1 LR, SVM, RandomForest, XGBoost,LightGBM各个模型的推导

2 XGBoost和LightGBM的不同点, LightGBM的改进点

3 XGBoost怎样进行分裂节点的

4 XGBoost和GBDT的不同点

5 XGBoost, LightGBM怎样进行并行化的

6 树分裂节点的公式ID3(信息增益), C4.5(信息增益率), CART(平方损失,Gini系数)

7 ID3, C4.5是怎样进行剪枝的, CART是怎样进行剪枝的 (两种剪枝的方法)

8 树结构的优点

计算快, 可解释性强直观, 可以处理缺失值, 可以处理混合特征；缺点：容易过拟合，对异常值特别敏感，不适合处理高维特征的数据

9 XGBoost怎样处理离散特征？

10 XGBoost怎样处理缺省值？

11 随机森林每次可以用到多少样本量？

12 从ID3到C4.5的原因,以及C4.5的分裂节点的启发式方法？

13 决策树剪枝分为?

14 决策树预剪枝？

15 决策树怎样处理连续值？

16 决策树怎样处理缺省值？如何选择属性(不考虑缺失值)？如何划分有缺失值的属性(给予权重, 权重是累积的吗？)？

二 XGBoost怎样用于文本分类:

计算词袋模型(bag of words, n-grams), 表示成向量直接输入到LR,XGBoost,RandomForest

怎样降维的呢？去停用词,去低频词处理掉一部分

用tf-idf来过滤掉一部分，特征选择：卡方检验和信息增益进行特征选择后输入到分类器中进行分类.

词向量的使用方法：先用特征选择后留下词，然后将这些词的词向量进行求和喂进分类器进行分类

共现矩阵：用一个窗口为5来统计一个词与另一个词的共现，然后用SVD进行奇异值矩阵分解得到每个词的词稠密的词向量,然后用于分解

三

深度学习

1 Dropout

2 batch notmalization

ELMO的使用方法

ELMO每层代表什么

高层特征提取器

Avg-pooling, max-pooling, attention, capsule的相比较