统计学习方法

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，在 Jupyter Notebook 上完成。
3. 课程老师：Miss.K 老师
4. 课程资料地址：

   https://pan.baidu.com/s/1PZH4FblVvrxGai0PfJPYYw

   提取码：hisk

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

第 1 周-1（ROC 曲线、L1/L2 范数）

任务名称：

书籍阅读：学习第一章内容，参考配套的 PPT

任务详解：对机器学习有大致了解，需要重点理解的部分是 1.4（误差及过拟合）,1.5（正则化和交叉验证）,1.8（分类问题的判别）

参考资料：

【统计学习方法之基础篇.PPT】以及 PPT 中推荐的视频和书籍

【第 1 章 统计学习方法概论.pdf】

作业

1. 理解 L1,L2 范式（主要概念和区别）
2. 理解 ROC 曲线，并解释代码（scikit-learn 官方代码）。【ROC curve.ipynb】

参考答案

1. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week1/MyHomeWork/homework_1.1.md
2. 见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week1/MyHomeWork/homework_1.2_ROC%20curve.ipynb

第 1 周-2（感知机、KNN）

任务名称：

观看录播视频并且理解感知机、KNN、KD Tree 的算法原理

任务详解：

1. 通过录播视频需要了解简单线性分类器的生成原理，并且尝试自己实现一个简单的分类器

2. KNN 主要理解算法原理，不要求自己实现代码，但是需要对代码进行理解和解释说明
3. 对于 KD Tree 需要了解算法 。

参考材料：【第 2 章 感知机.pdf】、【第 3 章 k 近邻法.pdf】

作业

1. 生成两个由 1000 个样本组成的二变量高斯分布，分别服从$m_1 = [0,2]^T, m_2 = [1.5,0]^T$，且具有同样的协方差矩阵$\mathrm{C} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$，计算两个分布的贝叶斯最优分类边界并画图。【simple linear perceptron.ipynb】
2. 理解 KNN 算法原理，并解释代码。【KNN.ipynb】

参考答案

1. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week1/MyHomeWork/homework_2.1_simple%20linear%20perceptron.ipynb
2. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week1/MyHomeWork/homework_2.2_KNN.ipynb

第 2 周-1（决策树）

任务名称：

1. 书籍阅读：学习第五章内容，参考配套的 PPT
2. 观看录播视频理解算法

作业

1. 根据所给的训练数据集，利用信息增益和信息增益比分别生成决策树；

day	outlook	temp	humidity	wind	tennis
1	sun	hot	high	weak	no
2	sun	hot	high	strong	no
3	cloud	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	norm	weak	yes
6	rain	cool	norm	strong	no
7	cloud	cool	norm	strong	yes
8	sun	mild	high	weak	no
9	sun	cool	norm	weak	yes
10	rain	mild	norm	weak	yes
11	sun	mild	norm	strong	yes
12	cloud	mild	high	strong	yes
13	cloud	hot	norm	weak	yes
14	rain	mild	high	strong	no

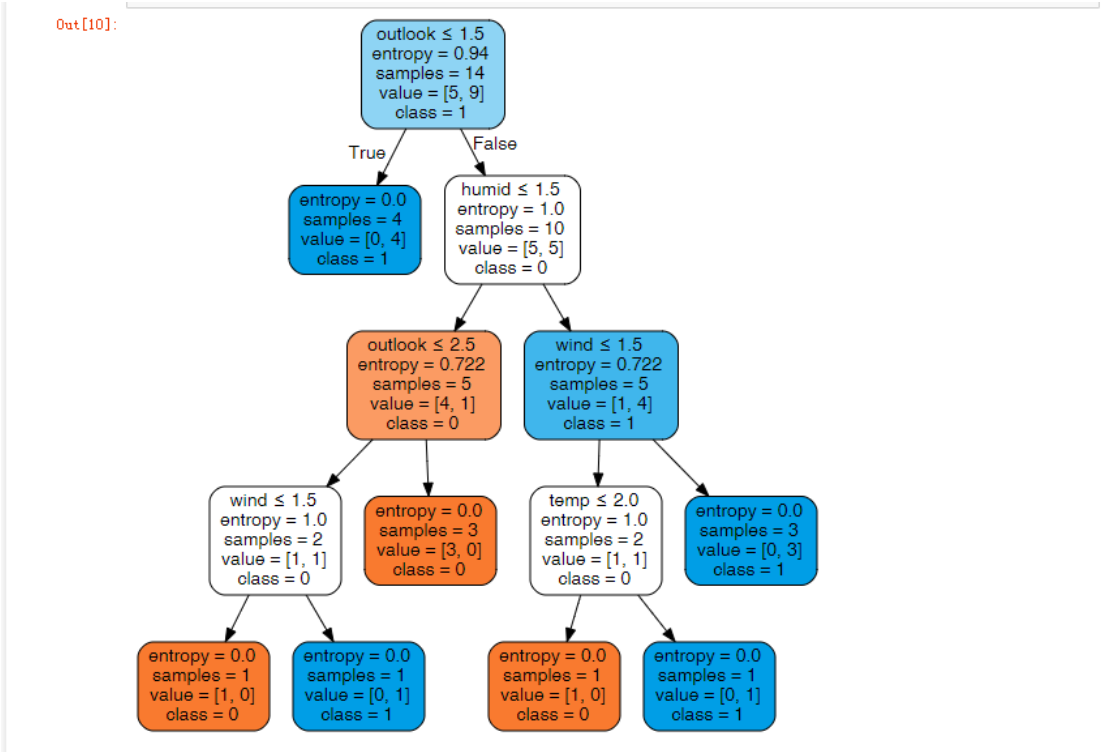
2. 理解 decision tree 代码
3. Sklearn DT 参数理解

(https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier)

4. 附加作业：利用上面学习的代码对项目进行分类

参考答案

1. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week2/MyHomeWork/homework_1.1.ipynb



2. 详见 github

https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week2/MyHomeWork/homework_1.2_decision%20tree.ipynb

3. Sklearn DT 参数理解

见官网，可以通过 criterion='entropy'来指定用什么方法来做评价标准。其中剪枝算法需要通过以下参数进行调参：

max_depth=None，树的最大深度

min_samples_split=2，分裂点的样本个数

min_samples_leaf =1，叶子节点的样本个数

max_leaf_nodes=None，最大的叶子节点数

4. 详见 github

https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week2/MyHomeWork/homework_1.4.ipynb

最终会得到 score 的值为 0.77

第 2 周-2（Adaboost 算法）

任务名称：

1. 书籍阅读：学习 8.1,8.2,8.3 内容，参考配套的 PPT

2. 观看录播视频理解算法

作业

1. 理解 Adaboost 代码

2. 附加作业：利用上面学习的代码对项目进行分类

参考答案

1. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week2/MyHomeWork/homework_2.1_Adaboost.ipynb
2. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week2/MyHomeWork/homework_2.2.ipynb

第 2 周-3（提升树算法）

任务名称：

1. 书籍阅读：学习 8.4 内容，参考配套的 PPT
2. 观看录播视频理解算法

作业

1. sklearn GBDT 参数理解(https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html#sklearn.ensemble.GradientBoostingClassifier)
2. 比较 ADABOOST 和 GBDT 算法

参考答案

1. 详见 github
https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week2/MyHomeWork/homework_3.1.ipynb
2. 和 AdaBoost 一样，Gradient Boosting 每次基于先前模型的表现选择一个表现一般的新模型并且进行调整。不同的是，AdaBoost 是通过提升错分数据点的权重来定位模型的不足，而 Gradient Boosting 是通过算梯度（gradient）来定位模型的不足。因此相比 AdaBoost, Gradient Boosting 可以使用更多种类的目标函数,而当目标函数是均方误差时，计算损失函数的负梯度值在当前模型的值即为残差。

第 3 周-1（朴素贝叶斯、逻辑斯蒂回归算法）

任务名称：

1. 书籍阅读：学习第四章内容及 6.1，参考配套的 PPT
2. 观看录播视频理解算法

作业

1. 理解朴素贝叶斯代码
2. 理解逻辑斯蒂回归模型代码
3. 尝试分类任务

参考答案

1. 详见 github

https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week3/MyHomeWork/homework_1.1_NaiveBayes.ipynb

2. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week3/MyHomeWork/homework_1.2_LogisticRegression.ipynb

3. 详见 github，采用朴素贝叶斯算法完成例 4.1 题目 https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week3/MyHomeWork/homework_1.3.ipynb

第 3 周-2（EM 算法）

任务名称：

1. 书籍阅读：学习 6.2+6.3+第九章，参考配套的 PPT
2. 观看录播视频理解算法

作业

1. 理解 EM 算法代码
2. 具体说明 E 步和 M 步的过程
3. 习题：P170 9.1 题 9.3 题

参考答案

1. 详见 github

https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week3/MyHomeWork/homework_2.1_EM.ipynb

2. 详见博客 https://blog.csdn.net/sinat_22594309/article/details/65629407

3. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week3/MyHomeWork/homework_2.3.ipynb

第 4 周-1（非线性 SVM 算法）

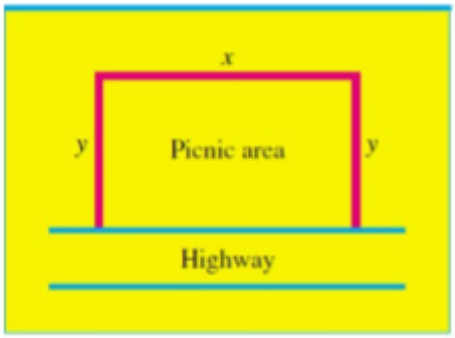
任务名称：

1. 书籍阅读：学习 7.3 和附录 C，参考配套的 PPT
2. 了解径向基、拉格朗日乘数法和 KKT 条件

作业

1. 完成下面两题截图提交

   （1）公路部门正计划在一条主要公路旁为驾车的人建立一个野餐区。这是长方形的，面积为五千平方米，在公路以外的三面用栅栏围起来。完成这项工作最少需要多少栅栏？



（2）一位编辑被给予 6 万英镑用于新书的开发和推广。一项实证研究发现，如果 x 千美元用于开发，y 千美元用于促销，那么大约$\mathrm{f}(\mathrm{x},\mathrm{y}) = 20x^{\frac{3}{2}}y$本书就会售出。编辑应该拨多少钱用于开发，多少钱用于促销，以便最大限度地提高销售?

2. 理解今天所说三个概念（我们会在下一关公布答案）

参考答案

1. （1）根据题意可以得到如下的 KKT 条件：

$$\min_{x,y} z = x + 2y \quad s.t.\ xy = 5000 \Rightarrow z = \frac{x^2+10000}{x}$$

可知就上式的 z 的导数，并令导数等于 0

$z' = \frac{x^2-10000}{x^2} = 0 \Rightarrow x = 100$ 根据约束条件可以得到 y=50

完成这项工作最少需要 200 米的栅栏

（2）根据题意可以得到如下 KKT 条件

$$\max_{x,y} f(x,y) = 20x^{\frac{3}{2}}y \quad s.t.\ x + y = 60 \Rightarrow f(x) = 20x^{\frac{3}{2}}(60 - x)$$

对 $f(x)$求导，并令导数等于 0

$f(x)' = x^{\frac{1}{2}}(1800 - 50x) = 0 \Rightarrow x = 36$ 可以得到 y=24

编辑应该拨 3.6 万用于开发，2.4 万用于促销，以便最大限度地提高销售，最大销售的书本量为 103680 本

2. （1）**径向基**

径向基函数是某种沿径向对称的标量函数,通常定义为样本到数据中心之间径向距离（通常是欧氏距离）的单调函数（由于距离是径向同性的）。

常见的径向基函数包括（定义$r = \|x - x_i\|$）：

- 高斯函数：$\phi(r) = e^{-(\varepsilon r)^2}$
- 多二次函数（multiquadric）：$\phi(r) = \frac{1}{1+(\varepsilon r)^2}$
- 逆二次函数（inverse quadratic）：$\phi(r) = \sqrt{1 + (\varepsilon r)^2}$
- 逆多二次函数（inverse multiquadric）：$\phi(r) = \frac{1}{1+(\varepsilon r)^2}$
- 多重调和样条（polyharmonic spline）：

$\phi(r) = r^k, k = 1,3,5, \ldots$

$\phi(r) = r^k \ln(r), k = 2,4,6, \ldots$

- 薄板样条（thin plate spline，为多重调和样条的特例）：

$\phi(r) = r^2 \ln(r)$

（2）**拉格朗日乘数法**

在数学最优问题中,拉格朗日乘数法(以数学家约瑟夫·路易斯·拉格朗日命名)是一种寻找变量受一个或多个条件所限制的多元函数的极值的方法。这种方法将一个有n个变量与k个约束条件的最优化问题转换为一个有$n + k$个变量的方程组的极值问题，其变量不受任何约束。这种方法引入了一种新的标量未知数，即拉格朗日乘数：约束

方程的梯度（gradient）的线性组合里每个向量的系数。此方法的证明牵涉到偏微分，全微分或链法，从而找到能让设出的隐函数的微分为零的未知数的值。

参考资料：

https://www.cnblogs.com/sddai/p/5728195.html

https://blog.csdn.net/the_lastest/article/details/78136692

（3）KKT 条件

关于 KKT 条件这一个理解，笔者这里直接借用链接中的进行理解。

参考资料：

https://blog.csdn.net/qq_32742009/article/details/81411151

https://blog.csdn.net/u014675538/article/details/77509342

第 4 周-2（线性可分 SVM、线性 SVM 算法）

任务名称：

1. 书籍阅读：学习 7.1,7.2，参考配套的 PPT
2. 观看录播视频理解算法

作业

1. 理解 SVM 代码
2. 习题：P134 7.2 题
3. 尝试分类任务

参考答案

1. 详见 github

https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week4/MyHomeWork/homework_2.1_SVM.ipynb

2. 详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week4/MyHomeWork/homework_2.2.ipynb

第 5 周-1（隐马尔科夫模型）

任务名称：

1. 书籍阅读：学习第十章，参考配套的 PPT
2. 观看录播视频理解算法

作业

1. 理解算法
2. 理解代码

参考答案

1. 代码理解详见 github https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week5/MyHomeWork/homework_1.2_HMM.ipynb

第 5 周-2（条件随机场）

任务名称：

1. 书籍阅读：学习第十一章，参考配套的 PPT
2. 观看录播视频理解算法
3. 理解条件随机场算法原理

作业

1. 理解算法
2. 习题：P209 11.3 题 11.4 题

参考答案

1. 代码理解详见 github
   https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week5/MyHomeWork/homework_2.1_CRF.ipynb
2. 详见 github
   https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week5/MyHomeWork/homework_2.2.ipynb