向前算法。
4. 理解隐马尔科夫模型参数的非监督学习算法。
5. 通过例题10.3，理解维特比算法。

**作业内容：**

作业10：

1. 给定盒子和求组成的隐马尔科夫模型，其中：

设，

1. 试用前向、后向算法计算
2. 试用维特比算法求最优路径
3. 尝试用自编程的方式完成上述题目。

打卡代码运行结果截图。

作业答案及代码讲解在本周日公布，助教会进行视频讲解。

**打卡要求：**图片2张

## 学习第11章条件随机场

**任务简介：**理解条件随机场和相应的算法。

**详细说明：**

第11章条件随机场，该算法可以用于标注问题。通过学习第1节，了解概率无向图的定义和因子分解形式；通过学习第2节，了解线性链条件随机场的定义及三种形式；通过学习第3节，理解条件概率和期望的前向-后向算法；通过学习第4节，掌握两个学习算法优化的目标函数，求解过程不做学习要求；通过学习第5节，掌握条件随机场的预测算法。

**学习目标：**

（1）导读视频

（2）理解概率无向图。

（3）通过例题11.1和例题11.2，掌握线性链条件随机场模型参数形式和矩阵形式。

（4）掌握条件概率和期望的前向-后向算法。

（5）理解拟牛顿法。

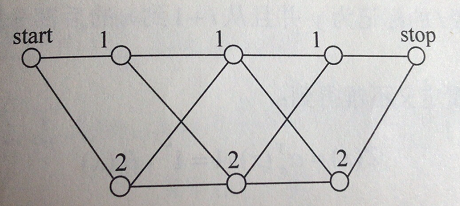
（6）理解预测的维特比算法。

**作业内容：**

作业11：

（1）完成习题11.4：

参考下图的状态路径图，假设随机矩阵分别是：



求以start=2为起点stop=2为终点的所有路径的状态序列y的概率及概率最大的状态序列。

（2）尝试用自编程的方式求解上述问题，打卡代码运行结果的截图。

**打卡要求：**图片2张

## 第5周作业讲解及代码公布

**任务简介：**观看助教录制的Chap10作业讲解视频。

代码附件：

链接：<https://pan.baidu.com/s/1kYkT_Ide8ycOODqbT0BqBA>

提取码：23g6

**任务简介：**观看助教录制的Chap11作业讲解视频。

代码附件：

链接：<https://pan.baidu.com/s/1jac4FYP33mai5s19LvgiGg>

提取码：28dy

**作业内容：**总结第10至11章内容，形成笔记。

**打卡要求：**图片1张