11. LR常见问题

1．逻辑回归概括

逻辑回归用一句话概括就是，逻辑回归假设数据服从伯努利分布，通过极大化似然函数的方法，运用梯度下降来求解参数，来解决二分类的问题。

2．逻辑回归的基本假设

- 逻辑回归的第一个基本假设就是**假设数据服从伯努利分布**。伯努利分布又名两点分布或 0-1 分布，即一个随机事件的可能结果只有两种，要么是 0，要么是 1。最简单的例子就是抛硬币，只有正面朝上和反面朝上两种结果，正面朝上的概率是 p，反面朝上的概率就是 1-p。此时逻辑回归的整个模型描述为 $h_\theta(x;\theta)=p$
- 逻辑回归的第二个假设是**假设样本为正的概率**是：

$$p=\frac{1}{1+e^{-\theta^T x}}$$

所有逻辑回归的最终形式是

$$h_\theta(x;\theta)=\frac{1}{1+e^{-\theta^T x}}$$

3．逻辑回归的损失函数

我们采用似然函数作为模型更新的loss，最大化似然函数

$$L(\theta) = p(\hat{y}|X;\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)$$

$$= \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{(1-y^{(i)})}$$

这个损失函数很难求导，于是我们将其取log，变成对数似然函数并转化为

$$loss(\theta) = log(L(\theta)) = \sum_{i=1}^{m} y^{(i)} log(h(x^{(i)})) + (1 - y^{(i)}) log(1 - h(x^{(i)}))$$

4．logistic回归和线性回归的关系

逻辑回归和线性回归首先都是广义的线性回归，逻辑回归使用了线性回归，并且多了个Sigmod函数，使样本映射到[0,1]之间的数值，从而来处理分类问题（同时也减少了离群点的影响）。**logistic回归分类模型的预测函数，是在用线性回归模型的预测值，去逼近真实标记的对数几率：**

$$\ln \frac{y}{1-y} = w^T x + b$$

从宏观来看，线性回归是在拟合输入向量x的分布，而逻辑回归中的线性函数是在拟合决策边界，它们的目标是不一样的。

线性回归模型：

$$f(x) = w_0x_0 + w_1x_1 + \cdots + w_nx_n + b$$

写成向量形式为：

$$f(x) = w^Tx + b$$

“广义线性回归”模型为：

$$y = g^{-1}(w^Tx + b)$$

上诉线性回归模型只能够进行回归学习，但是若要是做分类任务，就需要将分类任务的真实标记y与线性回归模型的预测值联系起来。

线性回归模型产生的预测值$z = wx + b$是一个实值，二分类问题的输出的标记y={0,1}，所以我们将实值y转化成0/1值便可，这样有一个可选函数便是“单位阶跃函数”：

$$y = \begin{cases} 0, & z < 0 \\ 0.5, & z = 0 \\ 1, & z > 0 \end{cases}$$

但是单位阶跃函数是非连续的函数，我们需要一个连续的函数，来输出0-1之间的分类概率，“Sigmoid函数”便可以很好的取代单位阶跃函数：

$$y = \frac{1}{1 + e^{-z}}$$

若将y视为样本x作为正例的可能性，则 1 – y 便是其反例的可能性。二者的比值便被称为“几率”，反映了x作为正例的相对可能性，再取对数就叫对数几率。这也是logistic回归又被称为对数几率回归的原因！

总结一下logistic回归和线性回归模型的关系：logistic回归分类模型的预测函数，是在用线性回归模型的预测值，去逼近真实标记的对数几率！这样也便实现了上面说的**将线性回归的预测值和分类任务的真实标记联系在了一起**！

5．线性回归和逻辑回归的异同

- 相同1：logistic回归分类模型的预测函数，是在用线性回归模型的预测值，去逼近真实标记的对数几率！不过，线性回归用最小二乘法拟合输入输出变量的关系，而逻辑回归中

的线性函数是在拟合决策边界，它们的目标是不一样的。

- 相同2：我们可以认为二者都使用了极大似然估计来对训练样本进行建模。线性回归使用最小二乘法，实际上就是在自变量x与超参数θ确定，因变量y服从正态分布的假设下，使用极大似然估计的一个化简；而逻辑回归中通过对似然函数的转化和求解，得到最佳参数θ。
- 相同3：二者在求解超参数的过程中，都可以使用梯度下降的方法，这也是监督学习中一个常见的相似之处
- 区别1：逻辑回归输出的是离散型变量，用于分类，线性回归输出的是连续性的，用于预测。
- 区别2：逻辑回归是假设变量服从伯努利分布，线性回归假设变量服从高斯分布。
- 区别3：逻辑回归是用最大似然法去计算假设函数中的最优参数值，而线性回归是用最小二乘法去对自变量因变量关系进行拟合。

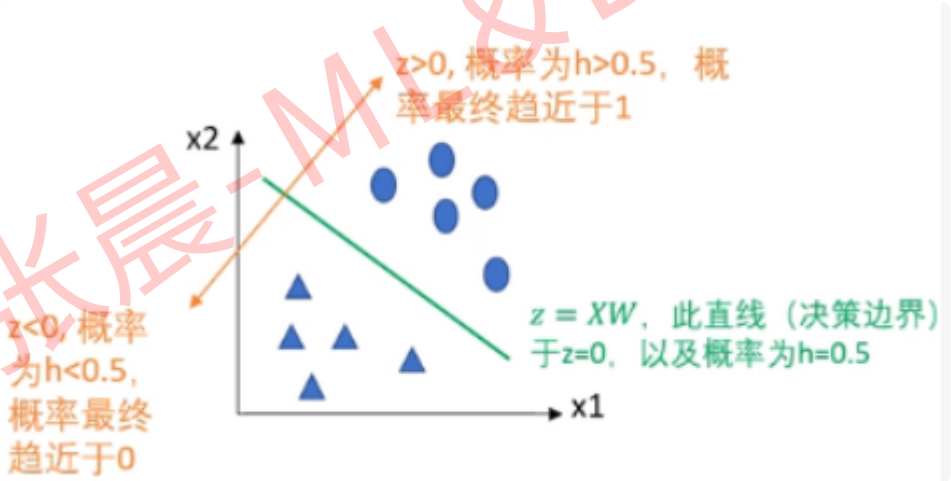
6．逻辑回归是线性模型吗

逻辑回归是一种广义线性模型，它引入了Sigmod函数，是非线性模型，但本质上还是一个线性回归模型，因为除去Sigmod函数映射关系，其他的算法原理，步骤都是线性回归的。逻辑回归的思路是，先拟合决策边界(这里的决策边界不局限于线性，还可以是多项式)，再建立这个边界与分类的概率联系，从而得到了二分类情况下的概率。

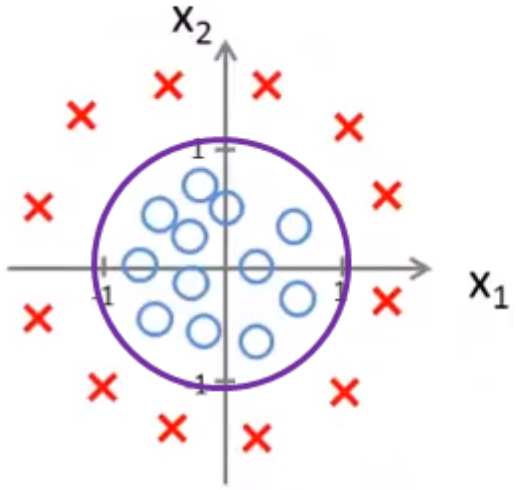
7．类别的分界线是什么

logistic回归分类模型的预测函数，是在用线性回归模型的预测值，去逼近真实标记的对数几率！

线性回归模型的预测值$w^T x$才是分隔曲线，它将样本点分为 $w^T x \geq 0$ 和 $w^T x < 0$ 两部分； sigmoid 函数不是样本点的分隔曲线，它表示的是逻辑回归的预测结果。



很多时候，直线并不能很好地作为决策边界，如下图所示：



此时需要使用多项式模型添加更多的特征：

$$h(x) = g\left(w^T x\right) = g\left(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2\right)$$

这相当于添加了两个新的特征：$w_3 = x_1^2$，$w_4 = x_2^2$

注意，此时仍然是线性模型，因为模型线性与否是针对参数来看的。
添加的特征越多，曲线越复杂，对训练样本的拟合度越高，同时也更容易导致过拟合而丧失泛化性。

8．逻辑回归输出值是概率吗

逻辑回归模型之所以是sigmoid 的形式，源于我们假设y服从伯努利分布，伯努利分布又属于指数分布族，经过推导，将伯努利分布变成指数分布族的形式后。我们发现伯努利分布的唯一参数Φ与指数分布族中的参数η具有sigmoid函数关系，于是我们转而求η与x的关系，此时，我们又假设η与x具有线性关系。至此，找到了我们要用的模型的样子，也就是逻辑回归。逻辑回归输出的到底是不是概率呢？**只有在满足 y服从伯努利分布，η和x之间存在线性关系时，输出值才是概率值（即满足假设的条件下）。不满足的情况下，得到的输出值，只是置信度**。

9．逻辑回归也可以处理多分类

方式一：修改逻辑回归的损失函数，使用 softmax 函数构造模型解决多分类问题，softmax 分类模型会有相同于类别数的输出，输出的值为对于样本属于各个类别的概率，最后对于样本进行预测的类型为概率值最高的那个类别。

方式二：根据每个类别都建立一个二分类器，本类别的样本标签定义为 0，其它分类样本标签定义为 1，则有多少个类别就构造多少个逻辑回归分类器

若所有类别之间有明显的互斥则使用 softmax 分类器，若所有类别不互斥有交叉的情况则构造相应类别个数的逻辑回归分类器。

10．逻辑回归的求解方法

由于该极大似然函数无法直接求解，我们一般通过对该函数进行梯度下降来不断逼近最优解。有随机梯度下降，批梯度下降，small batch 梯度下降三种方式，面试官可能会问这三种方式的优劣以及如何选择最合适的梯度下降方式。

- 批梯度下降会获得全局最优解，缺点是在更新每个参数的时候需要遍历所有的数据，计算量会很大，并且会有很多的冗余计算，导致的结果是当数据量大的时候，每个参数的更新都会很慢。
- 随机梯度下降是以高方差频繁更新，优点是使得sgd会跳到新的和潜在更好的局部最优解，缺点是使得收敛到局部最优解的过程更加的复杂。
- 小批量梯度下降结合了sgd和batch gd的优点，每次更新的时候使用n个样本。减少了参数更新的次数，可以达到更加稳定收敛结果，一般在深度学习当中我们采用这种方法。

其实这里还有一个隐藏的更加深的加分项，看你了不了解诸如Adam，动量法等优化方法。因为上述方法其实还有两个致命的问题。

第一个是如何对模型选择合适的学习率。自始至终保持同样的学习率其实不太合适。因为一开始参数刚刚开始学习的时候，此时的参数和最优解隔的比较远，需要保持一个较大的学习率尽快逼近最优解。但是学习到后面的时候，参数和最优解已经隔的比较近了，你还保持最初的学习率，容易越过最优点，在最优点附近来回振荡，通俗一点说，就很容易学过头了，跑偏了。

第二个是如何对参数选择合适的学习率。在实践中，对每个参数都保持的同样的学习率也是很不合理的。有些参数更新频繁，那么学习率可以适当小一点。有些参数更新缓慢，那么学习率就应该大一点。这里我们不展开，有空我会专门出一个专题介绍。

11．逻辑回归是如何分类的？

逻辑回归通过 logistic 函数将一个范围不定的连续值映射到 (0, 1) 的区间内，然后划定一个阈值，输出值大于这个阈值的是一类，小于这个阈值的是另一类。阈值会根据实际情况选择，一般会选择 0.5。

12．逻辑回归为什么使用极大似然函数作为损失函数，而不用平方损失函数？

使用极大似然函数取对数之后等同于对数损失函数，使用对数损失函数训练参数的速度比较快，其梯度下降过程中表示梯度的式子为 $\frac{dJ}{d\theta} = (\hat{y}_i - y_i)x_i$ ，梯度更新的速度只和 x_i, y_i 相关，与 logistic 函数本身的梯度无关，这样更新速度自始至终比较稳定。

如果使用平方损失函数，其梯度下降过程中表示梯度的式子为

$$\frac{dJ}{d\theta} = (y_i - \hat{y}_i)\hat{y}_i(1 - \hat{y}_i)x_i$$

这里引入了 $\hat{y}_i(1 - \hat{y}_i)$ 等价于 logistic 函数本身的梯度，logistic 函数在定义域内梯度都不大与 0.25，在输出接近于 0 或 1 时，梯度会变得非常小，出现梯度消失的问题，容易导致训练变慢。

另外对数损失函数只和分类正确的预测结果有关系，对于二元分类问题，使用极大似然估计的目的就是为了找到一个合适的参数 θ 使得把样本分为正确类的概率尽可能大，而不看重错误分类；相反，使用平方损失函数除了让正确的分类尽量变大，还会让错误的分类变得均匀，它对错误的输出惩罚比较大，但这在分类问题中是没有必要的。平方损失函数更适合回归问题。

13．概率和似然

在英文中，似然（likelihood）和概率（probability）是同义词，都指事件发生的可能性。但在统计中，似然与概率是不同的东西。

- 概率是已知参数，对结果可能性的预测。
- 似然是已知结果，对参数是某个值的可能性预测。

对于函数P(x|θ)， 输入有两个：x表示某一个具体的数据；θ 表示模型的参数。从不同的观测角度来看可以分为以下两种情况：

- 如果 θ 是已知确定的，x 是变量，则 P(x|θ) 称为概率函数 (probability function)，它描述对于不同的样本点 x ，其出现概率是多少。
- 如果 x 是已知确定的，θ 是变量，则 P(x|θ) 称为似然函数 (likelihood function)，它表示在不同的 θ 下，出现 x 这个样本点的概率是多少，也记作 L(θ|x) 或 L(x;θ) 或 f(x;θ)。

14．极大似然估计

给定一堆数据，假如我们知道它是从某一种分布中随机取出来的，可是我们并不知道这个分布具体的参，“模型已定，参数未知”。例如，我们知道这个分布是正态分布，但是不知道均值和方差；或者是二项分布，但是不知道均值。 最大似然估计（MLE，Maximum Likelihood Estimation）就可以用来估计模型的参数。**MLE的目标是找出一组参数，使得模型产生出观测数据的概率最大**:

$$\underset{\mu}{\operatorname{argmax}}\, p(\mathbf{X}; \mu)$$

其中 P(X; μ) 就是似然函数，表示在参数μ下出现观测数据的概率。我们假设每个观测数据是独立的，那么有

$$p(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} p(x_i; \mu)$$

为了求导方便，一般对目标取log。 所以最优化对似然函数等同于最优化对数似然函数：

$$\underset{\mu}{\operatorname{argmax}}\, p(\mathbf{X}; \mu) = \underset{\mu}{\operatorname{argmax}} \log p(\mathbf{X}; \mu)$$

15．特征相关性

逻辑回归在训练的过程当中，如果有很多的特征高度相关或者说有一个特征重复了100遍，会造成怎样的影响？

先说结论，如果在损失函数最终收敛的情况下，其实就算有很多特征高度相关也不会影响分类器的效果。但是对特征本身来说的话，假设只有一个特征，在不考虑采样的情况下，你现在将它重复100遍。训练完以后，数据还是这么多，但是这个特征本身重复了100遍，实质上将原来的特征分成了100份，每一个特征都是原来特征权重值的百分之一。如果在随机采样的情况下，其实训练收敛完以后，还是可以认为这100个特征和原来那一个特征扮演的效果一样，只是可能中间很多特征的值正负相消了。

为什么我们还是会在训练的过程当中将高度相关的特征去掉？

一方面，去掉高度相关的特征会让模型的可解释性更好。另一方面，可以大大提高训练的速度，如果模型当中有很多特征高度相关的话，就算损失函数本身收敛了，但实际上参数是没有收敛的，这样会拉低训练的速度，其次是特征多了，本身就会增大训练的时间。

16．特征离散化

在工业界，很少直接将连续值作为特征喂给逻辑回归模型，而是将连续特征离散化为一系列0、1特征交给逻辑回归模型，这样做的优势有以下几点：

- 稀疏向量内积乘法**运算速度快**，计算结果方便存储，容易scalable（扩展）。
- 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>30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。
- 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，不会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，所以怎么划分区间是门学问。
- 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，进一步引入非线性，提升表达能力。

李沐指出，模型是使用离散特征还是连续特征，其实是一个“**海量离散特征+简单模型**”同“**少量连续特征+复杂模型**”的权衡。既可以离散化用线性模型，也可以用连续特征加深度学习。就看是喜欢折腾特征还是折腾模型了。通常来说，前者容易，而且可以n个人一起并行做，有成功经验；后者目前看很赞，能走多远还须拭目以待。

17．逻辑回归的优缺点

- 优点：
  - 形式简单，模型的可解释性非常好。从**特征的权重可以看到不同的特征对最后结果的影响**，某个特征的权重值比较高，那么这个特征最后对结果的影响会比较大。
  - 模型效果不错。在工程上是可以接受的（作为baseline），**如果特征工程做的好，效果不会太差**，并且特征工程可以大家并行开发，大大加快开发的速度。
  - 训练速度较快。分类的时候，**计算量仅仅只和特征的数目相关**。并且逻辑回归的分布式优化sgd发展比较成熟，训练的速度可以通过堆机器进一步提高，这样我们可以在短时间内迭代好几个版本的模型。
  - 资源占用小,尤其是内存。因为**只需要存储各个维度的特征值**。
  - 输出结果方便调整。逻辑回归可以很方便的得到最后的分类结果，因为输出的是每个样本的概率分数，我们可以很容易的对这些概率分数进行cut off，也就是划分阈值(大于某个阈值的是一类，小于某个**阈值**的是一类)。
- 缺点:
  - 准确率并不是很高。因为形式非常的简单(非常类似线性模型)，**很难去拟合数据的真实分布**。
  - **很难处理数据不平衡的问题**。举个例子：如果我们对于一个正负样本非常不平衡的问题比如正负样本比 10000:1，就算我们把所有样本都预测为正也能使损失函数的值比较小。但是作为一个分类器，它对正负样本的区分能力不会很好。
  - 处理非线性数据较麻烦。逻辑回归在不引入其他方法的情况下，**只能处理线性可分的数据**，或者进一步说，处理二分类的问题。
  - 逻辑回归本身**无法筛选特征**。有时候，我们会用gbdt来筛选特征，然后再上逻辑回归。

18．L1正则和L2正则有什么区别

- 相同点：
  - 都用于避免过拟合
- 不同点：
  - L1正则是拉普拉斯先验，而L2正则则是高斯先验。
  - L1可以产生稀疏解，可以让一部分特征的系数缩小到0，从而间接实现特征选择。所以L1适用于特征之间有关联的情况。L2让所有特征的系数都缩小，但是不会减为0，它会使优化求解稳定快速。所以L2适用于特征之间没有关联的情况
  - 因为L1服从拉普拉斯分布，所以L1在0点处不可导，难以计算，这个方法可以使用Proximal Algorithms或者ADMM来解决。

19．逻辑回归与SVM

联系：

- 都是分类模型（二分类），本质上都在寻找分类超平面
- 都可加入正则化项，减轻过拟合的风险
- 都是判别模型
- 都是有监督的分类模型

区别：

- LR是参数模型，即假设样本服从某一分布；SVM是非参数模型
- 目标函数：LR的损失函数是交叉熵损失；SVM是hinge loss
- SVM依赖几个支持向量来分类；LR是通过非线性映射降低远离平面的点的权重，增大离平面近的点的权重
- LR–模型简单，好理解，精度低，可能局部最优；SVM–理解、优化复杂，精度高，全局最优，转化为对偶问题—>简化模型和计算
- LR可以做的SVM可以做（线性可分），SVM能做的LR不一定能做（线性不可分）
- LR产生类别的同时会产生概率，而SVM仅能产生类别
- LR不依赖于距离，而SVM是基于距离的
- **逻辑回归是经验风险最小化，svm是结构风险最小化。**这点体现在svm自带L2正则化项，逻辑回归并没有，但两个方法都可以增加不同的正则化项，如l1,l2等等。

20．选择LR还是SVM

假设： n = 特征数量，m = 训练样本数量

- 如果n相对于m更大，比如 n = 10,000，m = 1,000，则使用lr

理由：特征数相对于训练样本数已经够大了，使用线性模型就能取得不错的效果，不需要过于复杂的模型；

- 如果n较小，m比较大，比如n = 10，m = 10,000，则使用SVM（高斯核函数）

理由：在训练样本数量足够大而特征数较小的情况下，可以通过使用复杂核函数的SVM来获得更好的预测性能，而且因为训练样本数量并没有达到百万级，使用复杂核函数的SVM也不会导致运算过慢；

- 如果n较小，m非常大，比如n = 100, m = 500,000，则应该引入/创造更多的特征，然后使用 lr 或者线性核函数的SVM

理由：因为训练样本数量特别大，使用复杂核函数的SVM会导致运算很慢，因此应该考虑通过引入更多特征，然后使用线性核函数的SVM或者lr来构建预测性更好的模型。

21．逻辑回归与最大熵模型

逻辑回归跟最大熵模型MaxEnt没有本质区别。逻辑回归是最大熵对应类别为二类时的特殊情况，也就是当逻辑回归类别扩展到多类别时，就是最大熵模型。