链滴

梳理常见的机器学习面试题，你知道几个？

作者：julyedu

原文链接：https://ld246.com/article/1621499487397

来源网站：链滴

26、说说常见的损失函数?

对于给定的输入X，由f(X)给出相应的输出Y，这个输出的预测值f(X)与真实值Y可能一致也可能不一致要知道，有时损失或误差是不可避免的），用一个损失函数来度量预测错误的程度。损失函数记为L(Y,(X))，用来估量你模型的预测值f(x)与真实值Y的不一致程度。

27、为什么xgboost要用泰勒展开，优势在哪里?

xgboost使用了一阶和二阶偏导, 二阶导数有利于梯度下降的更快更准. 使用泰勒展开取得函数做自变的二阶导数形式, 可以在不选定损失函数具体形式的情况下, 仅仅依靠输入数据的值就可以进行叶子分优化计算, 本质上也就把损失函数的选取和模型算法优化/参数选择分开了. 这种去耦合增加了xgboos的适用性, 使得它按需选取损失函数, 可以用于分类, 也可以用于回归。

28、协方差和相关性有什么区别?

相关性是协方差的标准化格式。协方差本身很难做比较。例如：如果我们计算工资（$）和年龄（岁的协方差，因为这两个变量有不同的度量，所以我们会得到不能做比较的不同的协方差。

29、xgboost如何寻找最优特征？是有放回还是无放回的呢?

xgboost在训练的过程中给出各个特征的增益评分，最大增益的特征会被选出来作为分裂依据, 从而忆了每个特征对在模型训练时的重要性 -- 从根到叶子中间节点涉及某特征的次数作为该特征重要性序.

30、谈谈判别式模型和生成式模型?

判别方法：由数据直接学习决策函数 Y = f（X），或者由条件分布概率 P（Y|X）作为预测模型，即别模型。

生成方法：由数据学习联合概率密度分布函数 P（X,Y）,然后求出条件概率分布P(Y|X)作为预测的模，即生成模型。

由生成模型可以得到判别模型，但由判别模型得不到生成模型。

常见的判别模型有：K近邻、SVM、决策树、感知机、线性判别分析（LDA）、线性回归、传统的神网络、逻辑斯蒂回归、boosting、条件随机场

常见的生成模型有：朴素贝叶斯、隐马尔可夫模型、高斯混合模型、文档主题生成模型（LDA）、限玻尔兹曼机

31、线性分类器与非线性分类器的区别以及优劣

线性和非线性是针对，模型参数和输入特征来讲的；比如输入x，模型y=ax+ax^2那么就是非线性模，如果输入是x和X^2则模型是线性的。

线性分类器可解释性好，计算复杂度较低，不足之处是模型的拟合效果相对弱些。

非线性分类器效果拟合能力较强，不足之处是数据量不足容易过拟合、计算复杂度高、可解释性不好。

常见的线性分类器有：LR,贝叶斯分类，单层感知机、线性回归

常见的非线性分类器：决策树、RF、GBDT、多层感知机

SVM两种都有（看线性核还是高斯核）

32、L1和L2的区别

L1范数（L1 norm）是指向量中各个元素绝对值之和，也有个美称叫“稀疏规则算子”（Lasso regularization）。

比如 向量A=[1，-1，3]， 那么A的L1范数为 |1|+|-1|+|3|.

简单总结一下就是：

L1范数: 为x向量各个元素绝对值之和。

L2范数: 为x向量各个元素平方和的1/2次方，L2范数又称Euclidean范数或者Frobenius范数

Lp范数: 为x向量各个元素绝对值p次方和的1/p次方

33、L1和L2正则先验分别服从什么分布

面试中遇到的，L1和L2正则先验分别服从什么分布，L1是拉普拉斯分布，L2是高斯分布。

34、简单介绍下logistics回归?

逻辑回归（Logistic Regression）是机器学习中的一种分类模型，由于算法的简单和高效，在实际中用非常广泛。

比如在实际工作中，我们可能会遇到如下问题：

预测一个用户是否点击特定的商品

判断用户的性别

预测用户是否会购买给定的品类

判断一条评论是正面的还是负面的

这些都可以看做是分类问题，更准确地，都可以看做是二分类问题。要解决这些问题，通常会用到一已有的分类算法，比如逻辑回归，或者支持向量机。它们都属于有监督的学习，因此在使用这些算法前，必须要先收集一批标注好的数据作为训练集。有些标注可以从log中拿到（用户的点击，购买）有些可以从用户填写的信息中获得（性别），也有一些可能需要人工标注（评论情感极性）。

35、说一下Adaboost，权值更新公式。当弱分类器是Gm时，每个样本的的权重是w1，w2...，请出最终的决策公式。

给定一个训练数据集T={(x1,y1), (x2,y2)...(xN,yN)}

36、经常在网上搜索东西的朋友知道，当你不小心输入一个不存在的单词时，搜索引擎会提示你是不要输入某一个正确的单词，比如当你在Google中输入“Julw”时，系统会猜测你的意图：是不是要索“July”

梳理常见的机器学习面试题，你知道几个?

用户输入一个单词时，可能拼写正确，也可能拼写错误。如果把拼写正确的情况记做c（代表correct，拼写错误的情况记做w（代表wrong），那么"拼写检查"要做的事情就是：在发生w的情况下，试推断出c。换言之：已知w，然后在若干个备选方案中，找出可能性最大的那个c

37、为什么朴素贝叶斯如此“朴素”？

因为它假定所有的特征在数据集中的作用是同样重要和独立的。正如我们所知，这个假设在现实世界是很不真实的，因此，说朴素贝叶斯真的很“朴素”。

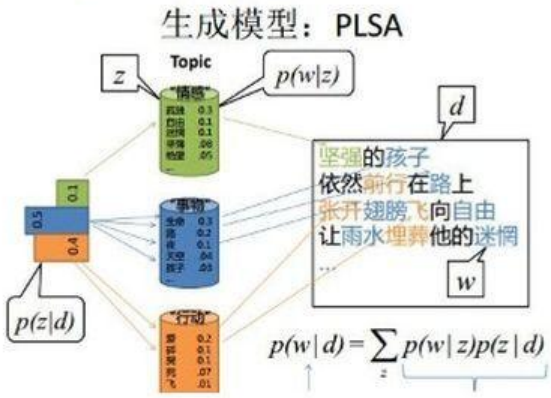
朴素贝叶斯模型(Naive Bayesian Model)的朴素(Naive)的含义是"很简单很天真"地假设样本特征彼独立. 这个假设现实中基本上不存在, 但特征相关性很小的实际情况还是很多的, 所以这个模型仍然能工作得很好。

38、请大致对比下plsa和LDA的区别

解析：

pLSA中，主题分布和词分布确定后，以一定的概率$P(z_k|d_i)$、$P(w_j|z_k)$分别选取具体的主题和词项，生成好文档。而后根据生成好的文档反推其主题分布、词分布时，最终用EM算法（极大似然估计思想）求解出了两个未知但固定的参数的值：$\phi_{k,j}$（由$P(w_j|z_k)$转换而来）和$\theta_{i,k}$（由$P(z_k|d_i)$转换而来）。
文档d产生主题z的概率，主题z产生单词w的概率都是两个固定的值。

举个文档d产生主题z的例子。给定一篇文档d，主题分布是一定的，比如{ P(zi|d), i = 1,2,3 }可能就是{0.4,0.5,0.1}，表示z1、z2、z3，这3个主题被文档d选中的概率都是个固定的值：P(z1|d) = 0.4、P(z2|d) = 0.5、P(z3|d) = 0.1，如下图所示（图截取自沈博PPT上）：



但在贝叶斯框架下的LDA中，我们不再认为主题分布（各个主题在文档中出现的概率分布）和词分布（各个词语在某个主题下出现的概率分布）是唯一确定的（而是随机变量），而是有很多种可能。但一篇文档总得对应一个主题分布和一个词分布吧，怎么办呢？LDA为它们弄了两个Dirichlet先验参数，这个Dirichlet先验为[illegible]文档随机地取出某个主题分布和词分布。

梳理常见的机器学习面试题，你知道几个?

39、请详细说说EM算法

到底什么是EM算法呢？Wikipedia给的解释是：

最大期望算法（Expectation-maximization algorithm，又译为期望最大化算法），是在概率模型中找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐性变量。

40、KNN中的K如何选取的?

关于什么是KNN，可以查看此文：《从K近邻算法、距离度量谈到KD树、SIFT+BBF算法》（链接：

http://blog.csdn.net/v_july_v/article/details/8203674）。KNN中的K值选取对K近邻算法的结果产生重大影响。如李航博士的一书「统计学习方法」上所说：如果选择较小的K值，就相当于用较小领域中的训练实例进行预测，“学习”近似误差会减小，只有与输入实例较近或相似的训练实例才会预测结果起作用，与此同时带来的问题是“学习”的估计误差会增大，换句话说，K值的减小就意味整体模型变得复杂，容易发生过拟合；

如果选择较大的K值，就相当于用较大领域中的训练实例进行预测，其优点是可以减少学习的估计误，但缺点是学习的近似误差会增大。这时候，与输入实例较远（不相似的）训练实例也会对预测器作，使预测发生错误，且K值的增大就意味着整体的模型变得简单。

K=N，则完全不足取，因为此时无论输入实例是什么，都只是简单的预测它属于在训练实例中最多的，模型过于简单，忽略了训练实例中大量有用信息。

在实际应用中，K值一般取一个比较小的数值，例如采用交叉验证法（简单来说，就是一部分样本做练集，一部分做测试集）来选择最优的K值。

41、防止过拟合的方法

过拟合的原因是算法的学习能力过强；一些假设条件（如样本独立同分布）可能是不成立的；训练样过少不能对整个空间进行分布估计。

处理方法：

1 早停止：如在训练中多次迭代后发现模型性能没有显著提高就停止训练

2 数据集扩增：原有数据增加、原有数据加随机噪声、重采样

3 正则化，正则化可以限制模型的复杂度

4 交叉验证

5 特征选择/特征降维

6 创建一个验证集是最基本的防止过拟合的方法。我们最终训练得到的模型目标是要在验证集上面有的表现，而不训练集。

42、机器学习中，为何要经常对数据做归一化

机器学习模型被互联网行业广泛应用，如排序（参见：排序学习实践

http://www.cnblogs.com/LBSer/p/4439542.html）、推荐、反作弊、定位（参见：基于朴素贝叶的定位算法http://www.cnblogs.com/LBSer/p/4020370.html）等。

一般做机器学习应用的时候大部分时间是花费在特征处理上，其中很关键的一步就是对特征数据进行一化。

为什么要归一化呢？很多同学并未搞清楚，维基百科给出的解释：1）归一化后加快了梯度下降求最解的速度；2）归一化有可能提高精度。

43、什么最小二乘法?

我们口头中经常说：一般来说，平均来说。如平均来说，不吸烟的健康优于吸烟者，之所以要加“平”二字，是因为凡事皆有例外，总存在某个特别的人他吸烟但由于经常锻炼所以他的健康状况可能会于他身边不吸烟的朋友。而最小二乘法的一个最简单的例子便是算术平均。

最小二乘法（又称最小平方法）是一种数学优化技术。它通过最小化误差的平方和寻找数据的最佳函匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平和为最小。

44、梯度下降法找到的一定是下降最快的方向么?

梯度下降法并不一定是全局下降最快的方向，它只是目标函数在当前的点的切平面（当然高维问题不叫平面）上下降最快的方向。在practical implementation中，牛顿方向（考虑海森矩阵）才一般被为是下降最快的方向，可以达到superlinear的收敛速度。梯度下降类的算法的收敛速度一般是linear至sublinear的（在某些带复杂约束的问题）。by林小溪（

https://www.zhihu.com/question/30672734/answer/139689869）。

45、简单说说贝叶斯定理的

解析：

在引出贝叶斯定理之前，先学习几个定义：

条件概率（又称后验概率）就是事件A在另外一个事件B已经发生条件下的发生概率。条件概率表示为P(A|B)，读作“在B条件下A的概率”。

比如，在同一个样本空间Ω中的事件或者子集A与B，如果随机从Ω中选出的一个元素属于B，那么这个随机选择的元素还属于A的概率就定义为在B的前提下A的条件概率，所以：P(A|B) = |A∩B|/|B|，接着分子、分母都除以|Ω|得到

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

联合概率表示两个事件共同发生的概率。A与B的联合概率表示为$P(A \cap B)$或者$P(A, B)$。

边缘概率（又称先验概率）是某个事件发生的概率。边缘概率是这样得到的：在联合概率中，把最终结果中那些不需要的事件通过合并成它们的全概率，而消去它们（对离散随机变量用求和得全概率，对连续随机变量用积分得全概率），这称为边缘化（marginalization），比如A的边缘概率表示为P(A)，B的边缘概率表示为P(B)。

接着，考虑一个问题：P(A|B)是在B发生的情况下A发生的可能性。

首先，事件B发生之前，我们对事件A的发生有一个基本的概率判断，称为A的先验概率，用P(A)表示；

其次，事件B发生之后，我们对事件A的发生概率重新评估，称为A的后验概率，用P(A|B)表示；

类似的，事件A发生之前，我们对事件B的发生有一个基本的概率判断，称为B的先验概率，用P(B)表示；

同样，事件A发生之后，我们对事件B的发生概率重新评估，称为B的后验概率，用P(B|A)表示。

贝叶斯定理便是基于下述贝叶斯公式：

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

上述公式的推导其实非常简单，就是从条件概率推出。

根据条件概率的定义，在事件B发生的条件下事件A发生的概率是$P(A|B) = \frac{P(A \cap B)}{P(B)}$

同样地，在事件A发生的条件下事件B发生的概率

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

整理与合并上述两个方程式，便可以得到：

$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A).$$

接着，上式两边同除以P(B)，若P(B)是非零的，我们便可以得到贝叶斯定理的公式表达式：

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

所以，贝叶斯公式可以直接根据条件概率的定义直接推出。即因为P(A,B) = P(A)P(B|A) = P(B)P(A|B)，所以P(A|B) = P(A)P(B|A) / P(B)。更多请参见此文：《从贝叶斯方法谈到贝叶斯网络》[链接：http://blog.csdn.net/v_july_v/article/details/40984699]。

梳理常见的机器学习面试题，你知道几个？

46、怎么理解决策树、xgboost能处理缺失值？而有的模型(svm)对缺失值比较敏感。

本题解析来源：

https://www.zhihu.com/question/58230411

首先从两个角度解释你的困惑：

工具包自动处理数据缺失不代表具体的算法可以处理缺失项

对于有缺失的数据：以决策树为原型的模型优于依赖距离度量的模型

回答中也会介绍树模型，如随机森林(Random Forest)和xgboost如何处理缺失值。文章最后总结了有缺失值时选择模型的小建议。

47、请举例说明什么是标准化、归一化

一、标准化（standardization）

简单来说，标准化是依照特征矩阵的列处理数据，其通过求z-score的方法，将样本的特征值转换到一量纲下。

公式一般为：(X-mean)/std，其中mean是平均值，std是方差。

从公式我们可以看出，标准化操作（standardization）是将数据按其属性（按列）减去平均值，然再除以方差。

这个过程从几何上理解就是，先将坐标轴零轴平移到均值这条线上，然后再进行一个缩放，涉及到的是平移和缩放两个动作。这样处理以后的结果就是，对于每个属性（每列）来说，所有数据都聚集在附近，方差为1。计算时对每个属性/每列分别进行。

48、随机森林如何处理缺失值?

@Yieshah：众所周知，机器学习中处理缺失值的方法有很多，然而，由题目“随机森林如何处理缺值”可知，问题关键在于随机森林如何处理，所以先简要介绍下随机森林吧。

随机森林是由很多个决策树组成的，首先要建立Bootstrap数据集，即从原始的数据中有放回地随机取一些，作为新的数据集，新数据集中会存在重复的数据，然后对每个数据集构造一个决策树，但是是直接用所有的特征来建造决策树，而是对于每一步，都从中随机的选择一些特征，来构造决策树，样我们就构建了多个决策树，组成随机森林，把数据输入各个决策树中，看一看每个决策树的判断结，统计一下所有决策树的预测结果，Bagging整合结果，得到最终输出。

那么，随机森林中如何处理缺失值呢？根据随机森林创建和训练的特点，随机森林对缺失值的处理还比较特殊的。

49、随机森林如何评估特征重要性?

衡量变量重要性的方法有两种，Decrease GINI 和 Decrease Accuracy：

1. Decrease GINI：

对于分类问题（将某个样本划分到某一类），也就是离散变量问题，CART使用Gini值作为评判标准定义为Gini=1-∑(P(i)*P(i)),P(i)为当前节点上数据集中第i类样本的比例。例如：分为2类，当前节点上100个样本，属于第一类的样本有70个，属于第二类的样本有30个，则Gini=1-0.7×07-0.3×03=0.4，可以看出，类别分布越平均，Gini值越大，类分布越不均匀，Gini值越小。在寻找最佳的分类特征阈值时，评判标准为：argmax（Gini-GiniLeft-GiniRight），即寻找最佳的特征f和阈值th，使得当节点的Gini值减去左子节点的Gini和右子节点的Gini值最大。

对于回归问题，相对更加简单，直接使用argmax(Var-VarLeft-VarRight)作为评判标准，即当前节点练集的方差Var减去减去左子节点的方差VarLeft和右子节点的方差VarRight值最大。

2. Decrease Accuracy：

对于一棵树Tb(x)，我们用OOB样本可以得到测试误差1；然后随机改变OOB样本的第j列：保持其他不变，对第j列进行随机的上下置换，得到误差2。至此，我们可以用误差1-误差2来刻画变量j的重要。基本思想就是，如果一个变量j足够重要，那么改变它会极大的增加测试误差；反之，如果改变它测误差没有增大，则说明该变量不是那么的重要。

50、请说说Kmeans的优化?

解析一

k-means：在大数据的条件下，会耗费大量的时间和内存。

优化k-means的建议：

1、减少聚类的数目K。因为，每个样本都要跟类中心计算距离。

2、减少样本的特征维度。比如说，通过PCA等进行降维。

3、考察其他的聚类算法，通过选取toy数据，去测试不同聚类算法的性能。

4、hadoop集群，K-means算法是很容易进行并行计算的。