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。

1 第 1 周

1.1 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$x_1,\cdots,x_n$来自正态分布$\mathrm{N}(\mu,\sigma^2)$，其中σ^2已知。

（1）根据样本$x_1,\cdots,x_n$写出μ的极大似然估计。

（2）假设μ的先验分别是正态分布$\mathrm{N}(0,\tau^2)$，根据样本$x_1,\cdots,x_n$写出μ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。

1.2 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$x_1 = (3,3)^T, x_2 = (4,3)^T$，负实例点是$x_3 = (1,1)^T$：

（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

1.3 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 $T = \{正实例: (5,4), (9,6), (4,7); 负实例: (2,3), (8,1), (7,2)\}$，试基于欧氏距离，找到数据点$S(5,3)$的最近邻$(k = 1)$，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

1.4 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

2 第 2 周

2.1 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$\lambda = 0.2$，试由下表的训练数据，利用先验概率的贝叶斯估计确定$x = (2, S)^T$的类标记 y。表中$X^{(1)}, X^{(2)}$为特征，取值的集合分别为$A_1 = \{1,2,3\}, A_2 = \{S, M, L\}$，Y 为类标记，$Y \in C = \{1, -1\}$。

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

2.2 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α确定时，存在唯一的最小子树T_α使损失函数$C_\alpha(T)$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

2.3 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

3 第 3 周

3.1 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$(\mathrm{Y}=1)$是$x_1=(3,3,3)^T, x_2=(4,3,2)^T, x_3=(2,1,2)^T$，负实例点$(\mathrm{Y}=1)$是$x_4=(1,1,1)^T, x_5=(-1,0,1)^T, x_6=(2,-2,1)^T$。

（a）用 python 自编程实现逻辑斯缔回归模型，并对点$(1,2,-2)^T$进行分类。

（b）试调用 sklearn.linear_model 的 LogisticRegression 模块，对点$(1,2,-2)^T$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

3.2 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$x_1=(1,2)^T, x_2=(2,3)^T, x_3=(3,3)^T$，负例点$x_4=(2,1)^T, x_5=(3,2)^T$，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：

$$\min_{w,b,\varepsilon} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{N}\varepsilon_i^2$$

$$\text{s.t. } y_i(wx_i+b) \geq 1-\varepsilon_i \quad i=1,2,\cdots,N$$

$$\varepsilon_i \geq 0 \quad i=1,2,\cdots,N$$

试求其对偶形式。

（3） 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

3.3 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_7RsiJ6u9bnHh0j-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

4 第 4 周

4.1 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

4.2 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。

（2） 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

4.3 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

5 第 5 周

5.1 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$\lambda = (A, B, \pi)$，其中：

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}, B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}, \pi = (0.2, 0.4, 0.4)^T$$

设$T = 4$，$O = (红, 白, 红, 白)$

（a） 试用前向、后向算法计算$P(O/\lambda)$

（b） 试用维特比算法求最优路径$I^* = (i_1^*, i_2^*, i_3^*, i_4^*)$

（2） 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

5.2 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：

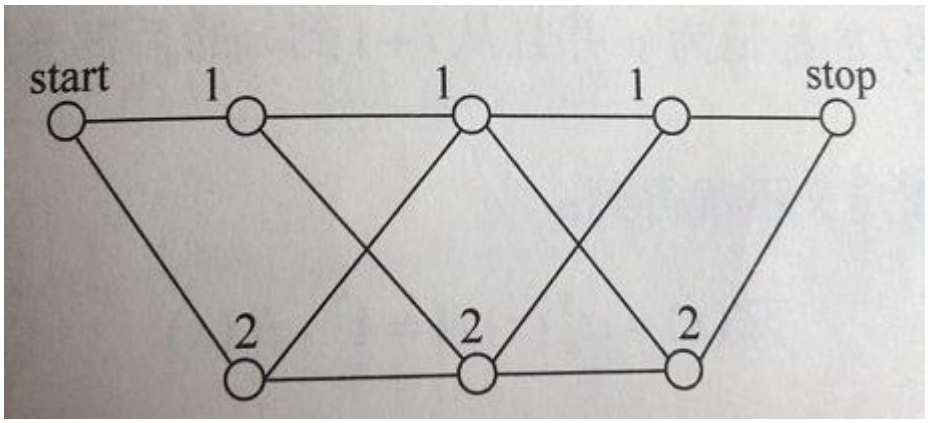
作业 11：

（1）完成习题 11.4：

参考下图的状态路径图，假设随机矩阵$M_1(x)$、$M_2(x)$、$M_3(x)$、$M_4(x)$分别是：

$$M_1(x)=\begin{bmatrix}0 & 0\\ 0.5 & 0.5\end{bmatrix}\quad M_2(x)=\begin{bmatrix}0.3 & 0.7\\ 0.7 & 0.3\end{bmatrix}$$

$$M_3(x)=\begin{bmatrix}0.5 & 0.5\\ 0.6 & 0.4\end{bmatrix}\quad M_4(x)=\begin{bmatrix}0 & 1\\ 0 & 1\end{bmatrix}$$



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

5.3 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_Ide8ycOODqbTOBqBA

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