美团猫眼电影机器学习工程师一面题目分享

猫眼电影一面分享,供大家参考，同时希望路过大神们能给我这个小白更圆满的答案

**1.基础算法题**

有序的数组，其中一个数出现一次，其中的数出现两次，找到这个出现一次的数？

答：很自然想到hashmap或者异或的做法，时间复杂度O（N）。回答的时候觉得有坑，因为有序的条件没用，心想肯定跟二分有关系，果然面试官要求降低复杂度。没思考出来二分的策略，后来在面试官的提醒下明白了，二分的时候跟数组的位置联系起来，既数组肯定是基数个，二分的时候如果mid不成立的话，要找的值一定落在跟左右两边相等元素的那一侧，时间复杂度O（logN）

**2.介绍下自己做的项目**

之前做过聊天式文本挖掘，主要介绍了针对短文本的特征所做的特征工程

针对项目提问：

*2.1.讲下LR模型*

答：在线性回归的基础上加了sigmoid函数，所有变成了分类模型，输出可以表示概率。采用似然估计构造目标函数，目标是优化最大似然估计，公式H(x)=-(ylogf(x)+(1-y)log(1-f(x)),一般优化方法采用的是梯度下降法。回答完毕。

*2.2.LR模型为什么采用似然估计损失函数*

答：因为最小二乘法是假设残差服从正太分布的，而LR在sigmoid 作用后就不会服从正态分布了，所以采用的是最大似然估计。

面试后思考：1.最小二乘法反映的是线性空间上的线性投影的最短距离，在非线性空间上表现不如MLE。（MLE可以看作一种特殊情况下的Bayesian 估计，具体来说，就是在prior 是 diffuse （无知的）情况下，让posterior 分布取得极大值的系数值）

2.如果采用均方差最损失函数的时候，梯度下降求偏导时会有一项导数项，这样会导致梯度在一定阶段会收敛的特别慢，而对数损失函数log正好能和sigmoid的exp抵消掉，会加快收敛速度。

*2.3.说下基本主题的LDA模型*

答：判别式模型，思想：文档一定概率上从属于某些个主题，主题一定概率上会选中某些相关的词，这样就构造了文档到主题到词的联系，同时可以解决同义词问题，因为同义词可能属于不同主题。算法流程回答的模糊，感觉面试官不太满意。

*2.4.说下项目用到的doc2vec怎么产生的？*

答：介绍了下word2vec的思想，然后讲传统上通常词向量简单的加权求和来表征一篇文档，而doc2vec训练方式是在word2vec的基础上，加入了段落ID，进行了一层训练，这样好处是保留了词的上下文信息。（ps：解释的不清楚，自己不太满意）

*2.5.说下论文中频繁序列挖掘prefixspan 算法？*

答：（ps：因为这个算法不做序列挖掘基本不知道，可能只了解apriori算法）对比apriori算法的过程和缺点，讲解该算法的优势，只需要扫描一次序列数据集，目标是挖掘出满足最小支持度的频繁序列，长度为1的前缀开始挖掘序列模式，搜索对应的投影数据库的频繁序列，然后递归的挖掘长度为2的前缀所对应的频繁序列。以此类推，一直递归到对应的投影数据库为空或者对应投影数据库中各项的支持度计数小于阈值为止。整个过程就是前缀不断的增长，产生1，2...N 频繁序列，对应的投影数据库不断缩小直至为空。

优点：PrefixSpan算法由于不用产生候选序列，且投影数据库缩小的很快，内存消耗比较稳定，作频繁序列模式挖掘的时候效果很高。

至此，项目上的提问到此结束。

**3.了解深度学习吗？能否讲下CNN的特点？**

答：（ps：内心是拒绝的，不怎么了解啊）神经网络模型前向传递，反向调节的特点（BP网络）,隐含层我觉得是一个特征做变换的过程，整个过程给人的感觉就是向前是特征拓展阶段，向后是参数调优阶段。回到CNN，特点是局部感受和权值共享，通过卷积核扫描原始数据能够学习到不同的局部的特征，接着通过池化进一步提取特征，这些做的能够让参数数目有量级的减少，同时权值共享是同一层隐含层共享权值，这样也是减少了隐含层的参数，很多卷积核学习的到特征最后传递到下一层网络，到最后一层采用分类器分类（扯不下去了，开始瞎扯）。深度学习解决了以往神经网络深度网络很多问题，梯度消失爆炸问题，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。

**4.说说RBM编码器**

答：（额，自己给自己挖坑了，看过早忘了）一种特征探测器，每一层学习的特征向上传递，然后反过来微调。好吧，只能回答这么多了。

面试后补充：RBM包括隐层，可见层和偏置层。可见层和隐含层可以双向传播。标准的RBM，隐含层和可见层都是二进制表示，既激活函数的激活值服从二项分布。每一层的节点没有链接，如果假设所有的节点都只能取0或者1，同时全概率分布p(v,h)满足伯努利分布。几个参数：1.可视层和隐含层的权重矩阵，二是可是节点的偏移量，三是隐层的偏移量。这几个参数决定将N维的样本编码M维的样本。

用途：

1.降维，类似稀疏自动编码器

2.用RBM训练得到的权重举证和偏移量作为BP神经网路的初始值，避免陷入局部极小值

3.可以估计联合分布P(v,h)，进而求出p（h|v）。生成式模型

4.直接计算p（h|v）进行分类。判别式模型

一个小时8分钟，面试结束。

**总结：总体感觉不是太满意，简历上写的有的细节也不是很清楚，这个一定得避免。深度学习知识薄弱。**