**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

1. 补充资料地址：

<https://pan.baidu.com/s/1Z3O-PRnZbhHAwpBquAmVJg>

提取码：65su

**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>

1. 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，且具有同样的协方差矩阵，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>

1. 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 |

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

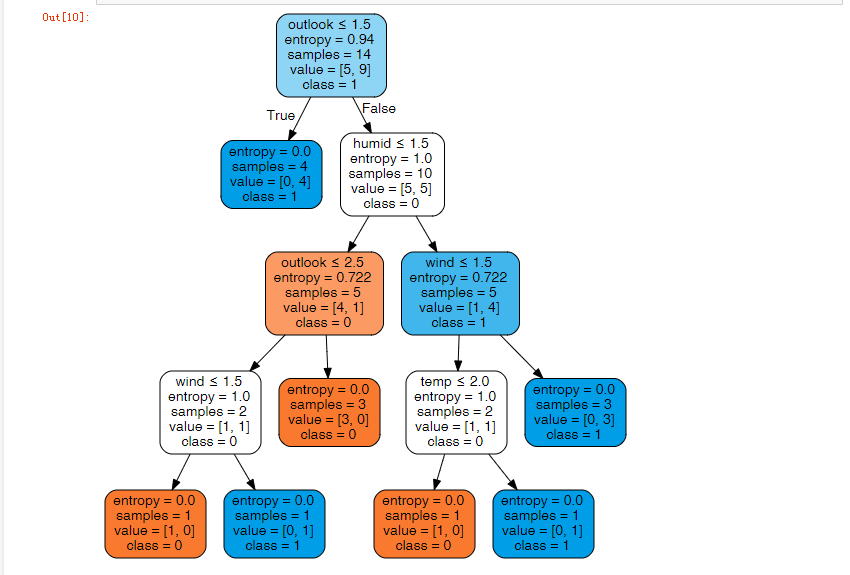
(<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier>)

1. 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>



1. 详见github

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

1. 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

1. 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>

1. 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>

1. 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>

1. 详见github

[https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week3/MyHomeWork/homework\_1.2\_LogisticRegression.ipynb](https://github.com/Relph1119/StatisticalLearningMethod-Camp/blob/master/Week3/MyHomeWork/homework_1.1_NaiveBayes.ipynb)

1. 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>

1. 详见博客

<https://blog.csdn.net/sinat_22594309/article/details/65629407>

1. 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）

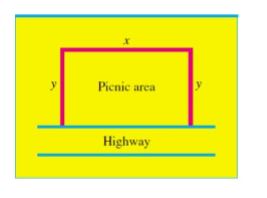
任务名称：

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?

1. 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：

可知就上式的z的导数，并令导数等于0

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

对 求导，并令导数等于0

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

1. （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）：

* [高斯函数](https://baike.baidu.com/item/%E9%AB%98%E6%96%AF%E5%87%BD%E6%95%B0)：
* 多二次函数（multiquadric）：
* 逆二次函数（inverse quadratic）：
* 逆多二次函数（inverse multiquadric）：
* 多重调和样条（polyharmonic spline）：

* 薄板样条（thin plate spline，为多重调和样条的特例）：

（2）**拉格朗日乘数法**

在数学最优[问题](https://baike.baidu.com/item/%E9%97%AE%E9%A2%98/1067365)中，拉格朗日乘数法（以数学家[约瑟夫·路易斯·拉格朗日](https://baike.baidu.com/item/%E7%BA%A6%E7%91%9F%E5%A4%AB%C2%B7%E8%B7%AF%E6%98%93%E6%96%AF%C2%B7%E6%8B%89%E6%A0%BC%E6%9C%97%E6%97%A5/7070424)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](https://baike.baidu.com/item/%E5%A4%9A%E5%85%83%E5%87%BD%E6%95%B0/2498728)的[极值](https://baike.baidu.com/item/%E6%9E%81%E5%80%BC/5330918)的方法。这种方法将一个有个变量与个[约束条件](https://baike.baidu.com/item/%E7%BA%A6%E6%9D%9F%E6%9D%A1%E4%BB%B6/1046571)的[最优化问题](https://baike.baidu.com/item/%E6%9C%80%E4%BC%98%E5%8C%96%E9%97%AE%E9%A2%98/9632567)转换为一个有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](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7%E7%BB%84%E5%90%88/8664061)里每个向量的系数。 此方法的证明牵涉到偏微分，[全微分](https://baike.baidu.com/item/%E5%85%A8%E5%BE%AE%E5%88%86/8155184)或链法，从而找到能让设出的[隐函数](https://baike.baidu.com/item/%E9%9A%90%E5%87%BD%E6%95%B0/6988840)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>

1. 详见github

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