精选机器学习面试题，斩获一线互联网公司机器学习岗 offer

作者：julyedu
原文链接：https://ld246.com/article/1622629907990
来源网站：链滴
许可协议：署名-相同方式共享 4.0 国际 (CC BY-SA 4.0)

评论有奖：评论区回复“11”，领取最新升级版《名企AI面试100题》电子书！

76、机器学习中，有哪些特征选择的工程方法?

目录
1 特征工程是什么?
2 数据预处理
2.1 无量纲化
2.1.1 标准化
2.1.2 区间缩放法
2.1.3 标准化与归一化的区别
2.2 对定量特征二值化
2.3 对定性特征哑编码
2.4 缺失值计算
2.5 数据变换
2.6 回顾
3 特征选择
3.1 Filter
3.1.1 方差选择法
3.1.2 相关系数法
3.1.3 卡方检验
3.1.4 互信息法
3.2 Wrapper
3.2.1 递归特征消除法
3.3 Embedded
3.3.1 基于惩罚项的特征选择法
3.3.2 基于树模型的特征选择法
3.4 回顾
4 降维
4.1 主成分分析法（PCA）
4.2 线性判别分析法（LDA）
4.3 回顾
5 总结
6 参考资料
知乎 @七月在线 七仔

77、用贝叶斯机率说明Dropout的原理

回想一下使用Bagging学习,我们定义 k 个不同的模型,从训练集有替换采样 构造 k 个不同的数据集,然在训练集上训练模型 i。 Dropout的目标是在指数 级数量的神经网络上近似这个过程。Dropout训练Bagging训练不太一样。在Bagging的情况下,所有模型是独立 的。 在Dropout的情况下,模型是共享数的,其中每个模型继承的父神经网络参 数的不同子集。参数共享使得在有限可用的内存下代表指数量的模型变得可能。

在Bagging的情况下,每一个模型在其相应训练集上训练到收敛。 在Dropout的情况下,通常大部分模都没有显式地被训练,通常该模型很大,以致到宇宙毁灭都不 能采样所有可能的子网络。取而代之的是,能的子网络的一小部分训练单个步骤,参数共享导致剩余的子网络能有好的参数设定。

78、对于维度极低的特征，选择线性还是非线性分类器?

非线性分类器，低维空间可能很多特征都跑到一起了，导致线性不可分。 1. 如果Feature的数量很大跟样本数量差不多，这时候选用LR或者是Linear Kernel的SVM 2. 如果Feature的数量比较小，样本量一般，不算大也不算小，选用SVM+Gaussian Kernel 3. 如果Feature的数量比较小，而样本数量多，需要手工添加一些feature变成第一种情况。

79、请问怎么处理特征向量的缺失值

一方面，缺失值较多.直接将该特征舍弃掉，否则可能反倒会带入较大的noise，对结果造成不良影响。另一方面缺失值较少,其余的特征缺失值都在10%以内，我们可以采取很多的方式来处理: 1) 把NaN直作为一个特征，假设用0表示； 2) 用均值填充； 3) 用随机森林等算法预测填充。

80、SVM、LR、决策树的对比

模型复杂度：SVM支持核函数，可处理线性非线性问题;LR模型简单，训练速度快，适合处理线性问题决策树容易过拟合，需要进行剪枝 损失函数：SVM hinge loss; LR L2 logistical loss（对数似然损）; adaboost 指数损失 数据敏感度：SVM添加容忍度对outlier不敏感，只关心支持向量，且需要先归一化; LR对异常点敏感 数据量：数据量大就用LR，数据量小且特征少就用SVM非线性核。

81、简述KNN最近邻分类算法的过程?

1. 计算测试样本和训练样本中每个样本点的距离（常见的距离度量有欧式距离，马氏距离等）；
2. 对上面所有的距离值进行排序；
3. 选前 k 个最小距离的样本；
4. 根据这 k 个样本的标签进行投票，得到最后的分类类别；

82、常用的聚类划分方式有哪些？列举代表算法。

1. 基于划分的聚类:K-means，k-medoids，CLARANS。
2. 基于层次的聚类：AGNES（自底向上），DIANA（自上向下），BIRCH(CF-Tree)，。
3. 基于密度的聚类：DBSACN，OPTICS，CURE。
4. 基于网格的方法：STING，WaveCluster。
5. 基于模型的聚类：EM,SOM，COBWEB。

83、什么是偏差与方差?

泛化误差可以分解成偏差的平方加上方差加上噪声。偏差度量了学习算法的期望预测和真实结果的偏程度，刻画了学习算法本身的拟合能力，方差度量了同样大小的训练集的变动所导致的学习性能的变，刻画了数据扰动所造成的影响，噪声表达了当前任务上任何学习算法所能达到的期望泛化误差下界刻画了问题本身的难度。偏差和方差一般称为bias和variance，一般训练程度越强，偏差越小，方差大，泛化误差一般在中间有一个最小值，如果偏差较大，方差较小，此时一般称为欠拟合，而偏差较，方差较大称为过拟合。

84、解决bias和Variance问题的方法是什么?

High bias解决方案:Boosting、复杂模型（非线性模型、增加神经网络中的层）、更多特征

High Variance解决方案：bagging、简化模型、降维

具体而言：

高偏差, 可以用boosting模型, 对预测残差进行优化, 直接降低了偏差. 也可以用高模型容量的复杂模型比如非线性模型, 深度神经网络), 更多的特征, 来增加对样本的拟合度.

高方差, 一般使用平均值法, 比如bagging, 或者模型简化/降维方法, 来降低方差。

85、采用 EM 算法求解的模型有哪些，为什么不用牛顿或梯度下降法？

用EM算法求解的模型一般有GMM或者协同过滤，k-means其实也属于EM。EM算法一定会收敛，是可能收敛到局部最优。由于求和的项数将随着隐变量的数目指数上升，会给梯度计算带来麻烦。

86、xgboost怎么给特征评分？

我们知道，在训练的过程中，cart树通过Gini指数选择分离点的特征，一个特征被选中的次数越多，么该特征评分越高。

但xgboost呢？对于一个叶子节点如何进行分裂，xgboost作者在其原始论文中给出了两种分裂节点方法

（1）枚举所有不同树结构的贪心法

（2）近似算法

87、什么是OOB？随机森林中OOB是如何计算的，它有么优缺点？

bagging方法中Bootstrap每次约有1/3的样本不会出现在Bootstrap所采集的样本集合中，当然也就有参加决策树的建立，把这1/3的数据称为袋外数据oob（out of bag）,它可以用于取代测试集误差计方法。

袋外数据(oob)误差的计算方法如下：

对于已经生成的随机森林,用袋外数据测试其性能,假设袋外数据总数为O,用这O个袋外数据作为输入,进之前已经生成的随机森林分类器,分类器会给出O个数据相应的分类,因为这O条数据的类型是已知的,用正确的分类与随机森林分类器的结果进行比较,统计随机森林分类器分类错误的数目,设为X,则袋外据误差大小=X/O;

优点：这已经经过证明是无偏估计的,所以在随机森林算法中不需要再进行交叉验证或者单独的测试集获取测试集误差的无偏估计。

缺点：当数据量较小时，自助法产生的数据集改变了初始数据集的分布，这会引入估计偏差。

88、推导朴素贝叶斯分类 P(c|d)，文档 d（由若干 word 组成），求该文档属于类别 c 的概率，并说明公式中哪概率可以利用训练集计算得到

根据贝叶斯公式P(c|d)=（P(c)P(d|c)/P(d)）

这里，分母P(d)不必计算，因为对于每个类都是相等的。 分子中，P(c)是每个类别的先验概率，可以训练集直接统计，P(d|c)根据独立性假设，可以写成如下 P(d|c)=￥P(wi|c)（￥符号表示对d中每个词在c类下概率的连乘），P(wi|c)也可以从训练集直接统计得到。 至此，对未知类别的d进行分类时，别为c=argmaxP(c)￥P(wi|c)。

89、请写出你了解的机器学习特征工程操作，以及它的意义

特征工程包括数据与特征处理、特征选择和降维三部分。

数据与特征处理包括：

1.数据选择、清洗、采样

- 数据格式化；
- 数据清洗，填充缺失值、去掉脏数据，将不可信的样本丢掉，缺省值极多的字段考虑不用；
- 采样：针对正负样本不平衡的情况，当正样本远大于负样本时，且量都很大时，使用下采样，量不时，可采集更多的数据或oversampling或修改损失函数；采样过程中可利用分层抽样保持不同类别据的比例。

90、请写出你对VC维的理解和认识

VC维是模型的复杂程度，模型假设空间越大，VC维越高。某种程度上说，VC维给机器学习可学性提了理论支撑。

1. 测试集合的loss是否和训练集合的loss接近？VC维越小，理论越接近，越不容易overfitting。
2. 训练集合的loss是否足够小？VC维越大，loss理论越小，越不容易underfitting。

我们对模型添加的正则项可以对模型复杂度(VC维)进行控制，平衡这两个部分。

91、kmeans聚类中，如何确定k的大小

这是一个老生常谈的经典问题，面试中也经常问。

K-均值聚类算法首先会随机确定k个中心位置，然后将各个数据项分配给最临近的中心点。待分配完之后，聚类中心就会移到分配给该聚类的所有节点的平均位置处，然后整个分配过程重新开始。这一程会一直重复下去，直到分配过程不再产出变化为止。

92、请用Python实现下线性回归，并思考下更高效的实方式

在数学中，线性规划 (Linear Programming，简称LP) 问题是目标函数和约束条件都是线性的最优化题。

线性规划是最优化问题中的重要领域之一。很多运筹学中的实际问题都可以用线性规划来表述。线性划的某些特殊情况，例如网络流、多商品流量等问题，都被认为非常重要，并有大量对其算法的专门究。很多其他种类的最优化问题算法都可以分拆成线性规划子问题，然后求得解。

在历史上，由线性规划引申出的很多概念，启发了最优化理论的核心概念，诸如“对偶”、“分解”“凸性”的重要性及其一般化等。同样的，在微观经济学和商业管理领域，线性规划被大量应用于解收入极大化或生产过程的成本极小化之类的问题。

93、怎么理解“机器学习的各种模型与他们各自的损失函一一对应？

首先你要明确 超参数 和 参数 的差别，超参数通常是你为了定义模型，需要提前敲定的东西(比如多式拟合的最高次数，svm选择的核函数)，参数是你确定了超参数(比如用最高3次的多项式回归)，学到的参数(比如多项式回归的系数)

另外可以把机器学习视作 表达 + 优化，其中表达的部分，各种模型会有各种不同的形态(线性回归 逻回归 SVM 树模型)，但是确定了用某个模型(比如逻辑回归)去解决问题，你需要知道当前模型要达到好的效果，优化方向在哪，这个时候就要借助损失函数了。

94、给你一个有1000列和1百万行的训练数据集。这个数据集是基于分类问题的。经理要求你来降低该数据集的维以减少模型计算时间。你的机器内存有限。你会怎么做？（你可以自由做各种实际操作假设）

答：你的面试官应该非常了解很难在有限的内存上处理高维的数据。以下是你可以使用的处理方法：

1.由于我们的RAM很小，首先要关闭机器上正在运行的其他程序，包括网页浏览器，以确保大部分内可以使用。

2.我们可以随机采样数据集。这意味着，我们可以创建一个较小的数据集，比如有1000个变量和30万，然后做计算。

3.为了降低维度，我们可以把数值变量和分类变量分开，同时删掉相关联的变量。对于数值变量，我将使用相关性分析。对于分类变量，我们可以用卡方检验。

4.另外，我们还可以使用PCA（主成分分析），并挑选可以解释在数据集中有最大偏差的成分。

5.利用在线学习算法，如VowpalWabbit（在Python中可用）是一个可能的选择。

6.利用Stochastic GradientDescent（随机梯度下降）法建立线性模型也很有帮助。

7.我们也可以用我们对业务的理解来估计各预测变量对响应变量的影响大小。但是，这是一个主观的法，如果没有找出有用的预测变量可能会导致信息的显著丢失。

注意：对于第4和第5点，请务必阅读有关在线学习算法和随机梯度下降法的内容。这些是高阶方法。

95、问2：在PCA中有必要做旋转变换吗？如果有必要，什么？如果你没有旋转变换那些成分，会发生什么情况？

答：是的，旋转（正交）是必要的，因为它把由主成分捕获的方差之间的差异最大化。这使得主成分容易解释。但是不要忘记我们做PCA的目的是选择更少的主成分（与特征变量个数相较而言），那些上的主成分能够解释数据集中最大方差。

通过做旋转，各主成分的相对位置不发生变化，它只能改变点的实际坐标。如果我们没有旋转主成分PCA的效果会减弱，那样我们会不得不选择更多个主成分来解释数据集里的方差。

注意：对PCA（主成分分析）需要了解更多。

96、给你一个数据集，这个数据集有缺失值，且这些缺失分布在离中值有1个标准偏差的范围内。百分之多少的数不会受到影响？为什么？

答：这个问题给了你足够的提示来开始思考！由于数据分布在中位数附近，让我们先假设这是一个正分布。

我们知道，在一个正态分布中，约有68%的数据位于跟平均数（或众数、中位数）1个标准差范围内，那样剩下的约32%的数据是不受影响的。

因此，约有32%的数据将不受到缺失值的影响。

97、给你一个癌症检测的数据集，你已经建好了分类模型取得了96%的精度。为什么你还是不满意你的模型性能？可以做些什么呢？

如果你分析过足够多的数据集，你应该可以判断出来癌症检测结果是不平衡数据。在不平衡数据集中精度不应该被用来作为衡量模型的标准，因为96%（按给定的）可能只有正确预测多数分类，但我们兴趣是那些少数分类（4%），是那些被诊断出癌症的人。

因此，为了评价模型的性能，应该用灵敏度（真阳性率），特异性（真阴性率），F值用来确定这个类器的“聪明”程度。如果在那4%的数据上表现不好，我们可以采取以下步骤：

1.我们可以使用欠采样、过采样或SMOTE让数据平衡。

2.我们可以通过概率验证和利用AUC-ROC曲线找到最佳阀值来调整预测阀值。

3.我们可以给分类分配权重，那样较少的分类获得较大的权重。

4.我们还可以使用异常检测。

注意：要更多地了解不平衡分类

98、解释朴素贝叶斯算法里面的先验概率、似然估计和边似然估计？

先验概率就是因变量（二分法）在数据集中的比例。这是在你没有任何进一步的信息的时候，是对分能做出的最接近的猜测。

例如，在一个数据集中，因变量是二进制的（1和0）。例如，1（垃圾邮件）的比例为70%和0（非圾邮件）的为30%。因此，我们可以估算出任何新的电子邮件有70%的概率被归类为垃圾邮件。

似然估计是在其他一些变量的给定的情况下，一个观测值被分类为1的概率。例如，“FREE”这个词以前的垃圾邮件使用的概率就是似然估计。边际似然估计就是，“FREE”这个词在任何消息中使用的率

99、你正在一个时间序列数据集上工作。经理要求你建立一个高精度的模型。你开始用决策树算法，因为你知道它在有类型数据上的表现都不错。后来，你尝试了时间序列回模型，并得到了比决策树模型更高的精度。这种情况会发吗？为什么

众所周知，时间序列数据有线性关系。另一方面，决策树算法是已知的检测非线性交互最好的算法。

为什么决策树没能提供好的预测的原因是它不能像回归模型一样做到对线性关系的那么好的映射。

因此，我们知道了如果我们有一个满足线性假设的数据集，一个线性回归模型能提供强大的预测。

100、给你分配了一个新的项目，是关于帮助食品配送公司节省更多的钱。问题是，公司的送餐队伍没办法准时送餐结果就是他们的客户很不高兴。最后为了使客户高兴，他只好以免餐费了事。哪个机器学习算法能拯救他们？

你的大脑里可能已经开始闪现各种机器学习的算法。但是等等！这样的提问方式只是来测试你的机器习基础。这不是一个机器学习的问题，而是一个路径优化问题。

机器学习问题由三样东西组成：

1.模式已经存在。

2.不能用数学方法解决（指数方程都不行）。

3.有相关的数据。

评论有奖：评论区回复“11”，领取最新升级版《名企AI面试100题》电子书！！