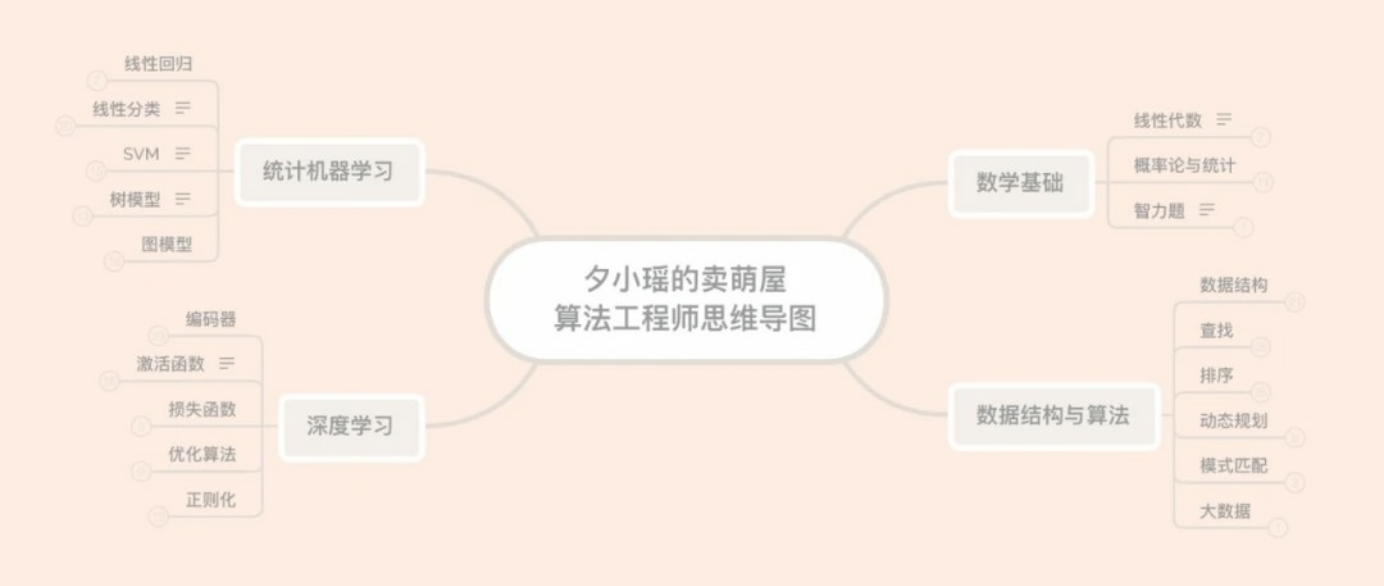
面试必备！卖萌屋算法工程师思维导图—统计机器学习篇

原创 rumor酱 夕小瑶的卖萌屋 6月10日

来自专辑

卖萌屋@算法工程师思维导图



卖萌屋的妹子们（划掉）作者团整理的**算法工程师思维导图**，求职/自我提升/查漏补缺神器。该手册一共分为**数据结构与算法、数学基础、统计机器学习和深度学习**四个部分。

下面是第二部分统计机器学习的内容~

公众号后台回复【思维导图】获取完整手册（Xmind脑图源文件，学习起来更方便(ง •_•)ง

统计机器学习

线性回归

- 回归方程：Y = WX
- 解析解：$(X^TX)^{-1}X^TY$
- 损失函数-最小二乘法

  用高斯概率密度函数表示出y，然后进行极大似然估计

  - 理解：频率派角度-误差复合高斯分布的最大似然估计
  - 求法：误差服从正太分布(0,sigma) => y服从正太分布(wx,sigma)
- 正则化

从两个角度理解：

- 频率角度：维度太大无法求逆矩阵，且容易过拟合，给w加上约束 X^TX是半正定，不一定可逆，$X^TX+lambdaI$ 为半正定加单位矩阵，是正定的，可逆
- 贝叶斯角度（最大后验）：参数符合laplace分布>L1正则，符合高斯分布>L2岭回归

线性分类

线性分类器是通过特征的线性组合来做出分类决定的分类器。数学上来说，线性分类器能找到权值向量w，使得判别公式可以写成特征值的线性加权组合。

- 硬分类
  - 感知机

    二分类模型，y为{-1, 1} 损失函数：误分类点到分类平面到距离，分对为0，分错>0； $L(w)=-sum(yi(wxi+b))$
  - Fisher判别分析

    把样本点投影到一个平面，类间均值差大，使得类内方差小
- 软分类

  P(Y|X) = P(X|Y)P(Y) / P(X)

  判别模型直接求P(Y|X) 生成模型求P(X,Y)=>P(X|Y)P(Y)=>P(Y|X)
  - 判别式: 逻辑回归

    由对数几率=>sigmoid：
    https://zhuanlan.zhihu.com/p/42656051

    公式推导：
    https://zhuanlan.zhihu.com/p/44591359

    简介：逻辑回归是使用sigmoid作为链接函数的广义线性模型，应用于二分类任务。它假设数据服从伯努利分布，对条件概率进行建模，通过极大似然估计的方法，运用梯度下降求解参数。

    $$y=\frac{1}{1+e^{-w^Tx}}$$

    目标函数：

    $$P_{all}=\prod_i p^{y_i}(1-p)^{1-y^i}$$

    求解：迭代法（为什么不求解析解？换成矩阵形式后，X和exp(X)同时存在，无法求出解析解。）逻辑回归的损失函数L是一个连续的凸函数，它只会有一个全局最优的点，不存在局部最优。可以用SGD。

Bias的可解释性:对于偏差b (Bias)，一定程度代表了正负两个类别的判定的容易程度。假如b是0，那么正负类别是均匀的。如果b大于0，说明它更容易被分为正类，反之亦然。

线性决策边界：

为什么不能用线性回归做分类?

https://www.zhihu.com/question/319865092/answer/661614886

平方差的意义和交叉熵的意义不一样。概率理解上，平方损失函数意味着模型的输出是以预测值为均值的高斯分布，损失函数是在这个预测分布下真实值的似然度，softmax损失意味着真实标签的似然度。

- **由来及其表达式：用线性回归拟合 $p > 1 - p$，得到对数几率回归**

- 生成式：朴素贝叶斯

  朴素贝叶斯是基于贝叶斯定理与特征条件独立假设大分类方法，对于给定的x，对x，y的联合分布建模(P(x|y)&P(y))，输出后验概率最大的Y，对P(x|y)采用了极大似然估计

  当特征离散时为线性分类：离散特征的朴素贝叶斯分类器判别公式能够写成特征值的加权线性组合。
  https://www.jianshu.com/p/469accb2e1a0

  假设：特征间相互独立,P(x1|y)与P(x2|y)相互独立 P(x1, x2, .., xn | Y) = P(x1|Y) * P(x2|Y) * ... * P(xn|Y)

  求解:对于给定的x，对x，y的联合分布建模(P(x|y)&P(y))，输出后验概率最大的Y，对P(x|y)采用了极大似然估计

  max[P(x|y)P(y)]，y服从伯努利分布，x|y服从categorial分布或高斯分布

  一般假设朴素贝叶斯的特征为离散值

  - 生成式：高斯判别分析

    假定已知类中的x的分别服从高斯分布，对于二分类，p(x|y=0)和p(x|y=1)分别服从两个高斯分布，方差一样，y服从bernoulli(p), $P(y) = p^y(1-p)^(1-y)$

    方差相同的情况下为线性分类（可以写成特征值x的线性加权组合）：
    https://www.jianshu.com/p/469accb2e1a0

    方差相同时把x^2消掉了，否则带有x^2就不是线性了

SVM

https://zhuanlan.zhihu.com/p/61123737

解读：https://zhuanlan.zhihu.com/p/49331510

考点：https://zhuanlan.zhihu.com/p/76946313

- 分类

  - 线性可分SVM

当训练数据线性可分时，通过硬间隔(hard margin，什么是硬、软间隔下面会讲)最大化可以学习得到一个线性分类器，即硬间隔SVM

- 线性SVM

  当训练数据不能线性可分但是可以近似线性可分时，通过软间隔(soft margin)最大化也可以学习到一个线性分类器，即软间隔SVM

- 非线性SVM

  当训练数据线性不可分时，通过使用核技巧(kernel trick)和软间隔最大化，可以学习到一个非线性SVM

- 线性可分SVM凸二次规划形式的推导
- 拉格朗日乘子法和KKT条件
- 凸二次规划求解
- 软间隔最大化
- 序列最小优化算法(SMO)
- 核函数
- 常见问题
  - **与感知机的区别**
  - **与逻辑回归的对比**
  - SVM**优缺点**
  - **二分类到多分类**

树模型

- 常见算法
  - XGBOOST

    优点：不用做特征标准化，可以处理缺失数据，对outlier不敏感

    理解泰勒展开：
    https://www.zhihu.com/question/25627482/answer/31229830

    理解GBDT：
    https://www.zybuluo.com/yxd/note/611571

    官方文档：
    https://github.com/dmlc/xgboost/tree/master/demo http://xgboost.readthedocs.io/en/latest/parameter.html#general-parametershttp://xgboost.readthedocs.io/en/latest/python/python_api.html

调参：

https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

https://www.dataiku.com/learn/guide/code/python/advanced-xgboost-tuning.html

源码剖析：

https://wenku.baidu.com/view/44778c9c312b3169a551a460.html min_child_weight:

https://www.zhihu.com/question/68621766

scale_pos_weight: https://blog.csdn.net/h4565445654/article/details/72257538

节点分裂：H，Weighted Quantile Sketch, h对loss有加权的作用

稀疏值处理：行抽样、列抽样

Shrinkage：学习速率减小，迭代次数增多，有正则化作用

系统设计：Columns Block, Cache Aware Access Gradient-based One Side Sampling (GOSS) Exclusive Feature Bundling (EFB)

- LightGBM

  官方文档：

  http://lightgbm.readthedocs.io/en/latest/ https://github.com/Microsoft/LightGBM

  改进：直方图算法；直方图差加速；Leaf-wise建树；特征并行和数据并行的优化

- Random Forest

  调参：

  http://www.cnblogs.com/pinard/p/6160412.html

  https://www.zhihu.com/question/34470160/answer/114305935

  https://zhuanlan.zhihu.com/p/25308120

  原理：

  http://www.cnblogs.com/pinard/p/6156009.html https://www.jianshu.com/p/dbf21ed8be88

  优化：

  https://stackoverflow.com/questions/23075506/how-to-improve-randomforest-performance

- 信息熵相关概念
- 生成
  - ID3:信息增益
  - C4.5:信息增益比
  - CART：回归-平方误差/分类-基尼指数
- 剪枝

- 叶节点个数
- 预剪枝/后剪枝

图模型

- 有向图

  https://www.zhihu.com/question/53458773/answer/554436625

  贝叶斯网络(Bayesian Networks, BNs)是有向图, 每个节点的条件概率分布表示为P(当前节点|父节点)

  从朴素贝叶斯到HMM：在输出序列的y时，依据朴素贝叶斯只有 $p(y_i, x_i) = P(x_i|y_i)P(y_i)$。没有考虑yi之间的关系，因此加入$P(y_i|y_i-1)$，得到HMM

  - HMM

    定义：HMM是关于时序的概率模型，由一个隐藏的马尔可夫链生成不可观测的状态随机序列，再由各个状态生成观测序列

    三要素：初始状态概率向量，状态转移矩阵A，观测/发射概率矩阵B

    假设：齐次马尔可夫&观测独立

    概率计算：给定三要素和观测序列，生成观测序列概率

    学习问题：给定观测序列，用极大似然估计三要素

    预测/解码：给定观测序列和三要素，求最可能的状态序列

  - 朴素贝叶斯

    https://www.zhihu.com/question/53458773/answer/554436625

- 无向图

  https://www.zhihu.com/question/53458773/answer/554436625

  马尔可夫网络则是无向图, 包含了一组具有马尔可夫性质的随机变量. 马尔可夫随机场(Markov Random Fields, MRF)是由参数(S,π,A)表示, 其中S是状态的集合，π是初始状态的概率, A是状态间的转移概率。一阶马尔可夫链就是假设t时刻的状态只依赖于前一时刻的状态，与其他时刻的状态和观测无关。这个性质可以用于简化概率链的计算。

  - 逻辑回归

    https://www.zhihu.com/question/265995680/answer/303148257

    朴素贝叶斯与逻辑回归的关系：都是对几率P/(1-P)进行拟合。朴素贝叶斯基于条件独立假设，另特征间相互独立，通过P(X|Y)P(Y)=>联合概率分布求得几率

    逻辑回归拟合特征间的关系：用线性回归逼近几率

  - CRF

    模型定义：

举例：

https://zhuanlan.zhihu.com/p/104562658

无向图：在给一个节点打标签时，把相邻节点的信息考虑进来（马尔可夫性：只与相邻的两个状态有关）

线性链条件随机场：P(Yi|X,Y1,...Yn) = P(Yi|X, Yi-1, Yi+1)，只考虑当前和前一个 由输入序列预测输出序列的判别模型，对条件概率建模

观测序列，状态/标记序列

特征函数：转移特征t（依赖当前和前一个位置），状态特征s（依赖当前位置），t和s对取值为1或0

特征函数：

转移特征t（依赖当前和前一个位置），状态特征s（依赖当前位置），t和s对取值为1或0

与逻辑回归比较：CRF是逻辑回归的序列化版本

与HMM比较：每一个HMM模型都可以用CRF构造出来。CRF更加强大：

1.CRF可以定义数量更多，种类更丰富的特征函数。HMM从朴素贝叶斯而来，有条件独立假设，每个观测变量只与状态变量有关。但是CRF却可以着眼于整个句子s定义更具有全局性的特征函数

2.CRF可以使用任意的权重。将对数HMM模型看做CRF时，特征函数的权重由于是log形式的概率

https://zhuanlan.zhihu.com/p/31187060

1.HMM是生成模型，CRF是判别模型

2.HMM是概率有向图，CRF是概率无向图

3.HMM求解过程可能是局部最优，CRF可以全局最优（对数似然为凸函数）

4.CRF概率归一化较合理，HMM则会导致label bias 问题