**5.逻辑回归与最大熵**

**5.1逻辑回归**

**5.1.1定义**

逻辑回归模型是用输入的线性函数表示输出的对数几率的分类模型。模型服从logistic分布，设X是连续随机变量，服从logistic分布的X具有下列分布函数和密度函数：

其中μ为位置参数，γ > 0为形状参数。函数曲线在中心附近增长速度较快，两端增长速度缓慢。形状参数γ越小曲线在中心附近增长越快。

逻辑回归模型可以用于二分类或多分类，二项逻辑回归模型由条件概率分布P(Y|X)表示，形式为参数化的logistic分布

随机变量X取值为实数，随机变量Y取值为1或0。上式第二个函数也称为sigmoid函数，常被用于神经网络或深度学习的激活函数，所以逻辑回归可以被看作神经网络的一个单元。

二项logistic回归模型也可以推广到多项logistic回归模型，假设离散型随机变量Y={1，2，…，K}，则多项logistic回归模型由以下条件概率分布表示

其中x为输入特征，w为特征的权值。

一个事件的几率（odds）是指该事件发生的概率与该事件不发生的概率的比值。假设事件发生概率为p，则该事件的几率是，该事件的对数几率（logit odds或logit函数）是

对于逻辑回归而言

所以输出Y=1的对数几率是由输入x的线性函数表示。

优点：计算代价不高、易于理解和实现。

缺点：容易欠拟合，分类精度可能不高。

**5.1.2模型参数估计**

逻辑回归模型可以应用**极大似然估计法**或**正则化的极大似然估计法**估计模型参数，设给定训练数据集T = {(x1, y1), (x2, y2), ......, (xN, yN)}（xi∈Rn，yi∈{0，1}），设P(Y=1 | x) = π(x)，P(Y=0 | x) = 1-π(x)，则似然函数为

对数似然函数为

对L(w)求极大值，得到w的估计值，于是模型参数估计就变成了以对数似然函数为目标函数的最优化问题，通常采用**梯度下降法**及**拟牛顿法**。目标函数的梯度为

**5.1.3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，逻辑回归特征离散化的优点：

1）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）；

2）离散化后的特征对异常数据有很强的鲁棒性；

3）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；

4）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；

5）特征离散化后，模型会更稳定。

总之特征离散化的好处在于计算简单、简化模型，增强模型的泛化能力，不易受噪声的影响几个方面。

续性变量转化成离散型变量大致有两类方法：（1）卡方检验方法；（2）信息增益方法。

（1）卡方检验方法：1）分裂方法；2）合并方法：就是找到一个分裂点看，左右2个区间，在目标值上分布是否有显著差异，有显著差异就分裂，否则就忽略。这个点可以每次找差异最大的点。合并类似，先划分如果很小单元区间，按顺序合并在目标值上分布不显著的相邻区间，直到收敛。

（2）信息增益方法1）分裂方法；2）合并方法：这个和决策树的学习很类似。就是找到一个分裂点看，左右2个区间，看分裂前后信息增益变化阈值，如果差值超过阈值（正值，分列前-分裂后信息熵），则分裂。每次找差值最大的点做分裂点，直到收敛。合并类似，先划分如果很小单元区间，按顺序合并信息增益小于阈值的相邻区间，直到收敛。

**5.2最大熵模型**

**5.2.1定义**

最大熵原理是统计学习的一般原理，学习概率模型时，在所有可能的概率模型（分布）中，熵最大的模型是最好的模型，在满足约束条件的模型集合中选取熵最大的模型（一般来说均匀分布时熵最大，最大熵原理通过熵的最大化来表示等可能性）。最大熵模型与logistic回归模型有类似的形式，它们又称为对数线性模型（log linear model）。

假设分类模型是一个条件概率分布P(Y | X)，对于给定的输入X以条件概率P(Y | X)输出Y。给定一个训练集

T = {(x1, y1), (x2, y2), ......, (xN, yN)}

可以确定联合分布P(X，Y)的经验分布和边缘分布P(X)的经验分布，分别以(X，Y)和(X)表示

其中v(X=x，Y=y)表示训练数据中样本(x，y)出现的频数，v(X=x)表示训练数据中输入x出现的频数，N表示训练样本容量。

特征函数f(x，y)描述输入x和输出y之间的某一个事实，其定义是

特征函数f(x，y)关于经验分布(X，Y)的期望值(f)为

特征函数f(x，y)关于模型P(Y | X)和经验分布(X)的期望值Ep(f)为

模型学习的约束条件为(f) = Ep(f)，假如有n个特征函数，就有n个约束条件。假设满足所有约束条件的模型集合为

定义在条件概率分布P(Y | X)上的条件熵为

则模型集合中条件熵H(P)最大的模型称为最大熵模型（式中的对数为自然对数）。

**5.2.2最大熵模型的学习**

最大熵模型的学习可以形式化为约束最优化问题：

s.t. (f) = Ep(f)，i = 1，2，…，n

然后将求最大值问题改写为等价的求最小值问题：

s.t. (f) - Ep(f) = 0，i = 1，2，…，n

再将约束最优化的原始问题转化为无约束最优化的对偶问题，通过求解对偶问题求解原始问题。引入拉格朗日乘子w0，w2，…，wn定义拉格朗日函数L(P,，w)：

最优化的原始问题是

对偶问题是

由于拉格朗日函数L(P，w)是P的凸函数，所以原始问题与对偶问题的解是等价的，所以可以通过求解对偶问题来求解原始问题。求解对偶问题分为两步，首先求解对偶问题内部的极小化问题求得最大熵模型，然后再求解对偶问题外部的极大化问题求解最优模型（最大熵模型）的参数。是w的函数，可以将其记作

Ψ(w)称为对偶函数，将其解记作

具体地，求L(P，w)对P(y | x)的偏导数：

令偏导数等于0，一般(x) > 0的情况下，使第二项为0，解得

因为，得

所以

所以exp(1-w0)起到了归一化的作用，令Zw(x) = exp(1-w0)，得

其中

上式Pw=Pw (y | w)就是最大熵模型。Zw(x)为规范化因子，fi(x，y)是特征函数，wi是特征的权值。之后应用最优化算法求解对偶函数的外部极大化问题学习得到模型的参数w\*用来表示P\*∈C。

最大熵模型的学习归结为对偶函数Ψ(w)的极大化。对偶函数极大化等价于最大熵模型的极大似然估计：

已知训练数据的经验概率分布(X，Y)，条件概率分布P(Y | X)的对数似然函数表示为

当条件概率分布P(Y | X)是最大熵模型时，代入上式可得对数似然函数为

再计算对偶函数Ψ(w)：

比较上面两个式子可知Ψ(w) = 。所以最大熵模型学习中对偶函数的极大化等价于最大熵模型的极大似然估计。

**5.2.3模型学习算法**

Logistic回归和最大熵模型的学习都可以认为是以似然函数为目标函数的最优化问题。常用的优化方法有改进的迭代尺度法、梯度下降法、牛顿法或拟牛顿法等。牛顿法或拟牛顿法因为是二阶梯度的，一般收敛速度更快。

**（1）改进的迭代尺度法（improved iterative scaling，IIS）**

根据上文的证明，最大熵模型的对数似然函数为

IIS的想法是：假设最大熵模型当前的参数向量是w = (w1，w2，…，wn)T，我们希望找到一个新的参数向量w + δ = (w1+δ1，w2+δ2，…，wn+δn)T，使得模型的对数似然函数值增大。如果能有这样一种参数向量更新的方法г：w→w +δ，那么就可以重复使用这一方法，直至找到对数似然函数的最大值。

IIS方法的关键在于如何确定每次参数更新的量δ。对于给定的经验分布，模型参数从w到w + δ，对数似然函数的改变量是

利用不等式

-log α ≥ 1-α，α > 0

得到对数似然函数改变量的下界

令右端记作

于是就有

L(w + δ) – L(w) ≥ A(δ | w)

提高对数似然改变量的下界A(δ | w)，那么对数似然函数也会提高，δ是一个向量，不易同时优化多个变量，IIS每次只优化其中一个变量δi，固定其他变量。IIS引进一个量

f#(x, y)表示所有特征在(x, y)出现的次数。于是A(δ | w)可以改写成

利用指数函数的凸性以及对任意i，有且，根据Jensen不等式，得到

因此A(δ | w)可以改写成

令上式右端记作

于是得到

L (w + δ) – L (w) ≥ B (δ | w)

B(δ | w)是对数似然函数改变量的一个新（相对不紧）下界。求B(δ | w)对δi的偏导数：

令偏导数为0可得

依次对δi求解上面的方程可以求出δ。

如果f#(x, y)是常数即对任何x，y都有f#(x, y)=M，那么δi就可以表示成

如果f#(x, y)不是常数则必须通过数值计算求δi。简单有效的方法就是牛顿法，以g(δi) = 0表示偏导方程，牛顿法通过迭代求得δi\*，使得g(δi\*) = 0。迭代公式是

**改进的迭代尺度算法IIS：**

输入：特征函数f1，f2，…，fn；经验分布(X, Y)，模型Pw(y | x)；

输出：最优参数值wi\*；最优模型Pw\*。

（1）对所有i∈{1，2，…，n}，取初值wi = 0；

（2）对每一i∈{1，2，…，n}：

a）令δi是方程

的解，这里；

b）更新wi值：wi ← wi + δi；

（3）如果不是所有wi都收敛，重复步（2）。

**（2）拟牛顿法**

对于最大熵模型而言

目标函数：

梯度：

其中

**拟牛顿法（BFGS算法）：**

输入：特征函数f1，f2，…，fn；经验分布(x, y)，目标函数f(w)，梯度函数g(w)= ▽f(w)，精度要求ε；

输出：最优参数值w\*；最优模型Pw\*(y | x)。

（1）选定初始点w(0)，取B0为正定对称矩阵，置k=0；

（2）计算gk = g(w(k))，若||gk|| < ε，则停止计算，得w\* = w(k)，否则转到（3）；

（3）由Bkpk = -gk求出pk；

（4）一维搜索：求λk使得

（5）置w(k+1) = w(k) + λkpk（6；

（6）计算gk+1 = g(w(k+1))，若||gk+1|| < ε，则停止计算，得w\* = w(k+1)；否则按下式求出Bk+1：

其中yk = gk+1 - gk，δk = w(k+1) + w(k)。

（7）置k = k+1，转（3）。