面试问题总结

黑体部分是一些答得不太好的。

1. 水滴公司（推荐广告、nlp）
2. **python中的list和numpy中的array有什么区别**

①list是Python中的普通列表对象，支持append和attend操作，没有shape属性；array和matrix是numpy数据库中的对象，不支持append和attend操作，具有shape属性。

②一个list中可以存放不同类型的数据，如int、float、str，或者布尔型；而array和matrix中只能存放相同类型的数据。

③list不支持乘法操作；array和matrix支持乘法操作（其中matrix只能是二维的。array和list可以是多维的）。

④list对象不支持一次性读取一行或一列，只能通过指针进行元素的索引；array和matrix即支持一次性读取一行或一列，也支持通过指针来进行元素的索引。

1. **python中的深拷贝和浅拷贝有什么区别**

<https://blog.csdn.net/weixin_30640291/article/details/95298577>





从上面的代码可以看出浅拷贝变更被拷贝对象的内容，拷贝方也会随之更改，而深拷贝则不受影响。

当需要数据共享时使用浅拷贝，数据独立不进行共享时使用深拷贝

1. 逻辑回归中的损失函数？ 推导出逻辑回归损失函数
2. **逻辑回归为什么要用sigmoid函数**

<https://zhuanlan.zhihu.com/p/59137998>

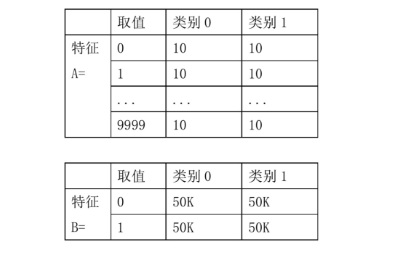
<https://blog.csdn.net/qq_40199232/article/details/106688521>

当因变量服从伯努利分布时，广义线性模型就为逻辑回归。

1. 说说常用决策树
2. **信息增益比对于信息增益来说改善了什么问题**

而在C4.5中，选择的是信息增益比来选择特征，以减少信息增益容易选择特征值多的特征的缺点。

<https://www.zhihu.com/question/22928442?sort=created>



两种特征，特征多的每种取值的总数较小，容易产生更纯的样本（抛硬币抛两次很可能抛出更纯的全正面，但一万次则不可能全正面）故特征A=0的样本只有20个，故更容易生成全为某个类别的样本（即更纯），而特征B=0的样本有100k个样本，故更接近实际的情况。

C4.5使用了信息增益率，在信息增益的基础上除了一项split information,来惩罚值更多的属性。数据量足够时则两者区别不大。

1. **CART树为什么要用GINI指数？**

基尼指数的意义是从数据集D中随机抽取两个样本类别标识不一致的概率。基尼指数越小，数据集的纯度越高。

相比于信息增益，信息增益比等作为特征选择方法，基尼指数省略了对数计算，运算量比较小，也比较容易理解，所以CART树选择使用基尼系数用来做特征选择。

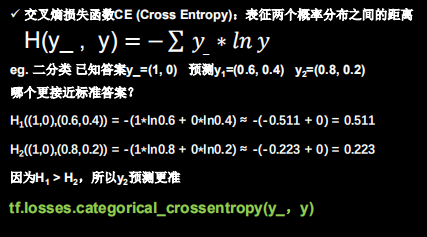
1. Bagging与boosting的区别
2. Boosting的常见树有哪些
3. Xgboost与gbdt相比改进了哪些
4. 介绍一下股票预测的项目
5. 介绍RNN和LSTM
6. LSTM改进了RNN的什么问题？为什么
7. 介绍协同过滤、协同过滤的共现矩阵中有缺失值怎么处理
8. 介绍矩阵分解、矩阵分解改善了协同过滤的哪些问题
9. 了解奇异值分解（svd）吗？
10. 样本不平衡的处理方法
11. 二分类模型有哪些评价指标？为什么不常用准确率
12. 介绍ROC曲线（AUC）、召回率、精确率
13. 编程题：股票买卖的最佳时机（leetcode常规题）
14. **tensorflow2.0和tensorflow1.0的区别**
15. 滴滴（国际化外卖，机器学习）

全是对着简历问，会扣细节，大多数是项目引申出来的问题。

1. embedding技术的embedding是用哪种方法做的？
2. **FM模型的k至关重要，deepfm模型实现该项目的时候有进行对k参数进行调参吗，调参的效果如何？**
3. **FM模型的损失函数是啥？**

交叉熵损失函数（一般分类模型的损失函数都是它）：

即求和，**注意负号（信息熵等也要加负号）**



1. RNN、LSTM、GRU的介绍以及改进
2. 为什么LSTM能解决RNN的梯度消失问题
3. RNN处理序列不一致的问题是如何处理的？
4. 常用的机器学习算法了解哪些？
5. 随机森林和gbdt的区别
6. Xgboost对gbdt的改进
7. **介绍Xgboost对缺失值处理算法：稀疏感知算法（详细见5.1的第六条）**

总的来说即xgboost在构建树的时候不考虑有缺失值的样本，在训练集中有缺失值的样本可以分别算左右子树归到最大的增益中去。并将其设置为默认的分支。如果仅仅在测试集中有缺失值，即没有学习到默认分支，则直接将其归到默认的右结点。

而CART树这类默认处理缺失值时，如果训练某个样本时，某个存在缺失值的特征恰好是当前的分裂增益最大的特征。则需重新遍历其他特征，去选择另外的特征作为当前结点的分类标准。

1. **随机森林和xgboost相关框架关于过拟合有关的参数有哪些，如何调节**

<https://www.cnblogs.com/TimVerion/p/11436001.html>

①随机森林有两个参数：n\_estimators和max\_features。n\_estimators表森林中树的个数，越大越好。Max\_featrues决定每棵树的随机性大小，较小的max\_features可以降低过拟合。

②和gbdt的n\_estimators表树的个数，树越多会容易导致模型过于复杂，即过分要求精度。同时max\_depth对于每个树而言都可以减少复杂度。

③其他模型如逻辑回归、svm默认用正则化参数降低过拟合。

④Xgboost的参数：

与过拟合相关：max\_depth、n\_estimators、subsample（这个参数控制对于每棵树，随机采样的比例，减小这个参数的值，算法会更加保守，避免过拟合）、colsample\_bytree（特征比例，相当于随机森林的max\_feature）。Gamma（指定了节点分裂所需的最小损失函数下降值。 这个参数的值越大，算法越保守，避免过拟合）。

⑤lightgbm参数

Bin的数量越少，越能防止过拟合。

1. Xgboost和lightgbm的区别
2. 逻辑回归损失函数的推导
3. 深度学习中常用的激活函数和特点
4. 神经网络中常用的解决过拟合方法？
5. 说说BN层具体怎么做的？
6. **对Dropout之后对预测模型还需要什么处理？**
7. Svm中软间隔和硬间隔？
8. Svm遇到非线性可分时怎么办？
9. 常用的核函数有哪些？

高斯径向基核函数 RBF、多项式核函数 POLY

1. **高斯核函数有什么特点，什么情况下使用高斯核函数？**

如果特征的数量大到和样本数量差不多，则选用LR或者线性核的SVM；

如果特征的数量小，样本的数量正常，则选用SVM+高斯核函数；

如果特征的数量小，而样本的数量很大，则需要手工添加一些特征从而变成第一种情况

1. 代码题：二叉树中是否存在和为固定值的路径、打印出路径方案的代码怎么写？
2. 海外starmake推荐算法

属实要diss一下，面试前十分钟放我鸽子说要临时开会，叫我额外找时间，我说晚上七点也不回复我行不行，面试连个邮件啥的都没。突然在玩手机被打电话说现在开始面试。什么垃圾玩意。

1. Fm模型的隐向量怎么得到的。

通过引入隐向量（对参数矩阵进行矩阵分解），完成对特征的参数估计

1. kmeans聚类算法迭代过程（注意与KNN区分）。

随机选取k个样本中心、计算每个样本到类中心的距离，将每个样本指派到与其最近的中心的类中。计算各个类中样本的均值，作为新的聚类中心。

1. 协同过滤找到K个相似用户后怎么生成最终的推荐集合。

找到相似的K个用户后，最常用的方式是利用用户相似度和相似用户评价的加权平均获得目标用户的评价预测。

1. 计算两个向量间相似度的常用方法

欧式距离、余弦相似度（夹角越小则余弦相似度越大）、皮尔逊系数（减小了用户评分偏置的影响）。

余弦相似度和欧式距离的效果区别：①欧式距离数值收维度影响很大，面对**文本相似度**这类的计算时，两个文本的内容相似但长度相差很大，此时用余弦相似度更合理。

②欧式距离更体现数值上的差异，而余弦距离更体现在方向上的差异，**比如统计两部剧的用户观看行为**，用户A为（0,1）而用户B为（1,0）则余弦相似度就很小（即夹角过大）。而用户A为（1,10）、用户B为（10,100）则不同。

1. wide&deep中的wide部分相当于什么模型

最后直接输入sigmoid为激活函数的单元，故相当于是逻辑回归模型。同理fm模型最后交叉了特征之后也要送入逻辑回归模型。

1. 清华教授的创业公司（名字属实难记）
2. 大数定律
3. 傅里叶变换
4. Layer normalization

但是有些场景是不能使用BN的，例如batchsize较小或者在RNN中，这时候可以选择使用LN，LN得到的模型更稳定且起到正则化的作用。

不同的是BN取的是不同样本的同一个特征，而LN取的是同一个样本的不同特征。在BN和LN都能使用的场景中，BN的效果一般优于LN，原因是基于不同数据，同一特征得到的归一化特征更不容易损失信息

1. 美团优选（销量预测、推荐算法）
2. fm模型的求解方法有哪些
3. FM模型处理回归问题的方法
4. GRU的两个门的名称

改名了，叫更新门和重置门。

1. 代码题：求根结点到叶子结点数字之和（dfs）
2. 滴滴国际化二面

全是在问项目细节

类别型特征测试集出现没有出现的类别

代码题：链表排序

先写了插入排序，后又被要求写归并排序。

1. 百度搜索策略部

（1）代码题：快排

（2）场景题三道：怎么统计一个省份的人数

**（3）2亿个数怎么找出不重复的数，不可以全部存入内存**

方案1：采用2-Bitmap（每个数分配2bit，00表示不存在，01表示出现一次，10表示多次，11无意义）进行，共需内存内存，还可以接受。然后扫描这2.5亿个整数，查看Bitmap中相对应位，如果是00变01，01变10，10保持不变。所描完事后，查看bitmap，把对应位是01的整数输出即可。

（4）在庞大的数据集中随机取n个数（储水池算法）

（5）聊项目

（6）介绍RNN和LSTM并说明区别

（7）随机森林和gbdt

（8）树模型有关防止过拟合的参数有哪些

（9）代码题：矩阵顺时针旋转

1. 创业公司二面
2. 如何根据每个数有个概率值得到每个数

注意记住C++中关于随机数的api

九、美团优选二面（已收到offer）

全都在聊项目，穿插着机器学习和深度学习知识问。

1. 了解过上亿数据怎么处理怎么做吗？
2. 尝试在电影评分预测的项目上加AUC曲线和ROC等指标

由于模型的准确度较高，并且电影数据集并非不平衡样本。

十、百度二面（我靠全场被虐）

（1）循环有序数组的二分查找

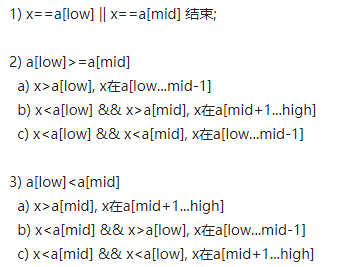
不要想着直接定位到转折点。

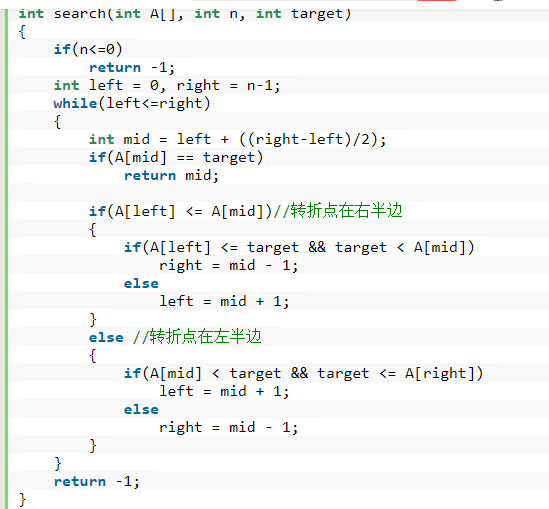
只需要知道转折点在中间点的哪一侧就行。因为不含转折点的一侧一定是单调递增的，依然能够帮助我判断出key在mid的哪一侧。

根据left和right下标，求得mid下标。

如果A[left]<=A[mid]，那么A[mid]一定不在转折点左侧，意味着从left到mid的整个左半部分都是严格递增的，因此我能够判断key是否在左半部分里。

如果A[left]>A[mid]，那么A[mid]一点在转折点左侧，意味着从mid到right的整个右半部分是严格递增的，因此我能够判断key是否在右半部分里。





（2）视频搜索项目设计（怎么设置特征、选什么模型、用什么损失函数）

1. 给你一个项目：svm、gbdt、神经网络三种选一个作为模型，该怎么选
2. Bert了解吗