京东云算法工程师一面分享

**总结下自己今天京东云面试的经历，希望能抛砖引玉，结识做机器学习方面的朋友交流学习**

**1.介绍一下自己**

答：从实验室项目介绍到实习的工作，重点介绍了实习文本分类项目特征工程的过程。

**2.针对项目提问**

**2.1 逻辑回归的目标函数和优化方法**

答：不多说，目标函数是服从二项分布的似然函数，优化常用的是梯度下降法

**2.3 讲下拟牛顿法**

答：对比了下梯度下降法只是泰勒的一阶展开式，而牛顿法是泰勒的二阶展开式，牛顿法主要问题在于海森矩阵求逆是一个很复杂的过程，所有才会有拟牛顿法以及相应的改进算法。

**3.讲下随机森林或者GDBT**

答：随机森林采用的是bagging的思想，bagging又称为bootstrap aggreagation，通过在训练样本集中进行有放回的采样得到多个采样集，基于每个采样集训练出一个基学习器，再将基学习器结合。随机森林在对决策树进行bagging的基础上，在决策树的训练过程中引入了随机属性选择。传统决策树在选择划分属性的时候是在当前节点属性集合中选择最优属性，而随机森林则是对结点先随机选择包含k个属性的子集，再选择最优属性，k作为一个参数控制了随机性的引入程度。

GDBT欢迎留言补充，就不写了

**4.上者两种算法的区别**

答：（面试后整理）

GBDT和随机森林的相同点：

1、都是由多棵树组成

2、最终的结果都是由多棵树一起决定

GBDT和随机森林的不同点：

1、组成随机森林的树可以是分类树，也可以是回归树；而GBDT只由回归树组成

2、组成随机森林的树可以并行生成；而GBDT只能是串行生成

3、对于最终的输出结果而言，随机森林采用多数投票等；而GBDT则是将所有结果累加起来，或者加权累加起来

4、随机森林对异常值不敏感，GBDT对异常值非常敏感

5、随机森林对训练集一视同仁，GBDT是基于权值的弱分类器的集成

6、随机森林是通过减少模型方差提高性能，GBDT是通过减少模型偏差提高性能

**5.随机森林怎么取最后的结果？**

答：对于分类任务，随机森林是多数表决；

对于回归任务，随机森林是简单平均

**6.随机森林是怎样避免ID3算法信息增益的缺点的？**

答：首先说下信息增益的过程，决策树算法本质上就是要找出每一列的最佳划分以及不同列划分的先后顺序及排布。信息增益的缺点是比较偏向选择取值多的属性。而gini系数每次都是二分，所以跟属性多少没有关系。

**7.为什么deep learning 能抑制梯度消失或者爆炸的问题？**

答:几个方面：一是激活函数不光是只用sigmoid函数，还有 ReLU函数 二是在参数并不是初始化的时候并不是随机选择的，而是在前面有自编码器做了特征特征器，这样避免了梯度下降法求解陷入局部最优解；三，深度学习一些手段，权值共享，卷积核，pooling等都能抑制梯度消失问题；四，二次代价函数换成交叉熵损失函数或者选用softmax+对数似然代价函数的组合。

**8.你可以讲下你熟悉的，不常见的算法？**

答：结合业务场景讲了一个无监督的社群算法

**9.你还有什么问题要问的？**

答：问了下他们的业务场景

整个过程40分钟左右

**总结：随机森林和GBDT没实际用过，但是还是知道一二的，有些答的比较浅，整个氛围还算比较轻松。**