**Logistic Regression**

**1. 如何推导Logistic Regression？**

LR的本质是一个基于条件概率的判别模型。LR的函数如下：

由于上述函数的特性，我们可作出如下的假设：

将两个公式合并成一个，如下：

假定样本与样本之间相互独立，那么整个样本集生成的概率即为所有样本生成概率的乘积：

为了简化问题，我们对整个表达式求对数，(将指数问题对数化是处理数学问题常见的方法)：

通过梯度下降法对参数进行更新：

**2. LR用到的sigmoid函数是如何推导出来的？**

最大熵原理：对于概率模型而言，在缺乏先验的情况下，条件熵最大的模型是最好的模型。很好理解，熵最大意味着不确定度最高，在没有先验知识情况下，自然这是最好的假设。

目标：使得条件熵最大

这里上面波浪号可以理解为是联合概率密度和样本特征分布已知（根据样本先验可知），最终的目标就是求出让H函数最大的P(Y|X)。

在这里定义一个特征函数的概念，在样本特征与Label之间满足一定关系下，对于样本特征的描述，定义为：。

由贝叶斯相关定理，对于每个特征函数i而言，其期望值显然有 :

而对于最终的条件概率而言，所有label取值概率和为1：

有上面的信息，可以的得到一个等式约束下的拉格朗日最优化问题：

由拉格朗日乘子法，问题转化为求如下式子的最小值：

此时，对L求P(y|x)的导数（指定x和y，去掉求和符号，方便求解），并让导数值为0：

进一步，可以得到：

可以求出：

由所有label概率和为1：

上式合并可得：

上式极为softmax的公式。由于在最大熵模型中，特征函数可以随意定义，因为，定义：

则

因为从分类角度考虑，sigmoid函数其实就是softmax的二分类特例，假定y取值为0和1，且定义

代入上面推导出来的公式，可得：

这个就是所谓的sigmoid函数。

**3. LR的损失函数的实质是什么？**

极大似然函数（使联合概率最大），交叉熵。

**4. LR的优缺点分别是什么？**

优点：计算代价低；速度快；容易理解和实现；适合需要得到一个分类概率的场景。

缺点：容易欠拟合；分类和回归的精度不高；当特征空间很大时，逻辑回归的性能不是很好；依赖于全部的数据特征，当特征有缺失的时候表现效果不好。

**5. sigmoid函数有什么优点？**

①sigmoid的输入范围是，而其结果刚好为，正好满足概率分布为的要求。二分类问题要求用概率去描述分类器，自然比单纯的某个阈值要方便很多；

②sigmoid是一个单调上升的函数，具有良好的连续性，不存在不连续点。

**6. 工业界上使用LR前为什么要对特征进行离散化？有哪些离散化的方法？**

进行离散化的原因：

1. 离散特征的增加和减少都很容易，易于模型的快速迭代。
2. 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. 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。
4. 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。
5. 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，举例，一个变量如果服从对数正态分布，那么对于不同区间，重要度不同。因此通过离散化，可以实现对不同区间学习不同的权重。
6. 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。
7. 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. 使用连续值的LR模型用一个权值去管理该特征，而one-hot后有三个权值管理了这个特征，这样使得参数管理的更加精细，所以这样拓展了LR模型的非线性能力。
9. 这样做了我们至少不用再去对变量进行归一化，也可以加速参数的更新速度；再者使得一个很大权值管理一个特征，拆分成了许多小的权值管理这个特征多个表示，这样做降低了特征值扰动对模型为稳定性影响，也降低了异常数据对模型的影响，进而使得模型具有更好的鲁棒性。

离散化的方法：

1. 无监督方法：分箱法又分为等宽分箱法和等频分箱法，前者制定定长的间隔将特征放入不同箱子内，后者根据频率划分箱子，这两种做法都有一定的缺陷。等宽分箱法对异常点比较敏感，等频分箱法的缺陷是完全按照频率划分会出现实例特征相同却不在同一个箱体内的情况。
2. 基于熵的离散方法：使用合成的方法或者分裂的方法根据熵计算和阈值判定来决定合成或分裂督的方法。

**7. 如何简单介绍下LR模型？**

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，来达到将数据二分类的目的。

**8. LR和SVM之间有什么区别？**

联系：都是线性分类器，模型求解的就是一个超平面；SVM和LR的性能都会收到outlier的影响。

区别：

①支持向量机只考虑局部的边界线附近的点，而逻辑回归考虑全局（远离的点对边界线的确定也起作用）；

②SVM不直接依赖数据分布，分类平面不受一类点影响；LR则受到所有数据点的影响，如果数据不同类别出现强烈的不平衡，一般需要先对数据进行平衡处理；

③SVM依赖数据表达的距离测度，所以需要对数据先做normalization；LR不受其影响。

**9. LR和最大熵模型有哪些联系和区别？**

①逻辑回归跟最大熵模型没有本质区别。逻辑回归是最大熵对应类别为二类时的特殊情况；

②指数簇分布的最大熵等价于其指数形式的最大似然；

③二项式分布的最大熵解等价于二项式指数形式(sigmoid)的最大似然；

④多项式分布的最大熵等价于多项式分布指数形式(softmax)的最大似然。

**10. 逻辑回归的损失函数为什么要使用极大似然函数作为损失函数？**

损失函数一般有四种，平方损失函数，对数损失函数，HingeLoss0-1损失函数，绝对值损失函数。将极大似然函数取对数以后等同于对数损失函数。在逻辑回归这个模型下，对数损失函数的训练求解参数的速度是比较快的。至于原因大家可以求出这个式子的梯度更新，这个式子的更新速度只和相关。和sigmoid函数本身的梯度是无关的。这样更新的速度是可以自始至终都比较的稳定。

为什么不选平方损失函数的呢？因为如果使用平方损失函数，梯度更新的速度和sigmoid函数本身的梯度是很相关的。sigmoid函数在它在定义域内的梯度都不大于0.25。这样训练会非常的慢。

**11. 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。训练以后完以后，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。

**12. 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？**

①去掉高度相关的特征会让模型的可解释性更好。

②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。