

# Github大佬笔记（机器学习算法篇） 2

- 从特征空间角度
- Logistic Regression:
  - 特征的线性组合求交叉熵的最小化，也就是对特征的线性组合做 logistic，使得 Logistic Regression 会在特征空间中做线性分界面，适用于分类任务；
- GBDT:
  - 采用 CART 树作为基分类器，其每轮树的特征拟合都是对特征空间做平行于坐标轴的空间分割，所以自带特征选择和可解释性，GBDT 即可处理分类问题也可解决回归问题，只是其统一采用回归思路进行求解（试想，如果不将分类转换为回归问题，GBDT 每轮目标函数旨在拟合上一轮组合模型的负梯度，分类信息无法求梯度，故而依旧是采用 softmax 转换为回归问题进行求解）；
- 相关知识
  - 线性分类器：
    - 感知器准则函数：
      - 原理：以使错分类样本到分界面距离之和最小为原则；
      - 优点：通过错分类样本提供的信息对分类器函数进行修正，这种准则是人工神经网络多层感知器的基础；
    - SVM：
      - 基本思想：在两类线性可分条件下，所设计的分类器界面使两类之间的间隔为最大；
      - 基本出发点：使期望泛化风险尽可能小。（使用核函数可解决非线性问题）
    - Fisher 准则【线性判别分析（LDA）】：
      - 介绍：将所有样本投影到一条原点出发的直线，使得同类样本距离尽可能小，不同类样本距离尽可能大，具体为最大化“广义瑞利商”；
      - 特点：两类样本一般类内密集，类间分离；
      - 方式：寻找线性分类器最佳的法线向量方向，使两类样本在该方向上的投影满足类内尽可能密集，类间尽可能分开。这种度量通过类内离散矩阵  $S_w$  和类间离散矩阵  $S_b$  实现。
- 从正则角度
- Logistic Regression:
  - 方式：采用一种约束参数稀疏的方式；
    - L2 正则整体约束权重系数的均方和，使得权重分布更均匀；
    - L1 正则约束权重系数绝对值和，其自带特征选择特性；
- GBDT:
  - 弱算法的个数  $T$ ，就是迭代  $T$  轮。 $T$  的大小就影响着算法的复杂度；

- 步长 (Shrinkage) 在每一轮迭代中, 原来采用  $F_t(x) = F_{t-1}(x) + \alpha_t h_t(x; wt)$  进行更新, 可以加入步长  $v$ , 使得一次不更新那么多:
- 区别:
  - LR 采用对特征系数进行整体的限定;
  - GBDT 采用迭代的误差控制本轮参数的增长;
- XGBoost的正则是在 GBDT 的基础上又添加了一棵树里面节点的个数, 以及每个树叶子节点上面输出分数的 L2 模平方
- 从算法角度
  - Logistic Regression
    - 若采用 SGB, Momentum, SGD with Nesterov Acceleration 等算法, 只用到了二阶导数信息,
    - 若用 AdaGrad, AdaDelta / RMSProp, Adam, Nadam, 牛顿法则用到了二阶导数信息,
  - GBDT 直接拟合上一轮组合函数的特梯度, 只用到了二阶导数信息, XGBoost 则是用到了二阶导数信息。
- 从特征的角度
  - 特征组合
    - LR 特征选择方法: 不具有特征组合的能力, 只是一次性地寻求最大化熵的过程, 对每一维的特征都假设独立, 因此只具备对已有特征空间进行分割的能力, 更不会对特征空间进行升维 (特征组合)
    - GBDT 特征选择方法: 采用最小化均方损失来寻找分裂特征及对应分裂点, 所以自动会在当前根据特征 A 分裂的子树下寻求其他能使负梯度最小的其他特征 B, 这样就自动具备寻求好的特征组合的性能, 因此也能给出哪些特征比较重要 (根据该特征被选作分裂特征的次数)。
  - 特特征的稀疏性
    - Logistic Regression不具有特征组合的能力, 并假设特征各个维度独立, 因此只具有线性分界面, 实际应用中, 多数特征之间有相关性, 只有维度特别大的稀疏数据中特征才会近似独立, 所以适合应用在特征稀疏的数据上;
    - GBDT: 其更适合处理稠密特征, 如 GBDT+LR 的Facebook论文中, 对于连续型特征导入 GBDT 做特征组合来代替一部分手工特征工程, 而对于 ID 类特征的做法往往是 one-hot 之后直接传入 LR, 或者先 hash, 再 one-hot 传入树中进行特征工程, 而目前的主流做法是直接 one-hot + embedding 来将高维稀疏特征压缩为低维稠密特征, 也进一步引入了语意信息, 有利于特征的表达。
  - 数据假设不同
    - LR
    - 第一个假设: 假设数据服从伯努利分布。
    - 伯努利分布有一个简单的例子是抛硬币, 抛中为正面的概率是  $p$ , 抛中为负面的概率是  $1-p$ 。在逻辑回归这个模型里面是假设  $h_{\theta}$  为样本为正的,  $1 - h_{\theta}$  为样本为负的概率。那么整个模型可以描述为:

- 第二个假设：假设样本为正的的概率是：
  - 第三个假设：噪声是高斯分布的
  - GBDT
  - 未对数据做出上述假设。
- 机器学习面试150题：不只是考SVM xgboost 特征工程 (1-50) - 知乎 (zhihu.com).
  - 机器学习面试150题：不只是考SVM xgboost 特征工程 (51-100) |附送【AI工程师大礼包】 - 知乎 (zhihu.com).
  - 机器学习面试150题：不只是考SVM xgboost 特征工程 (101-153) 附送【AI工程师大礼包】 - 知乎 (zhihu.com).

以上内容整理于 [幕布文档](#)