

作者华校专，曾任阿里巴巴资深算法工程师，现任智易科技首席算法研究员，《Python 大战机器学习》的作者。

这是作者多年以来学习总结的笔记，经整理之后开源于世。目前还有约一半的内容在陆续整理中，已经整理好的内容放置在此。曾有出版社约稿，但是考虑到出版时间周期较长，而且书本购买成本高不利于技术广泛传播，因此作者采取开源的形式。笔记内容仅供个人学习使用，非本人同意不得应用于商业领域。

笔记内容较多，可能有些总结的不到位的地方，欢迎大家探讨。联系方式:huaxz1986@163.com

数学基础

- [1.线性代数基础](#)
 - 一、基本知识
 - 二、向量操作
 - 三、矩阵运算
 - 四、特殊函数
- [2.概率论基础](#)
 - 一、概率与分布
 - 二、期望和方差
 - 三、大数定律及中心极限定理
 - 五、常见概率分布
 - 六、先验分布与后验分布
 - 七、信息论
 - 八、其它
- [3.数值计算基础](#)
 - 一、数值稳定性
 - 二、梯度下降法
 - 三、二阶导数与海森矩阵
 - 四、牛顿法
 - 五、拟牛顿法
 - 六、约束优化

统计学习

- [0.机器学习简介](#)
 - 一、基本概念
 - 二、监督学习
 - 三、机器学习三要素
- [1.线性代数基础](#)
 - 一、线性回归
 - 二、广义线性模型
 - 三、对数几率回归

- 四、线性判别分析
 - 五、感知机
- [2.支持向量机](#)
 - 一、线性可分支持向量机
 - 二、线性支持向量机
 - 三、非线性支持向量机
 - 四、支持向量回归
 - 五、SVDD
 - 六、序列最小最优化方法
 - 七、其它讨论
- [3.朴素贝叶斯](#)
 - 一、贝叶斯定理
 - 二、朴素贝叶斯法
 - 三、半朴素贝叶斯分类器
 - 四、其它讨论
- [4.决策树](#)
 - 一、原理
 - 二、特征选择
 - 三、生成算法
 - 四、剪枝算法
 - 五、CART 树
 - 六、连续值、缺失值处理
 - 七、多变量决策树
- [5.knn](#)
 - 一、k 近邻算法
 - 二、kd树
- [6.集成学习](#)
 - 一、集成学习误差
 - 二、Boosting
 - 三、Bagging
 - 四、集成策略
 - 五、多样性分析
- [7.梯度提升树](#)
 - 一、提升树
 - 二、xgboost
 - 三、LightGBM
- [8.特征工程](#)
 - 一、缺失值处理
 - 二、特征编码
 - 三、数据标准化、正则化
 - 四、特征选择
 - 五、稀疏表示和字典学习
 - 六、多类分类问题

- 七、类别不平衡问题
- [9.模型评估](#)
 - 一、泛化能力
 - 二、过拟合、欠拟合
 - 三、偏差方差分解
 - 四、参数估计准则
 - 五、泛化能力评估
 - 六、训练集、验证集、测试集
 - 七、性能度量
 - 七、超参数调节
 - 八、传统机器学习的挑战
- [10.降维](#)
 - 一、维度灾难
 - 二、主成分分析 PCA
 - 三、核化线性降维 KPCA
 - 四、流形学习
 - 五、度量学习
 - 六、概率PCA
 - 七、独立成分分析
 - 八、t-SNE
 - 九、LargeVis
- [11.聚类](#)
 - 一、性能度量
 - 二、原型聚类
 - 三、密度聚类
 - 四、层次聚类
 - 五、谱聚类
- [12.半监督学习](#)
 - 半监督学习
 - 一、生成式半监督学习方法
 - 二、半监督 SVM
 - 三、图半监督学习
 - 四、基于分歧的方法
 - 五、半监督聚类
 - 六、总结
- [13.EM算法](#)
 - 一、示例
 - 二、EM算法原理
 - 三、EM算法与高斯混合模型
 - 四、EM 算法与 kmeans 模型
 - 五、EM 算法的推广
- [14.最大熵算法](#)
 - 一、最大熵模型MEM

- 二、分类任务最大熵模型
- 三、最大熵的学习
- [15. 隐马尔可夫模型](#)
 - 一、隐马尔可夫模型HMM
 - 二、HMM 基本问题
 - 三、最大熵马尔科夫模型MEMM
- [16. 概率图与条件随机场](#)
 - 一、概率图模型
 - 二、贝叶斯网络
 - 三、马尔可夫随机场
 - 四、条件随机场 CRF

深度学习

- [0. 深度学习简介](#)
 - 一、介绍
 - 二、历史
- [1. 机器学习基础](#)
 - 一、基本概念
 - 二、点估计、偏差方差
 - 三、最大似然估计
 - 四、贝叶斯估计
 - 五、随机梯度下降
 - 七、传统机器学习的挑战
 - 八、低维流形
- [2. 深度前馈神经网络](#)
 - 一、基础
 - 二、损失函数
 - 三、输出单元
 - 四、隐单元
 - 五、结构设计
 - 六、历史小记
- [3. 反向传播算法](#)
 - 一、链式法则
 - 二、反向传播
 - 三、深度前馈神经网络
 - 四、实现
 - 五、应用
 - 六、自动微分
- [4. 正则化](#)
 - 一、基本概念
 - 二、参数范数正则化
 - 三、约束正则化

- 四、数据集增强
- 五、噪声鲁棒性
- 六、早停
- 七、参数共享
- 八、dropout
- 九、稀疏表达
- 十、半监督学习与多任务学习
- 十一、对抗训练
- 十二、正切传播算法
- 十三、正则化和欠定问题
- [5.最优化基础](#)
 - 一、代价函数
 - 二、神经网络最优化挑战
 - 三、mini-batch
 - 四、基本优化算法
 - 五、自适应学习率算法
 - 六、二阶近似方法
 - 七、共轭梯度
 - 八、优化策略和元算法
 - 九、参数初始化策略
- [6.卷积神经网络](#)
 - 一、卷积运算
 - 二、卷积层、池化层
 - 三、基本卷积的变体
 - 四、算法细节
 - 五、历史和现状
- [7.循环神经网络](#)
 - 一、RNN计算图
 - 二、循环神经网络
 - 三、长期依赖
 - 四、序列到序列架构
 - 五、递归神经网络
 - 六、回声状态网络
 - 七、LSTM 和其他门控RNN
 - 八、外显记忆
- [8.工程实践指导原则](#)
 - 一、性能度量
 - 二、默认的基准模型
 - 三、决定是否收集更多数据
 - 四、选择超参数
 - 五、调试策略
 - 六、示例：数字识别系统
 - 七、数据预处理

- 八、变量初始化
- 九、结构设计

自然语言处理

- [主题模型](#)
 - 一、Unigram Model
 - 二、pLSA Model
 - 三、LDA Model
 - 四、模型讨论
- [词向量](#)
 - 一、向量空间模型 VSM
 - 二、LSA
 - 三、Word2Vec
 - 四、GloVe

计算机视觉

- [图片分类网络](#)
 - 一、LeNet
 - 二、AlexNet
 - 三、VGG-Net
 - 四、Inception
 - 五、ResNet
 - 六、SENet
 - 七、DenseNet
 - 八、小型网络
 - 九、趋势

工具

CRF

- [CRF++](#)
 - 一、安装
 - 二、使用
 - 三、Python接口
 - 四、常见错误

lightgbm

- [lightgbm使用指南](#)
 - 一、安装
 - 二、调参

- 三、进阶
- 四、API
- 五、Docker

xgboost

- [xgboost使用指南](#)
 - 一、安装
 - 二、调参
 - 三、外存计算
 - 四、GPU计算
 - 五、单调约束
 - 六、DART booster
 - 七、Python API

scikit-learn

- [1.预处理](#)
 - 一、特征处理
 - 二、特征选择
 - 三、字典学习
 - 四、PipeLine
- [2.降维](#)
 - 一、PCA
 - 二、MDS
 - 三、Isomap
 - 四、LocallyLinearEmbedding
 - 五、FA
 - 六、FastICA
 - 七、t-SNE
- [3.监督学习模型](#)
 - 一、线性模型
 - 二、支持向量机
 - 三、贝叶斯模型
 - 四、决策树
 - 五、KNN
 - 六、AdaBoost
 - 七、梯度提升树
 - 八、Random Forest
- [4.模型评估](#)
 - 一、数据集切分
 - 二、性能度量
 - 三、验证曲线 && 学习曲线
 - 四、超参数优化

- [5. 聚类模型](#)
 - 一、KMeans
 - 二、DBSCAN
 - 三、MeanShift
 - 四、AgglomerativeClustering
 - 五、BIRCH
 - 六、GaussianMixture
 - 七、SpectralClustering
- [6. 半监督学习模型](#)
 - 一、标签传播算法
- [7. 隐马尔可夫模型](#)
 - 一、Hmmlearn
 - 二、seqlearn

spark

- [1. 基础概念](#)
 - 一、核心概念
 - 二、安装和使用
 - 三、pyspark shell
 - 四、独立应用
- [2. rdd使用](#)
 - 一、概述
 - 二、创建 RDD
 - 三、转换操作
 - 四、行动操作
 - 五、其他方法和属性
 - 六、持久化
 - 七、分区
 - 八、混洗
- [3. dataframe使用](#)
 - 一、概述
 - 二、SparkSession
 - 三、DataFrame 创建
 - 四、DataFrame 保存
 - 五、DataFrame
 - 六、Row
 - 七、Column
 - 八、GroupedData
 - 九、functions
- [4. 累加器和广播变量](#)
 - 一、累加器
 - 二、广播变量

numpy

- [numpy 使用指南](#)
 - 一、 ndarray
 - 二、 ufunc 函数
 - 三、 函数库
 - 四、 数组的存储和加载

scipy

- [scipy 使用指南](#)
 - 一、 常数和特殊函数
 - 二、 拟合与优化
 - 三、 线性代数
 - 四、 统计
 - 五、 数值积分
 - 六、 稀疏矩阵

matplotlib

- [matplotlib 使用指南](#)
 - 一、 matplotlib配置
 - 二、 matplotlib Artist
 - 三、 基本概念
 - 四、 布局
 - 五、 Path
 - 六、 path effect
 - 七、 坐标变换
 - 八、 3D 绘图
 - 九、 技巧

pandas

- [pandas 使用指南](#)
 - 一、 基本数据结构
 - 二、 内部数据结构
 - 三、 下标存取
 - 四、 运算
 - 五、 变换
 - 六、 数据清洗
 - 七、 字符串操作
 - 八、 聚合与分组
 - 九、 时间序列
 - 十、 DataFrame 绘图
 - 十一、 移动窗口函数

- 。 十二、 数据加载和保存