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

<h2 id="添加微信-julyedufu77-回复--11---领取最新升级版-名企AI面试100题-电子书--">添加微信：julyedufu77，回复“11”，领取最新升级版《名企 AI 面试 100 题》电子书！！</h2>
<h2 id="51-KMeans算法k值及初始类簇中心点的选取">51、KMeans 算法 k 值及初始类簇中心点选取</h2>
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，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，根据一个簇内的所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，直至类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很大的影响。
<h2 id="52-解释对偶的概念">52、解释对偶的概念</h2>
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，一般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的最优下界，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 问题进行求解，从而进一步引入核函数的思想。
<h2 id="53-如何进行特征选择-">53、如何进行特征选择？</h2>
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，使模型泛能力更强，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：

去除方差较小的特征
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非零。
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，通常采用的是方或者最小二乘拟合。一般不需要 feature engineering、调参等繁琐的步骤。它的两个主要问题，1 重要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（偏向题）。
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 其他类似的方法。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，终汇总特征选择结果，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所在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非 0 的数，而最无用的特征得分将会接近于 0。

<h2 id="54-衡量分类器的好坏-">54、衡量分类器的好坏？</h2>
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）
<h2 id="55-机器学习和统计里面的auc的物理意义是啥-">55、机器学习和统计里面的 auc 的物理义是啥？</h2>
auc 是评价模型好坏的常见指标之一，本题解析来自：
<a href="https://ld246.com/forward?goto=https%3A%2F%2Fwww.zhihu.com%2Fquestion%239840928" target="_blank" rel="nofollow ugc">https://www.zhihu.com/question/39840928<a>
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，第二部分逻辑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—如何根据 auc 值来计算真正的类别，换句话说，就是对 auc 的反向工程。
<h2 id="56-数据预处理">56、数据预处理</h2>

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
连续值：离散化。有的模型（如决策树）需要离散值
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。图像操作

皮尔逊相关系数，去除高度相关的列

<h2 id="57-观察增益gain--alpha和gamma越大-增益越小-">57、观察增益 gain, alpha 和 gamma 越大，增益越小？ </h2>
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率太低，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, 可以由正则化项数调整(lamda 为叶子权重平方和的系数, gama 为叶子数量)..
<h2 id="58-什麽造成梯度消失问题-">58、什麽造成梯度消失问题?</h2>
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，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。
<h2 id="59-到底什么是特征工程-">59、到底什么是特征工程？ </h2>
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，而是做特征工程，如下图所示（图来自：
http://www.julyedu.com/video/play/18

<h2 id="60-你知道有哪些数据处理和特征工程的处理-">60、你知道有哪些数据处理和特征工程的理？ </h2>

<h2 id="61-准备机器学习面试应该了解哪些理论知识-">61、准备机器学习面试应该了解哪些理论识？ </h2>

<h2 id="62-数据不平衡问题">62、数据不平衡问题</h2>
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
<h2 id="63-特征比数据量还大时-选择什么样的分类器-">63、特征比数据量还大时，选择什么样的类器？ </h2>
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。</
>
<h2 id="64-常见的分类算法有哪些-他们各自的优缺点是什么-">64、常见的分类算法有哪些？他们自的优缺点是什么？ </h2>
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。
<h2 id="65-常见的监督学习算法有哪些-">65、常见的监督学习算法有哪些？</h2>
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
著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。
<h2 id="66-说说常见的优化算法及其优缺点-">66、说说常见的优化算法及其优缺点？</h2>
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
<h2 id="67-特征向量的归一化方法有哪些-">67、特征向量的归一化方法有哪些？</h2>
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
<h2 id="68-RF与GBDT之间的区别与联系-">68、RF 与 GBDT 之间的区别与联系？</h2>
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 回树，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）
<h2 id="69-试证明样本空间中任意点-x-到超平面--w-b--的距离公式">69、试证明样本空间中任意 x 到超平面 (w,b) 的距离公式</h2>

<h2 id="70-请比较下EM算法-HMM-CRF">70、请比较下 EM 算法、HMM、CRF</h2>
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。注意重点关注算法思想。
（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（xpectation）；M 步，求极大（maxmization）。本质上 EM 算法还是一个迭代算法，通过不断用一代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化的算法，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。
<h2 id="71-带核的SVM为什么能分类非线性问题-">71、带核的 SVM 为什么能分类非线性问题？</h2>

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线性问题, SVM 得到超平面是高维空间的线性分类平面。
<h2 id="72-请说说常用核函数及核函数的条件">72、请说说常用核函数及核函数的条件</h2>
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 Gra 矩阵要是半正定矩阵。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是径向基核。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（RBF），sigmoid 核。
<h2 id="73-请具体说说Boosting和Bagging的区别">73、请具体说说 Boosting 和 Bagging 的区</h2>
（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，这主要是由两个操所优化的：　　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（通为 sqr(n)）。　　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：直接取颗树结果的平均值。
<h2 id="74-逻辑回归相关问题">74、逻辑回归相关问题</h2>
（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 Be noulli 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 的个数作为规则项，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服从 Laplace 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来解答这个问题的原理就在这。　　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 损失（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 更侧重全优化，而树模型主要是局部的优化。
... ...
<h2 id="75-什么是共线性--跟过拟合有什么关联-">75、什么是共线性, 跟过拟合有什么关联?</h2>
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。 共线性会成冗余，导致过拟合。 解决方法：排除变量的相关性 / 加入权重正则。
<h2 id="添加微信-julyedufu77-回复--11---领取最新升级版-名企AI面试100题-电子书---">添加微：julyedufu77，回复 “11 ”，领取最新升级版《名企 AI 面试 100 题》电子书！！ </h2>