链滴

数据分析转岗 AI 薪资翻 3 倍多 | 机器学习面试都问些什么?

作者：julyedu
原文链接：https://ld246.com/article/1621416129311
来源网站：链滴

<h6 id="添加微信-julyedukefu14-回复-11-领取最新升级版-名企AI面试100题-">添加微信：julyedukefu14，回复【11】领取最新升级版【名企 AI 面试 100 题】</h6>
1、请详细说说支持向量机（support vector machine，SVM）的原理
支持向量机，因其英文名为 support vector machine，故一般简称 SVM，通俗来讲，它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化，终可转化为一个凸二次规划问题的求解。
2、哪些机器学习算法不需要做归一化处理？
在实际应用中，需要归一化的模型：
1.基于距离计算的模型：KNN。
2.通过梯度下降法求解的模型：线性回归、逻辑回归、支持向量机、神经网络。
但树形模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率，决策树、随机森林(Random Forest)。
3、树形结构为什么不需要归一化？
因为数值缩放不影响分裂点位置，对树模型的结构不造成影响。
按照特征值进行排序的，排序的顺序不变，那么所属的分支以及分裂点就不会有不同。而且，树模型不能进行梯度下降的，因为构建树模型（回归树）寻找最优点时是通过寻找最优分裂点完成的，因此模型是阶跃的，阶跃点是不可导的，并且求导没意义，也就不需要归一化。
4、在 k-means 或 kNN，我们常用欧氏距离来计算最近的邻居之间的距离，有时也曼哈顿距离，请对比下这两种距离的差别
欧氏距离，最常见的两点之间或多点之间的距离表示法，又称之为欧几里得度量，它定义于欧几得空间中..
5、数据归一化（或者标准化，注意归一化和标准化不同）的原因
能不归一化最好不归一化，之所以进行数据归一化是因为各维度的量纲不相同。而且需要看情况行归一化。
有些模型在各维度进行了不均匀的伸缩后，最优解与原来不等价（如 SVM）需要归一化。
有些模型伸缩有与原来等价，如：LR 则不用归一化，但是实际中往往通过迭代求解模型参数，如果标函数太扁（想象一下很扁的高斯模型）迭代算法会发生不收敛的情况，所以最好进行数据归一化。
6、请简要说说一个完整机器学习项目的流程
抽象成数学问题
明确问题是进行机器学习的第一步。机器学习的训练过程通常都是一件非常耗时的事情，胡乱尝试时成本是非常高的。
这里的抽象成数学问题，指的我们明确我们可以获得什么样的数据，目标是一个分类还是回归或者是类的问题，如果都不是的话，如果划归为其中的某类问题。
获取数据
数据决定了机器学习结果的上限，而算法只是尽可能逼近这个上限。
数据要有代表性，否则必然会过拟合。
而且对于分类问题，数据偏斜不能过于严重，不同类别的数据数量不要有数个数量级的差距。
而且还要对数据的量级有一个评估，多少个样本，多少个特征，可以估算出其对内存的消耗程度，判训练过程中内存是否能够放得下。如果放不下就得考虑改进算法或者使用一些降维的技巧了。如果数量实在太大，那就要考虑分布式了。
特征预处理与特征选择
良好的数据要能够提取出良好的特征才能真正发挥效力
7、逻辑斯蒂回归为什么要对特征进行离散化
如七月在线老师所说：
① 非线性！非线性！非线性！逻辑回归属于广义线性模型，表达能力受限；单变量离散化为 N 个后每个变量有单独的权重，相当于为模型引入了非线性，能够提升模型表达能力，加大拟合； 离散特征增加和减少都很容易，易于模型的快速迭代；
② 速度快！速度快！速度快！稀疏向量内积乘法运算速度快，计算结果方便存储，容易扩展；
③ 鲁棒性！鲁棒性！鲁棒性！离散化后的特征对异常数据有很强的鲁棒性：比如一个特征是年龄 > 0 是 1，否则 0。如果特征没有离散化，一个异常数据“年龄 300 岁”会给模型造成很大的干扰；
④ 方便交叉与特征组合：离散化后可以进行特征交叉，由 M+N 个变量变为 M*N 个变量，进一步引

非线性，提升表达能力；
⑤ 稳定性：特征离散化后，模型会更稳定，比如如果对用户年龄离散化，20-30 作为一个区间，不因为一个用户年龄长了一岁就变成一个完全不同的人。当然处于区间相邻处的样本会刚好相反，所以么划分区间是门学问；
⑥ 简化模型：特征离散化以后，起到了简化了逻辑回归模型的作用，降低了模型过拟合的风险。
8、简单介绍下 LR
@rickjin：把 LR 从头到脚都给讲一遍。建模，现场数学推导，每种解法的原理，正则化，LR 和 maxent 模型啥关系。有不少会背答案的人，问逻辑细节就糊涂了。
原理都会? 那就问工程，并行化怎么做，有几种并行化方式，读过哪些开源的实现。还会，那就准备了吧，顺便逼问 LR 模型发展历史。
9、overfitting 怎么解决
overfitting 就是过拟合, 其直观的表现如下图所示，随着训练过程的进行，模型复杂度增加，在 training data 上的 error 渐渐减小，但是在验证集上的 error 却反而渐渐增大——因为训练出来的网过拟合了训练集, 对训练集外的数据却不 work, 这称之为泛化(generalization)性能不好。泛化性能是练的效果评价中的首要目标，没有良好的泛化，就等于南辕北辙, 一切都是无用功。
10、LR 和 SVM 的联系与区别解析一
LR 和 SVM 都可以处理分类问题，且一般都用于处理线性二分类问题（在改进的情况下可以处多分类问题）
区别：
1、LR 是参数模型，svm 是非参数模型，linear 和 rbf 则是针对数据线性可分和不可分的区别；
2、从目标函数来看，区别在于逻辑回归采用的是 logistical loss，SVM 采用的是 hinge loss，这两损失函数的目的都是增加对分类影响较大的数据点的权重，减少与分类关系较小的数据点的权重。<b>
3、SVM 的处理方法是只考虑 support vectors，也就是和分类最相关的少数点，去学习分类器。而辑回归通过非线性映射，大大减小了离分类平面较远的点的权重，相对提升了与分类最相关的数据点权重。
4、逻辑回归相对来说模型更简单，好理解，特别是大规模线性分类时比较方便。而 SVM 的理和优化相对来说复杂一些，SVM 转化为对偶问题后,分类只需要计算与少数几个支持向量的距离,这个进行复杂核函数计算时优势很明显,能够大大简化模型和计算。
5、logic 能做的 svm 能做，但可能在准确率上有问题，svm 能做的 logic 有的做不了。
11、什么是熵
从名字上来看，熵给人一种很玄乎，不知道是啥的感觉。其实，熵的定义很简单，即用来表示随变量的不确定性。之所以给人玄乎的感觉，大概是因为为何要取这样的名字，以及怎么用。
熵的概念最早起源于物理学，用于度量一个热力学系统的无序程度。在信息论里面，熵是对不确性的测量。
12、说说梯度下降法
1 什么是梯度下降法
经常在机器学习中的优化问题中看到一个算法，即梯度下降法，那到底什么是梯度下降法呢？
维基百科给出的定义是梯度下降法（Gradient descent）是一个一阶最优化算法，通常也称为最速下法。 要使用梯度下降法找到一个函数的局部极小值，必须向函数上当前点对应梯度（或者是近似梯度的反方向的规定步长距离点进行迭代搜索。如果相反地向梯度正方向迭代进行搜索，则会接近函数的部极大值点；这个过程则被称为梯度上升法。
13、牛顿法和梯度下降法有什么不同？
牛顿法（Newton's method）
牛顿法是一种在实数域和复数域上近似求解方程的方法。方法使用函数 f (x)的泰勒级数的前面几项来找方程 f (x) = 0 的根。牛顿法最大的特点就在于它的收敛速度很快。
14、熵、联合熵、条件熵、相对熵、互信息的定义
为了更好的理解，需要了解的概率必备知识有：
大写字母 X 表示随机变量，小写字母 x 表示随机变量 X 的某个具体的取值；
P(X)表示随机变量 X 的概率分布，P(X,Y)表示随机变量 X、Y 的联合概率分布，P(Y|X)表示已知随机量 X 的情况下随机变量 Y 的条件概率分布；
p(X = x)表示随机变量 X 取某个具体值的概率，简记为 p(x)；
p(X = x, Y = y) 表示联合概率，简记为 p(x,y)，p(Y = y|X = x)表示条件概率，简记为 p(y|x)，且有：p x,y) = p(x) * p(y|x)。

15、说说你知道的核函数
<img src="https://ld246.com/images/img-loading.svg" alt="数据分析转岗 AI 薪资翻 3 倍多 | 机器学习面试都问些什么？" data-src="https://b3logfile.com/file/2021/05/d75a80956c0e4326626933434b45a61.jpeg?imageView2/2/interlace/1/format/jpg">数据分析转岗 AI 薪翻 3 倍多 | 机器学习面试都问些什么？
16、什么是拟牛顿法（Quasi-Newton Methods）？
拟牛顿法是求解非线性优化问题最有效的方法之一，于 20 世纪 50 年代由美国 Argonne 国家验室的物理学家 W.C.Davidon 所提出来。Davidon 设计的这种算法在当时看来是非线性优化领域最创造性的发明之一。不久 R. Fletcher 和 M. J. D. Powell 证实了这种新的算法远比其他方法快速和可，使得非线性优化这门学科在一夜之间突飞猛进。
拟牛顿法的本质思想是改善牛顿法每次需要求解复杂的 Hessian 矩阵的逆矩阵的缺陷，它使用定矩阵来近似 Hessian 矩阵的逆，从而简化了运算的复杂度。拟牛顿法和最速下降法一样只要求每一迭代时知道目标函数的梯度。通过测量梯度的变化，构造一个目标函数的模型使之足以产生超线性收性。这类方法大大优于最速下降法，尤其对于困难的问题。
17、kmeans 的复杂度？
时间复杂度：O(tKmn)，其中，t 为迭代次数，K 为簇的数目，m 为记录数（也可认为是样本数，n 为维数
空间复杂度：O((m+K)n)，其中，K 为簇的数目，m 为记录数（也可认为是样本数），n 为维数
18、请说说随机梯度下降法的问题和挑战？
<img src="https://ld246.com/images/img-loading.svg" alt="数据分析转岗 AI 薪资翻 3 倍多 | 机器学习面试都问些什么？" data-src="https://b3logfile.com/file/2021/05/eee6b1a4be2d40ffaae2d8a44dbc7ba.jpeg?imageView2/2/interlace/1/format/jpg">数据分析转岗 AI 薪翻 3 倍多 | 机器学习面试都问些什么？
那到底如何优化随机梯度法呢？详情请点击：论文公开课第一期：详解梯度下降等各类优化算法含视频和 PPT 下载）（链接：
https://ask.julyedu.com/question/7913）

19、说说共轭梯度法？
共轭梯度法是介于梯度下降法（最速下降法）与牛顿法之间的一个方法，它仅需利用一阶导数信息，克服了梯度下降法收敛慢的缺点，又避免了牛顿法需要存储和计算 Hessian 矩阵并求逆的缺点，共轭度法不仅是解决大型线性方程组最有用的方法之一，也是解大型非线性最优化最有效的算法之一。在种优化算法中，共轭梯度法是非常重要的一种。其优点是所需存储量小，具有逐步收敛性，稳定性高而且不需要任何外来参数。
20、对所有优化问题来说, 有没有可能找到比現在已知算法更好的算法？
没有免费的午餐定理：
对于训练样本（黑点），不同的算法 A/B 在不同的测试样本（白点）中有不同的表现，这表示：对于个学习算法 A，若它在某些问题上比学习算法 B 更好，则必然存在一些问题，在那里 B 比 A 好。

也就是说：对于所有问题，无论学习算法 A 多聪明，学习算法 B 多笨拙，它们的期望性能相同。

但是：没有免费午餐定理假设所有问题出现几率相同，实际应用中，不同的场景，会有不同的问题分，所以，在优化算法时，针对具体问题进行分析，是算法优化的核心所在。
21、什么是最大熵
熵是随机变量不确定性的度量，不确定性越大，熵值越大；若随机变量退化成定值，熵为 0。如没有外界干扰，随机变量总是趋向于无序，在经过足够时间的稳定演化，它应该能够达到的最大程度熵。
为了准确的估计随机变量的状态，我们一般习惯性最大化熵，认为在所有可能的概率模型（分布的集合中，熵最大的模型是最好的模型。换言之，在已知部分知识的前提下，关于未知分布最合理的断就是符合已知知识最不确定或最随机的推断，其原则是承认已知事物（知识），且对未知事物不做何假设，没有任何偏见
22、LR 与线性回归的区别与联系
LR 工业上一般指 Logistic Regression(逻辑回归)而不是 Linear Regression(线性回归). LR 在线

回归的实数范围输出值上施加 sigmoid 函数将值收敛到 0~1 范围, 其目标函数也因此从差平方和函变为对数损失函数, 以提供最优化所需导数（sigmoid 函数是 softmax 函数的二元特例, 其导数均为数值的 f*(1-f)形式）。
请注意, LR 往往是解决二元 0/1 分类问题的, 只是它和线性回归耦合太紧, 不自觉也冠了个回归名字(马甲无处不在). 若要求多元分类,就要把 sigmoid 换成大名鼎鼎的 softmax 了。
23、简单说下有监督学习和无监督学习的区别
有监督学习：对具有标记的训练样本进行学习，以尽可能对训练样本集外的数据进行分类预测。LR,SVM,BP,RF,GBDT）
无监督学习：对未标记的样本进行训练学习，比发现这些样本中的结构知识。(KMeans,PCA)
24、请问（决策树、Random Forest、Boosting、Adaboot）GBDT 和 XGBoost 的别是什么？
集成学习的集成对象是学习器. Bagging 和 Boosting 属于集成学习的两类方法. Bagging 方法放回地采样同数量样本训练每个学习器, 然后再一起集成(简单投票); Boosting 方法使用全部样本(可权重)依次训练每个学习器, 迭代集成(平滑加权).
决策树属于最常用的学习器, 其学习过程是从根建立树, 也就是如何决策叶子节点分裂. ID3/C4.5 策树用信息熵计算最优分裂, CART 决策树用基尼指数计算最优分裂, xgboost 决策树使用二阶泰勒展系数计算最优分裂。
25、机器学习中的正则化到底是什么意思？
经常在各种文章或资料中看到正则化，比如说，一般的目标函数都包含下面两项

数据分析转岗 AI 薪翻 3 倍多 | 机器学习面试都问些什么？
其中，误差/损失函数鼓励我们的模型尽量去拟合训练数据，使得最后的模型会有比较少的 bias 而正则化项则鼓励更加简单的模型。因为当模型简单之后，有限数据拟合出来结果的随机性比较小，容易过拟合，使得最后模型的预测更加稳定。
但一直没有一篇好的文章理清到底什么是正则化？
说到正则化，得先从过拟合问题开始谈起。
<h5 id="添加微信-julyedukefu14-回复-11-领取最新升级版-名企AI面试100题--">添加微信：julyedukefu14，回复【11】领取最新升级版【名企 AI 面试 100 题】</h5>