链滴

机器学习高频面试真题整理

作者：julyedu

原文链接：https://ld246.com/article/1622545058149

来源网站：链滴

51、KMeans算法k值及初始类簇中心点的选取

KMeans算法是最常用的聚类算法，主要思想是:在给定K值和K个初始类簇中心点的情况下，把每个点亦即数据记录)分到离其最近的类簇中心点所代表的类簇中，所有点分配完毕之后，根据一个类簇内的有点重新计算该类簇的中心点(取平均值)，然后再迭代的进行分配点和更新类簇中心点的步骤，直至簇中心点的变化很小，或者达到指定的迭代次数。

KMeans算法本身思想比较简单，但是合理的确定K值和K个初始类簇中心点对于聚类效果的好坏有很的影响。

52、解释对偶的概念

一个优化问题可以从两个角度进行考察，一个是primal 问题，一个是dual 问题，就是对偶问题，一情况下对偶问题给出主问题最优值的下界，在强对偶性成立的情况下由对偶问题可以得到主问题的最下界，对偶问题是凸优化问题，可以进行较好的求解，SVM中就是将primal问题转换为dual问题进行解，从而进一步引入核函数的思想。

53、如何进行特征选择?

特征选择是一个重要的数据预处理过程，主要有两个原因：一是减少特征数量、降维，使模型泛化能更强，减少过拟合;二是增强对特征和特征值之间的理解。

常见的特征选择方式：

1. 去除方差较小的特征

2. 正则化。1正则化能够生成稀疏的模型。L2正则化的表现更加稳定，由于有用的特征往往对应系数零。

3. 随机森林，对于分类问题，通常采用基尼不纯度或者信息增益，对于回归问题，通常采用的是方差者最小二乘拟合。一般不需要feature engineering、调参等繁琐的步骤。它的两个主要问题，1是重的特征有可能得分很低（关联特征问题），2是这种方法对特征变量类别多的特征越有利（偏向问题。

4. 稳定性选择。是一种基于二次抽样和选择算法相结合较新的方法，选择算法可以是回归、SVM或其类似的方法。它的主要思想是在不同的数据子集和特征子集上运行特征选择算法，不断的重复，最终总特征选择结果，比如可以统计某个特征被认为是重要特征的频率（被选为重要特征的次数除以它所的子集被测试的次数）。理想情况下，重要特征的得分会接近100%。稍微弱一点的特征得分会是非0数，而最无用的特征得分将会接近于0。

54、衡量分类器的好坏?

这里首先要知道TP、FN（真的判成假的）、FP（假的判成真）、TN四种（可以画一个表格）。

55、机器学习和统计里面的auc的物理意义是啥?

auc是评价模型好坏的常见指标之一，本题解析来自：

https://www.zhihu.com/question/39840928

分三部分，第一部分是对AUC的基本介绍，包括AUC的定义，解释，以及算法和代码，第二部分用逻回归作为例子来说明如何通过直接优化AUC来训练，第三部分，内容完全由@李大猫原创——如何根auc值来计算真正的类别，换句话说，就是对auc的反向工程。

56、数据预处理

1. 缺失值，填充缺失值fillna：

i. 离散：None,

ii. 连续：均值。

iii. 缺失值太多，则直接去除该列

2. 连续值：离散化。有的模型（如决策树）需要离散值

3. 对定量特征二值化。核心在于设定一个阈值，大于阈值的赋值为1，小于等于阈值的赋值为0。如图操作

4. 皮尔逊相关系数，去除高度相关的列

57、观察增益gain, alpha和gamma越大，增益越小?

xgboost寻找分割点的标准是最大化gain. 考虑传统的枚举每个特征的所有可能分割点的贪心法效率低，xgboost实现了一种近似的算法。大致的思想是根据百分位法列举几个可能成为分割点的候选者然后从候选者中计算Gain按最大值找出最佳的分割点。它的计算公式分为四项, 可以由正则化项参数整(lamda为叶子权重平方和的系数, gama为叶子数量)..

58、什麽造成梯度消失问题?

Yes you should understand backdrop - Andrej Karpathy

How does the ReLu solve the vanishing gradient problem?

神经网络的训练中，通过改变神经元的权重，使网络的输出值尽可能逼近标签以降低误差值，训练普使用BP算法，核心思想是，计算出输出与标签间的损失函数值，然后计算其相对于每个神经元的梯度进行权值的迭代。

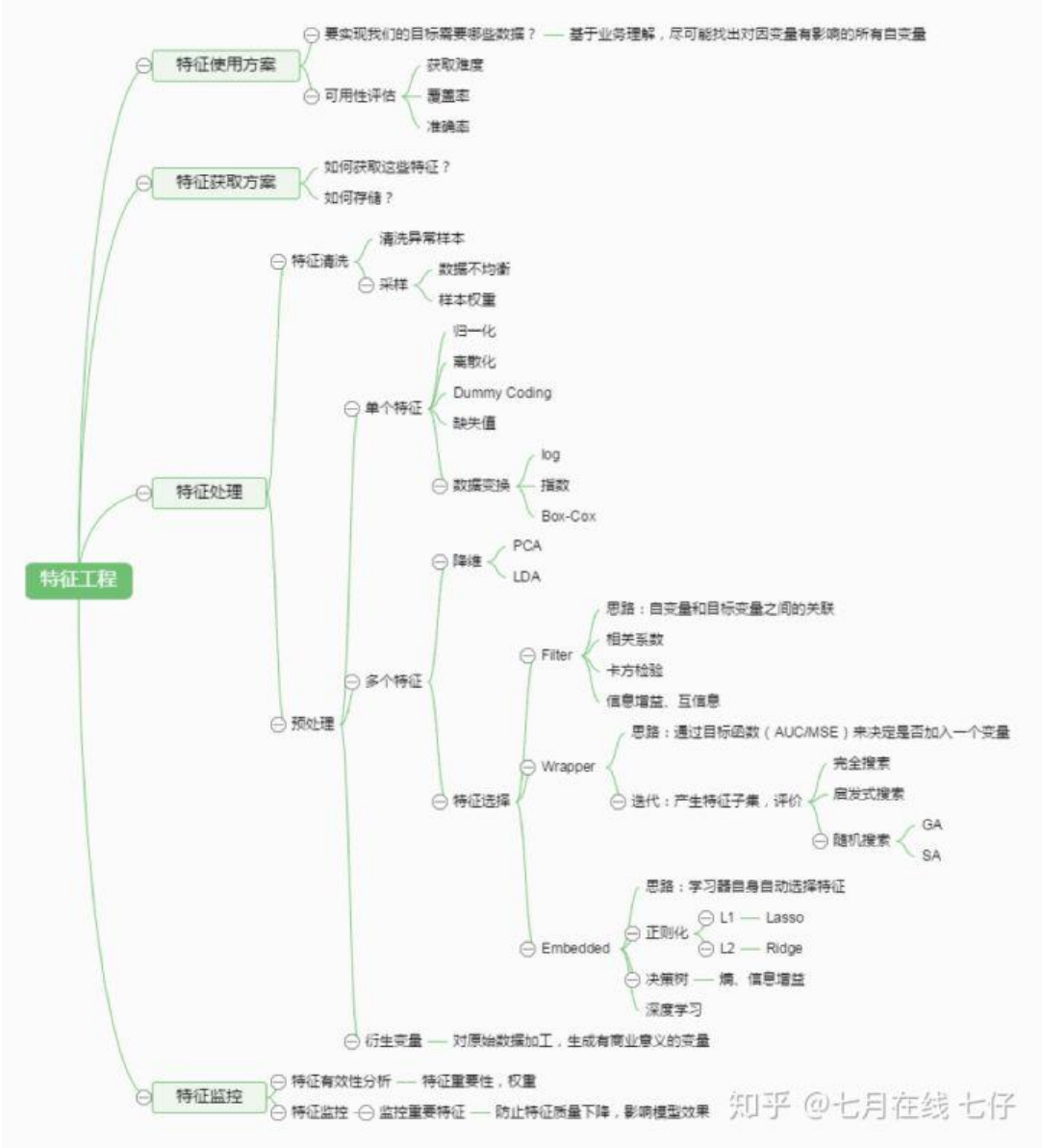
梯度消失会造成权值更新缓慢，模型训练难度增加。造成梯度消失的一个原因是，许多激活函数将输值挤压在很小的区间内，在激活函数两端较大范围的定义域内梯度为0，造成学习停止。

59、到底什么是特征工程?

首先，大多数机器学习从业者主要在公司做什么呢？不是做数学推导，也不是发明多高大上的算法，是做特征工程，如下图所示（图来自：

http://www.julyedu.com/video/play/18)

60、你知道有哪些数据处理和特征工程的处理?



61、准备机器学习面试应该了解哪些理论知识?

原文链接：机器学习高频面试真题整理

1. 【理论功底】主要考察对机器学习模型的理解，***选择性提问***（如果遇到面试者的研究方向自己不了解但感兴趣，会很欣喜，可以趁机学习一个哈哈）这块儿的问题会比较细碎，都是我自己深入思考过的（***背书是没用的，这里任何一个点我都可以给你展开问下去***），在此**全部手敲**
    1. 过拟合欠拟合（举几个例子让判断下，顺便问问交叉验证的目的、超参数搜索方法、EarlyStopping）、L1正则和L2正则的做法、正则化背后的思想（顺便问问BatchNorm、Covariance Shift）、L1正则产生稀疏解原理、逻辑回归为何线性模型（顺便问问LR如何解决低维不可分、从图模型角度看LR和朴素贝叶斯和无监督）、几种参数估计方法MLE/MAP/贝叶斯的联系和区别、简单说下SVM的支持向量（顺便问问KKT条件、为何对偶、核的通俗理解）、 GBDT随机森林能否并行（顺便问问bagging boosting）、 生成模型判别模型举个例子、聚类方法的掌握（顺便问问Kmeans的EM推导思路、谱聚类和Graph-cut的理解）、梯度下降类方法和牛顿类方法的区别（顺便问问Adam、L-BFGS的思路）、半监督的思想（顺便问问一些特定半监督算法是如何利用无标签数据的、从MAP角度看半监督）、常见的分类模型的评价指标（顺便问问交叉熵、ROC如何绘制、AUC的物理含义、类别不均衡样本）
    2. CNN中卷积操作和卷积核作用、maxpooling作用、卷积层与全连接层的联系、梯度爆炸和消失的概念（顺便问问神经网络权值初始化的方法、为何能减缓梯度爆炸消失、CNN中有哪些解决办法、LSTM如何解决的、如何梯度裁剪、dropout如何用在RNN系列网络中、dropout防止过拟合）、为何卷积可以用在图像/语音/语句上（顺便问问channel在不同类型数据源中的含义）
    3. 如果面试者跟我一样做NLP、推荐系统，我会继续追问 CRF跟逻辑回归 最大熵模型的关系、CRF的优化方法、CRF和MRF的联系、HMM和CRF的关系（顺便问问 朴素贝叶斯和HMM的联系、LSTM+CRF 用于序列标注的原理、CRF的点函数和边函数、CRF的经验分布）、WordEmbedding的几种常用方法和原理（顺便问问language model、perplexity评价指标、word2vec跟Glove的异同）、topic model说一说、为何CNN能用在文本分类、syntactic和semantic问题举例、常见Sentence embedding方法、注意力机制（顺便问问注意力机制的几种不同情形、为何引入、seq2seq原理）、序列标注的评价指标、语义消歧的做法、常见的跟word有关的特征、factorization machine、常见矩阵分解模型、如何把分类模型用于商品推荐（包括数据集划分、模型验证等）、序列学习、wide&deep model（顺便问问为何wide和deep)
2. 【**代码能力**】主要考察实现算法和优化代码的能力，我一般会先看面试者的github repo（如果简历给出来），看其代码风格、架构能力（**遇到大神会认真学习一个哈哈**），如果没有github，我会避免问典型的应试题，而是问一些 我本人从实际问题中抽象出的小算法题，比如：
    1. 给出节点的矩阵和边的矩阵，求路径和最大的路径（来源于 Viterbi 算法，本质就是个动态规划），至少给个思路和伪代码（顺便聊聊前向传播和反向传播）
    2. 给出一数组，数组元素是pair对儿，表示一个有向无环图的<父亲节点, 孩子节点>，用最优的方法，将其变成一个新的有序数组，数组元素是该有向无环图所有节点，数组的有序性体现在：父亲节点在孩子节点前面（来源于 贝叶斯网络实现时的小trick）
3. 【**项目能力**】主要考察解决实际问题的思路、填坑能力，这部分最考验面试官功底，要能从面试者浮夸的描述中寻找有意义的点，并一步步深挖。另外很多dirty work(数据预处理、文本清

62、数据不平衡问题

这主要是由于数据分布不平衡造成的。解决方法如下：

采样，对小样本加噪声采样，对大样本进行下采样

数据生成，利用已知样本生成新的样本

进行特殊的加权，如在Adaboost中或者SVM中

采用对不平衡数据集不敏感的算法

改变评价标准：用AUC/ROC来进行评价
采用Bagging/Boosting/ensemble等方法
在设计模型的时候考虑数据的先验分布

63、特征比数据量还大时，选择什么样的分类器?

线性分类器，因为维度高的时候，数据一般在维度空间里面会比较稀疏，很有可能线性可分。

64、常见的分类算法有哪些？他们各自的优缺点是什么?

贝叶斯分类法
优点：
1）所需估计的参数少，对于缺失数据不敏感。
2）有着坚实的数学基础，以及稳定的分类效率。
缺点：
1）假设属性之间相互独立，这往往并不成立。（喜欢吃番茄、鸡蛋，却不喜欢吃番茄炒蛋）。
2）需要知道先验概率。
3）分类决策存在错误率。

65、常见的监督学习算法有哪些?

感知机、svm、人工神经网络、决策树、逻辑回归

作者：七月在线 七仔
链接：
https://zhuanlan.zhihu.com/p/217494137
来源：知乎

66、说说常见的优化算法及其优缺点?

1）随机梯度下降
优点：容易陷入局部最优解
缺点：收敛速度较快
2）批量梯度下降
优点：可以一定程度上解决局部最优解的问题

67、特征向量的归一化方法有哪些?

线性函数转换，表达式如下：
y=(x-MinValue)/(MaxValue-MinValue)
对数函数转换，表达式如下：

y=log10 (x)

反余切函数转换，表达式如下：

y=arctan(x)*2/PI

减去均值，除以标准差：

y=(x-means)/ Standard Deviation

68、RF与GBDT之间的区别与联系?

1）相同点：都是由多棵树组成，最终的结果都是由多棵树一起决定。

2）不同点：

a 组成随机森林的树可以分类树也可以是回归树，而GBDT只由回归树组成；

b 组成随机森林的树可以并行生成，而GBDT是串行生成

c 随机森林的结果是多数表决表决的，而GBDT则是多棵树累加之和

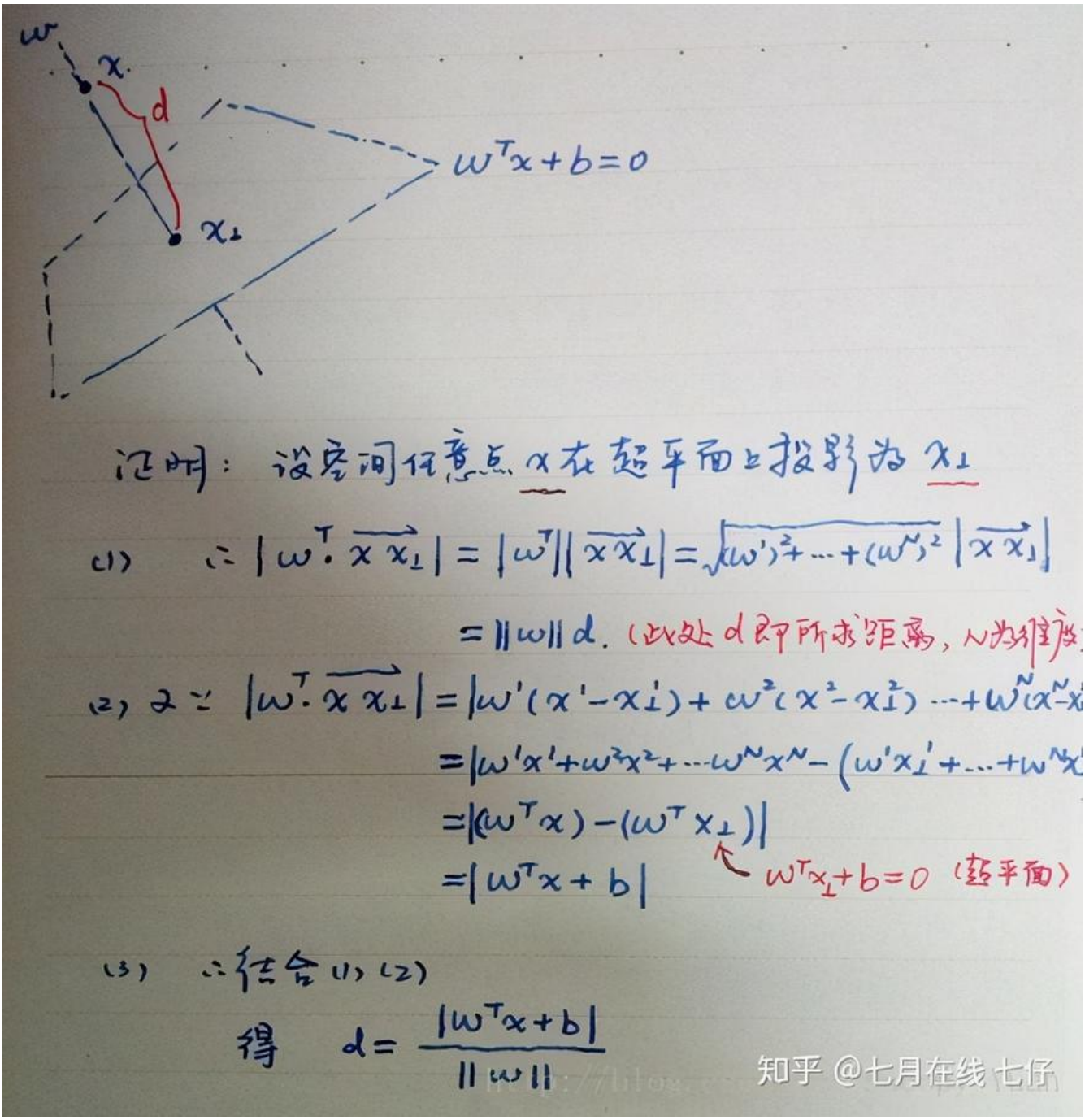
d 随机森林对异常值不敏感，而GBDT对异常值比较敏感

e 随机森林是减少模型的方差，而GBDT是减少模型的偏差

f GBDT的会累加所有树的结果，而这种累加是无法通过分类完成的，因此GBDT的树都是CART回归，而不是分类树（尽管GBDT调整后也可以用于分类但不代表GBDT的树为分类树）

69、试证明样本空间中任意点 x 到超平面 (w,b) 的距离式

$$r = \frac{|w^T x + b|}{||w||} \qquad (西瓜书, 6.2)$$



70、请比较下EM算法、HMM、CRF

这三个放在一起不是很恰当，但是有互相有关联，所以就放在这里一起说了。注意重点关注算法的思
。

（1）EM算法

EM算法是用于含有隐变量模型的极大似然估计或者极大后验估计，有两步组成：E步，求期望（e pectation）；M步，求极大（maxmization）。本质上EM算法还是一个迭代算法，通过不断用上一参数对隐变量的估计来对当前变量进行计算，直到收敛。

注意：EM算法是对初值敏感的，而且EM是不断求解下界的极大化逼近求解对数似然函数的极大的算法，也就是说EM算法不能保证找到全局最优值。对于EM的导出方法也应该掌握。

71、带核的SVM为什么能分类非线性问题?

核函数的本质是两个函数的内积，通过核函数将其隐射到高维空间，在高维空间非线性问题转化为线问题, SVM得到超平面是高维空间的线性分类平面。

72、请说说常用核函数及核函数的条件

我们通常说的核函数指的是正定和函数，其充要条件是对于任意的x属于X，要求K对应的Gram矩阵是半正定矩阵。RBF核径向基，这类函数取值依赖于特定点间的距离，所以拉普拉斯核其实也是径向核。SVM关键是选取核函数的类型，常用核函数主要有线性内核，多项式内核，径向基内核（RBF）sigmoid核。

73、请具体说说Boosting和Bagging的区别

（1） Bagging之随机森林　　随机森林改变了决策树容易过拟合的问题，这主要是由两个操作所优的：　　1）Boostrap从袋内有放回的抽取样本值　　2）每次随机抽取一定数量的特征（通常为sqrn））。　　分类问题：采用Bagging投票的方式选择类别频次最高的　　回归问题：直接取每颗树果的平均值。

74、逻辑回归相关问题

（1）公式推导一定要会

（2）逻辑回归的基本概念　　这个最好从广义线性模型的角度分析，逻辑回归是假设y服从Bernoull分布。

（3）L1-norm和L2-norm　　其实稀疏的根本还是在于L0-norm也就是直接统计参数不为0的个数为规则项，但实际上却不好执行于是引入了L1-norm；而L1norm本质上是假设参数先验是服从Laplac分布的，而L2-norm是假设参数先验为Gaussian分布，我们在网上看到的通常用图像来解答这个问的原理就在这。　　但是L1-norm的求解比较困难，可以用坐标轴下降法或是最小角回归法求解。

（4）LR和SVM对比　　首先，LR和SVM最大的区别在于损失函数的选择，LR的损失函数为Log损（或者说是逻辑损失都可以）、而SVM的损失函数为hinge loss。　　其次，两者都是线性模型。　最后，SVM只考虑支持向量（也就是和分类相关的少数点）

（5）LR和随机森林区别　　随机森林等树算法都是非线性的，而LR是线性的。LR更侧重全局优化而树模型主要是局部的优化。

… …

75、什么是共线性, 跟过拟合有什么关联?

共线性：多变量线性回归中，变量之间由于存在高度相关关系而使回归估计不准确。 共线性会造成冗，导致过拟合。 解决方法：排除变量的相关性 / 加入权重正则。