0813

1. 微积分和线性代数
2. 基本的Python能力（廖雪峰的Python教程，到面向对象编程一节即可）
3. 基本的Python数据处理能力（pandas和numpy）

pandas官方文档

Github上的pandas练习项目：<https://www.kesci.com>/home/project/5a8afe517f2d695222327e14

numpy官方文档

numpy菜鸟教程

常用机器学习算法：

* 线性模型原理和实现
* 集成学习模型（随机森林、GBDT、xgboost、lightgbm）原理和实现
* 深度学习模型、Tensorflow的使用
* 特征选择、模型调参与模型融合
* 不包含支持向量机（算法复杂度较高，在比较大的数据集上很难运行）

比赛平台：天池、Kaggle

腾讯广告赛、京东赛 – 数据量非常大、结果比较真实。

微信公众号：麻婆豆腐AI

Github：https://iphysresearch.github.io/DataSciComp?sub=PF,AC,DM

比赛解决方案汇总Github：<https://github.com/Smilexuhc/Data-Competition-TopSolution>

机器学习的目标：以机器算法替代人工劳动，以自动地完成预测或决策。

机器学习：从大量的数据中学习规律，需要依赖于统计学工具。大部分机器学习问题都可以转化为预测问题。

计算机程序的逻辑仍然由人类编写和实现，但机器学习中的规律是由算法从大量数据中自动学习的。

线性回归：性能上的优越性、良好的可解释性。

系数b1：在其他条件相同的条件下，x1每增加1，y每增加b1。

练习1、使用sklearn估计孩子身高和父母身高之间的线性回归关系。

男女分开预测、正则化后再预测

计算机程序设计课程：<https://cn.udacity.com/course/design-of-computer-programs--cs212>

作业提交邮箱：[xmurpi@163.com](mailto:xmurpi@163.com)

0818

EDA

查看类分布：看类别分布是不是类别不平衡的。

查看特征的skew，对skew较大的特征取log。对一些值较大的特征也可取log可提升模型预测效果。

取log：保留特征原始的相对大小，但是差距变小。

pandas画图：<https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html>

数据透视表功能：<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.pivot_table.html>

时间特征处理：考虑时间的周期性，1天=24小时=24\*3600秒，划分为不同的hour, week, holiday, minute, day。

特征相关性分析：data.corr()，一般分析预测目标与哪个特征的相关性比较高。

若数据不平衡，其对相关性会存在很大的误导性。

正负数据平衡的比例：1:1最好，可容忍范围4:6。若类别数据比例低于3，则可认为数据不平衡问题。

异常值检测：

1. 画图
2. 四分位数法：IQR=Q\_3-Q\_1，Outliers>Q\_3+k\*IQR，或<Q\_1-k\*IQR，但主要找extreme outliers。

分类问题数据不平衡处理方法：

1. 收集更多数据
2. using weight（class\_weight属性）：样本不平衡时，模型会更倾向于去学习占比大的类别
3. 改变评价指标：混淆矩阵（多分类需转化为二分类），accuracy（正确率），precision（精确率），recall（召回率），F1 Score，AUC

Precision：预测为正样本中有多少是预测正确的。

Recall：原始有多少正样本被预测为正样本。

1. 采样

回归问题评价指标：MAE、MSE、Log Loss（神经网络）。

numba库：加速代码运行。

作业：

1. 根据不同权重class\_weight，画出LogisticRegression评价指标。
2. 补充完整评价指标函数定义(TP\FP\TN\FN)，思考fast\_auc()。
3. 复习所有东西。
4. 预习Resampling，Cross-validation。