用于分类的线性模型

- 分类的目标：将输入变量x分到K个离散的类别C_k中的某一类。我们考虑分类的线性模型，也就是指决策面是输入向量x的线性函数，因此被定义为D维输入空间中的$(D-1)$维超平面。如果数据集可以被线性决策面精确分类，那么就称这个数据集是线性可分的。
- 分类的不同方法：
  - 构造判别函数，直接将向量x分到具体类别当中
  - 在推断阶段对条件概率$p(C_k|x)$直接建模，然后使用该概率分布进行最优决策。其中：
    - 一种方法直接对条件概率进行建模，将条件概率分布表示为参数模型，然后使用训练集来优化参数。
    - 另一种是采用生成式的方法，对类条件概率密度$p(x|C_k)$以及类的先验概率分布$p(C_k)$进行建模，然后使用贝叶斯定理计算后验概率分布：

$$p(C_k|x)=\frac{p(x|C_k)p(C_k)}{p(x)}$$

对于分类问题，我们想得到的是离散的类别标签，或者预测位于$(0,1)$之间的后验概率分布。所以我们使用非线性函数$f(.)$对w的线性函数进行变换，即

$$y(x)=f(w^Tx+w_0)$$

其中$f(.)$被称为激活函数，而其反函数被称为链接函数。决策面对应于$y(x)=$常数,因此，决策面是x的线性函数，因此y函数被称为推广的线性模型。但是不再是参数的线性模型。

4.1 判别函数

4.1.1 二分类

线性函数的最简形式：

$$y(x)=w^Tx+w_0$$

其中w称为权向量，w_0被称为偏置。偏置的相反数有时被称为阈值。对于一个输入向量x，如果$y(x)\geqslant 0$则被分到C_1,否则被分到C_2中。决策边界为$y(x)=0$,对于两个决策面上的点x_A,x_B，由于$y(x_A)=y(x_B)=0$,所以$w^T(x_A-x_b)=0$,所以w是决策面的法向量，所以从原点到决策面的距离为

$$\frac{w^Tx}{||w||}=-\frac{w_0}{||w||}$$

对于任意一点x以及它在决策面上的投影$x_\perp$,我们设x到决策面的距离为r，则可以得到

$$x=x_\perp+r\frac{w}{||w||}$$

所以

$$r=\frac{y(x)}{||w||}$$

另注：为了使符号更加简洁，我们引入一个虚输入$x_0=1$,定义$\tilde{w}=(w_0,w)$以及$\tilde{x}=(x_0,x)$,从而 $y(x)=\tilde{w}^T\tilde{x}$

4.1.2 多分类

可以使用K-1个分类器，每一个都是二分类器，分别表示是否属于该类，这被称作“1对其他”分类器。
另一种方法是引入$\frac{K(K-1)}{2}$个二元判别函数，被称为"1对1"分类器，每一个点的类别根据判别函数的大多数输出类确定。
我们引入K类判别函数，由K个线性判别函数组成：

$$y_k(x) = w_k^T x + w_{k0}$$

如果对于点x，对于所有的$j \neq k$,都有$y_k(x) > y_j(x)$,就将它分为C_k.于是C_j与C_k的决策面为$y_k(x) = y_j(x)$,对应于一个$D-1$维超平面，形式为

$$(w_k - w_j)^T x + (w_{k0} - w_{j0}) = 0$$

4.1.3 用于分类的最小平方方法

使用向量记号，可以很容易的将这些量聚集在一起，即

$$y(x) = \tilde{W}^T \tilde{x}$$

这样一个新输入量会被分配到输出$y_k = \tilde{w}_k^T x \tilde{x}$最大的类别中。
现在最小化平方误差函数来确定参数矩阵$\tilde{W}$,平方误差函数可以写作

$$E_D(\tilde{W}) = \frac{1}{2}\mathbf{Tr}\{(\tilde{X}\tilde{W} - T)^T(\tilde{X}\tilde{W} - T)\}$$

求导可得：

$$\tilde{W} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T T = \tilde{X}^\dagger T$$

其中$\tilde{X}^\dagger$是矩阵$\tilde{X}$的伪逆矩阵，判别函数为

$$y(x) = \tilde{W}^T\tilde{x} = T^T(\tilde{X}^\dagger)^T\tilde{x}$$

另外，如果训练集中每一个目标向量都满足线性性质

$$a^T t_n + b = 0$$

其中a和b为常数，那么对于任意的x值，都有：

$$a^T y(x) + b = 0$$

最小平方方法对于判别函数的参数给出精确的解析解，但是对于离群点缺少鲁棒性。

4.1.4 Fisher线性判别函数

将D维输入向量x使用下式投影到一维

$$y = w^T x$$

在y上设置阈值将$y \geq -w_0$的样本分为C_1类，其余样本分为C_2类。
对于C_1中N_1个点，均值向量

$$m_1 = \frac{1}{N_1}\sum_{n\in C_1} x_n$$
$$m_2 = \frac{1}{N-2}\sum_{n\in C_2} x_n$$

投影到w上后，我们想要使得两类投影之间均值相差较大，但是类内的方差较小。而对于类别C_k的数据点变换后的类内方差为

$$s_k^2 = \sum_{n\in C_k}(y_n - m_k)^2$$

Fisher准则为：

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

另外定义

$$\begin{aligned} S_B &= (m_2 - m_1)(m_2 - m_1)^T \\ S_W &= \sum_{n\in C_1}(x_n - m_1)(x_n - m_1)^T + \sum_{n\in C_2}(x_n - m_2)(x_n - m_2)^T \\ J(w) &= \frac{w^T S_B w}{w^T S_W w} \end{aligned}$$

所以求导数可以得到，$J(w)$取的最大值的条件为：

$$(w^T S_B S_W w) S_W w = (w^T S_W w) S_B w$$

而$S_B w$总是在$(m_2 - m_1)$的方向上，我们不关心w的大小，而是关心其方向，所以忽略标量因子$(w^T S_B w)$和$(w^T S_W w)$,我们可以得到：

$$w \propto S_W^{-1}(m_2 - m_1)$$

如果类内协方差矩阵是各向同性的，而S_W正比于单位矩阵，那么我们看到w正比于类均值的差。

4.1.5 与最小平方的关系

最小平方方法确定线性判别函数的目标是使模型的预测尽可能地与目标值接近。相反，Fisher判别准则的目标是使输出空间的类别有最大的区分度。对于二分类问题，Fisher准则可以看成最小平方的一个特例。

4.1.6 多分类Fisher线性判别函数

将类内协方差矩阵推广到K类，可以得到：

$$S_W = \sum_{k=1}^{K} S_k$$

其中

$$S_k = \sum_{n\in C_K}(x_n - m_k)(x_n - m_k)^T$$

$$m_k = \frac{1}{N_k}\sum_{n\in C_k} x_n$$

整体的协方差矩阵

$$S_T = \sum_{n=1}^{N}(x_n - m)(x_n - m)^T$$
$$m = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{k=1}^{K} N_k m_k$$

所以可以得到：

$$S_T = S_B + S_W$$
$$S_B = \sum_{k=1}^{K}\sum_{n\in C_k} N_k(m_k - m)(m_k - m)^T$$

我们在投影后的D'维y空间中定义类似的矩阵：

$$s_W = \sum_{k=1}^{K}\sum_{n\in C_k} N_k(y_n - \mu_k)(y_n - \mu_k)^T$$
$$\mu_k = \frac{1}{N_k}\sum_{n\in C_k} y_n$$
$$\mu = \frac{1}{N}\sum_{k=1}^{K} N_k\mu_k$$
$$s_B = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T$$

此时我们选择的准则为$J(w) = \mathbf{Tr}\{s_W^{-1}s_B\}$,$J(W) = \mathbf{Tr}\{(W^T S_W W)^{-1}(W^T S_B W)\}$.
值得强调的时，有一个重要的结果对于所有的这些判别准则都成立。首先S_B 由K个矩阵的和组成，每一个矩阵都是两个向量的外积，因此秩等于1。此外，由于这些矩阵中只有(K -1)个是相互独立的。因此S_B 的秩最大等于(K -1)，因此最多有(K - 1)个非零特征值。这表明，向由S_B 张成的(K - 1)维空间上的投影不会改变$J(W)$的值，因此通过这种方法我们不能够找到多于(K - 1)个线性“特征”.

4.1.7 感知器算法

感知器收敛定理（perceptron convergence theorem）表明，如果存在一个精确的解（即，如果训练数据线性可分），那么感知器算法可以保证在有限步骤内找到一个精确解。
需要注意的是，达到收敛状态所需的步骤数量可能非常大，并且在实际应用中，在达到收敛状态之前，我们不能够区分不可分问题与缓慢收敛问题。
即使数据集是线性可分的，也可能有多个解，并且最终哪个解会被找到依赖于参数的初始化以及数据点出现的顺序。此外，对于线性不可分的数据集，感知器算法永远不会收敛。

4.2 概率生成式模型

考虑二分类的情况下，类别C_1的后验概率可以写成

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + P(x|C_2)p(C_2)}$$
$$= \frac{1}{1+e^{-a}} = \sigma(a)$$

其中

$$a = \ln \frac{p(x|C_1)p(C_1)}{P(x|C_2)p(C_2)}$$

σ函数满足以下性质：

$$\sigma(-a) = 1 - \sigma(a)$$

logistic sigmod 的反函数被称为logit函数，表示两类别的概率比值的对数$\ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$也被称为log odds 函数：

$$a = \ln \frac{\sigma}{1-\sigma}$$

对于K>2个类别的情况，我们可以得到:

$$\begin{aligned} p(C_k|x) &= \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} \\ &= \frac{\exp(a_k)}{\sum_j \exp(a_j)} \end{aligned}$$

它被称为归一化指数，可以被当作logistic sigmod 函数对于多类情况的推广。这里，$a_k = \ln p(x|C_k)p(C_k)$，归一化指数也被称为softmax函数。

4.2.1 连续输入

假设类条件概率密度是高斯分布，然后我们假设所有类别的协方差矩阵相同，这样类别C_k的类条件概率为：

$$p(x|C_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k))$$

首先考虑两类别的情形，可以得到：

$$p(C_1|x) = \sigma(w^Tx + w_0)$$

其中已经定义:

$$\begin{aligned} w &= \Sigma^{-1}(\mu_1 - \mu_2) \\ w_0 &= -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{p(C_1)}{p(C_2)} \end{aligned}$$

对于K个类别的一半情形，我们可以得到：

$$a_k(x) = w_k^Tx + w_{k0}$$

其中我们定义了

$$\begin{aligned} w_k &= \Sigma^{-1}\mu_k \\ w_{k0} &= -\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln p(C_k) \end{aligned}$$

由于我们假设各个类别的协方差矩阵相同，决策边界是线性的，最终的决策边界，对应于最小错误分类率，会出现在后验概率最大的两个概率相等的位置，因此由x的线性函数定义，从而再次得到了一个一般的线性模型。
如果不假设各个类别的协方差矩阵相同，则我们会得到x的二次函数，引出了二次判别函数。

4.2.2 最大似然解

首先考虑两个类别的情形，每一个类别都有一个高斯类条件概率密度，且协方差矩阵相同。假设我们有数据集$\{x_n, t_n\}$,将先验概率$p(C_1) = \pi, p(C_2) = 1 - \pi$,对于一个属于类别一的数据点$x_n$，其$t_n = 1$,所以

$$p(x_n, C_1) = p(C_1)p(x|C_1) = \pi\mathcal{N}(x_n|\mu_1, \Sigma)$$

类似的，我们可以得到：

$$p(x_n, C_2) = p(C_2)p(x|C_2) = (1 - \pi)\mathcal{N}(x_n|\mu_2, \Sigma)$$

于是似然函数为：

$$p(\mathbf{t}, X|\pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N}[\pi\mathcal{N}(x_n|\mu_1, \Sigma)]^{t_n}[(1 - \pi)\mathcal{N}(x_n|\mu_2, \Sigma)]^{1-t_n}$$

我们取对数，对于π的最大化而言，取对数得到

$$\sum_{n=1}^{N}\{t_n \ln \pi + (1 - t_n) \ln (1 - \pi)\}$$

求导，使之导数为0，可以得到：

$$\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N}$$

同时，我们对于μ_1进行最大化，可以得到：

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}(x_n|\mu_1, \Sigma) = -\frac{1}{2}\sum_{n=1}^{N} t_n(x_n - \mu_1)^T\Sigma^{-1}(x_n - \mu_1) + 常数$$

解得：

$$\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n x_n, \mu_2 = \frac{1}{N_2}\sum_{n=1}^{N}(1 - t_n)x_n$$

同样地，对于Σ,我们可以得到：

$$-\frac{1}{2}\sum_{n=1}^{N} t_n \ln |\Sigma| - \frac{1}{2}\sum_{n=1}^{N} t_n(x_n - \mu_1)^T\Sigma^{-1}(x_n - \mu_1) - \frac{1}{2}\sum_{n=1}^{N}(1 - t_n)(x_n - \mu_2)^T\Sigma^{-1}(x_n - \mu_2) - \frac{1}{2}\sum_{n=1}^{N}(1 - t_n)\ln |\Sigma|$$
$$= -\frac{N}{2}\ln |\Sigma| - \frac{N}{2}\mathbf{Tr}\{\Sigma^{-1}S\}$$

其中

$$S = \frac{N_1}{N}S_1 + \frac{N_2}{N}S_2$$
$$S_1 = \frac{1}{N_1}\sum_{n\in C_1}(x - \mu_1)(x - \mu_1)^T$$
$$S_2 = \frac{1}{N_2}\sum_{n\in C_2}(x - \mu_2)(x - \mu_2)^T$$

使用高斯分布的最大似然解的标准结果，我们可以看到$\Sigma = S$,表示一个对于两类都有关系的协方差矩阵求加权平均。
这个结果很容易推广到K类，但是拟合类高斯分布的方法对于离群点并不鲁棒，因为高斯的最大似然估计是不鲁棒的。

4.2.3 离散特征

我们做出朴素贝叶斯的假设，特征值看成是相互独立的，以类别C_k为条件，因此得到类条件分布的形式为：

$$p(x|C_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$

其中对于某个类别，都有D个独立的参数。我们可以得到：

$$a_k(x) = \sum_{i=1}^{D}\{x_i \ln \mu_{ki} + (1-x_i)\ln \mu_{ki}\} + \ln p(C_k)$$

与之前一样，这是输入x_i的线性函数。

4.2.4 指数族分布

可以看到，将注意力集中于$u(x) = x$的分布上，然后引入一个缩放参数s，这样就得到指数族类条件概率分布的一个子集：

$$p(x|\lambda_k, s) = \frac{1}{s}h(\frac{1}{s}x)g(\lambda_k)\exp(\frac{1}{s}\lambda_k^T x)$$

每一个类别的参数向量是不同的λ_k,但是缩放参数s是一样的。对于二分类问题，将类条件概率密度表达式带入到$a = \frac{\ln p(x|C_1)p(C_1)}{\ln p(x|C_2)p(C_2)}$,得到：

$$a(x) = \frac{1}{s}(\lambda_1 - \lambda_2)^T x + \ln g(\lambda_1) - \ln g(\lambda_2) + \ln p(C_1) - \ln p(C_2)$$

对于K类问题，可以得到：

$$a_k(x) = \frac{1}{s}\lambda_k^T x + \ln g(\lambda_k) + \ln p(C_k)$$

4.3 概率判别式模型

我们可以使用一般的线性模型的函数形式，然后直接使用最大似然估计得到参数。在寻找一般的线性模型参数时，一种间接的方式是分别寻找类条件概率密度和类别先验概率，然后使用贝叶斯定理就可以求出后验类概率。
在直接方法中，我们最大化由条件概率分布定义的似然函数，这种方法代表了判别式方法的一种形式。判别式方法的优点是有更少的可调节参数需要确定，预测表现也会提升，尤其是当类条件概率密度的假设没有很好地近似真实的分布时更是如此。

4.3.1 固定基函数

我们可以使用一个基函数向量$\phi(x)$对于输入变量进行一个固定的非线性变换，应用到算法中依然是成立的，最终的决策边界在特征空间中是线性的因此对应于原始x空间中的非线性决策边界。基函数中通常会设置为一个常数，例如$\phi_0(x) = 1$,使得对应的参数w_0扮演偏置的作用。恰当选择非线性变换会让后验概率建模过程更简单。

4.3.2 logistic 回归

首先讨论二分类问题，可以看到在一般假设条件下，类别C_1的后验概率可以写成作用在ϕ的线性函数上的 logistic sigmod 函数的形式：

$$p(C_1|\phi) = y(\phi) = \sigma(w^T\phi)$$
$$p(C_2) = 1 - p(C_1|\phi)$$

对于一个M维特征空间ϕ,这个模型有M个可调节参数，相反如果使用最大似然方法来调节高斯类条件概率密度，那么我们有$2M$个参数来描述均值，$\frac{M(M+1)}{2}$个参数来描述共享协方差矩阵，总参数随着M增长而二次增长。

对于logistic sigmod 函数，其导数为:

$$\frac{\partial \sigma}{\partial a} = \sigma(1-\sigma)$$

对于一个数据集，似然函数可以写为：

$$p(\mathbf{t}|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}$$

然后通过取负对数的形式，可以得到交叉熵误差函数;

$$E(w) = -\ln p(\mathbf{t}|w) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln 1 - y_n\}$$

其中$y_n = \sigma(a_n), a_n = w^T\phi(x)$，两侧对w求梯度，可以得到：

$$\nabla E(w) = \sum_{n=1}^{N}(y_n - t_n)\phi_n$$

我们可以根据上面的结果对参数的值进行更新。，可以看到的是，数据点对于梯度的贡献为目标值和模型预测值之间误差与基函数向量的乘积，函数形式与线性回归模型中平方和误差函数的梯度的函数形式完全相同。
但是，最大似然方法对于现行可分的数据集会产生严重的过拟合。

4.3.3 迭代重加权最小平方

误差函数可以通过一种高效的迭代方法求出最小值，这种方法使用了对数似然函数的局部二次近似，更新的形式为：

$$w^{新} = w^{旧} - H^{-1}\nabla E(w)$$

其中H是一个hessian矩阵，元素由$E(w)$关于w的二阶导数组成。
将该方法用于线性回归模型中，误差函数为平方和误差函数，其梯度和Hessian矩阵为

$$\nabla E(w) = \sum_{n=1}^{N}(w^T\phi(x) - t_n)\phi_n = \Phi^T\Phi w - \Phi^T\mathbf{t}$$
$$\mathbf{H} = \nabla\nabla E(w) = \sum_{n=1}^{N}\phi_n\phi_n^T = \Phi^T\Phi$$

所以

$$w^{新} = w^{旧} - (\Phi^T\Phi)^{-1}\{\Phi^T\Phi w^{旧} - \Phi^T\mathbf{t}\} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$$

这就是标准的最小平方解。

我们把它应用到logistic回归模型的交叉熵误差函数，可以看到其梯度和hessian矩阵分别为：

$$\nabla E(w) = \sum_{n=1}^{N}(y_n - t_n)\Phi^T(y-t)$$
$$H = \nabla\nabla E(w) = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^T = \Phi^T R\Phi$$

并且我们引入了一个N*N的对角矩阵R,元素为

$$R_{nn} = y_n(1-y_n)$$

可以看到，Hessian不再是常数，而是依赖于w的值，而对于logistic sigmod 的结果y_n来说，$0 < y_n < 1$,可以看到对于任意的向量u,都有$u^T H u > 0$，所以Hessian矩阵是正定的，所以误差函数是w的一个凸函数，所以有唯一的最小值。
所以更新公式如下：

$$\begin{aligned} w^{新} &= w^{旧} - (\Phi^T R\Phi)^{-1}\Phi^T(\mathbf{y}-\mathbf{t}) \\ &= (\Phi^T R\Phi)^{-1}\{\Phi R\Phi w^{旧} - \Phi^T(\mathbf{y}-\mathbf{t})\} \\ &= (\Phi^T R\Phi)^{-1}\Phi^T R\mathbf{z} \end{aligned}$$

其中**z**是一个N维向量，元素为：

$$\mathbf{z} = \Phi w^{旧} - R^{-1}(\mathbf{y}-\mathbf{t})$$

由于权矩阵R不是常量，而是依赖于参数w,所以必需迭代更新参数，该算法被称为迭代重加权最小平方，或者简称为IRLS。与加权最小平方问题一样，对角矩阵R可以看作是方差，因为 logistic 回归模型的均值和方差是

$$\mathbb{E}[t] = \sigma(x) = y$$
$$\mathbf{var}[t] = \mathbb{E}[t^2] - \mathbb{E}[t]^2 = \sigma(x) - \sigma(x)^2 = y(1-y)$$

其中，当$t \in \{0,1\}$，$t = t^2$,事实上，可以将IRLS看成变量空间$a = w^T\phi$的线性问题的解。这样，**z**的第n个元素z_n就可以看作是这个空间中的有效目标值,z_n可以通过对当前操作点$w^{旧}$附近的logistic sigmod函数的局部线性近似方法得到：

$$\begin{aligned} a_n(w) &\simeq a_n(w^{旧}) + \frac{\partial a_n}{\partial y_n}|w^{旧}(t_n - y_n) \\ &= \phi_n^T w^{旧} - \frac{y_n - t_n}{y_n(1-y_n)} \\ &= z_n \end{aligned}$$

4.3.4 多类logistic回归

对于多分类问题，后验概率：

$$p(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

其中，激活$a_k = w^T\phi$
我们想要用最大似然方法直接确定确定这个模型中的参数$\{w_k\}$,为此，我们需要求出y_k关于所有激活a_j的导数，导数为：

$$\frac{y_k}{a_j} = y_k(I_{kj} - y_j)$$

其中I_{kj}为单位矩阵的元素。
接下来对于似然函数我们使用onehot编码，因此对于属于C_k的特征向量ϕ_k的目标函数t_n是一个二元向量，其第k个元素为1，其余元素为0，因此，似然函数：

$$p(T|w_1,\cdots,w_K)=\prod_{n=1}^{N}\prod_{k=1}^{K}p(C_k|\phi_n)^{t_{nk}}=\prod_{n=1}^{N}\prod_{k=1}^{K}y_{nk}^{t_nk}$$

其中$y_{nk}=y_k(\phi_n)$,T为目标变量的一个N*K的矩阵，取负对数：

$$E(w_1,\cdots,w_K)=-\ln p(T|w_1,\cdots,w_K)=-\sum_{n=1}^{N}\sum_{k=1}^{K}t_{nk}\ln y_{nk}$$

它被称为多分类的交叉熵误差函数

给出导数：

$$\nabla_{w_j}E(w_1,\cdots,w_K)=\sum_{n=1}^{K}(y_{nj}-t_{nj})\phi_n$$

其中，我们使用了性质

$$\sum_k t_{nk}=1$$

为了找到一个批处理算法，我们再次使用Newton-Raphson更新来获得多类问题的对应IRLS算法，这需要求出大小M*M的hessian矩阵，其中块i，j 为：

$$\nabla_{w_i}\nabla_{w_j}E(w_1,\cdots,w_K)=\sum_{n=1}^{N}y_{nk}(I_{kj}-y_{nj})\phi_n\phi_n^T$$

多类logistic回归的Hessian矩阵依然是正定的，因此误差函数有唯一的最小值。

4.3.5 probit 回归

对于很多类条件概率密度而言，其并没有很简单的后验概率函数形式，我们对于二分类的情况，使用一般的线性模型的框架：

$$p(t=1|a)=f(a)$$

其中，$a=w^T\phi$.
对于每一个输入ϕ_n,我们选择$a_n=w^T\phi_n$,然后按以下方式设置目标值：

$$\begin{cases}t_n=1, & \text{如果}a_n\geq\theta\\ t_n=0, & \text{其他情况}\end{cases}$$

如果θ从概率密度$p(\theta)$中抽取，则对应激活函数为：

$$f(a)=\int_{-\infty}^{a}p(\theta)\mathrm{d}\theta$$

假设$p(\theta)$是一个0均值，单位方差的高斯概率密度，对应的累计分布函数为：

$$\Phi(a)=\int_{-\infty}^{a}\mathcal{N}(\theta|0,1)\mathrm{d}\theta$$

这被称为逆probit函数，而对于

$$\mathbf{erf}(a)=\frac{2}{\sqrt{\pi}}\int_0^a \exp(-\theta^2)\mathrm{d}\theta$$

有这样的关系：

$$\Phi(a)=\frac{1}{2}\{1+\mathbf{erf}(\frac{a}{\sqrt{2}})\}$$

而基于probit激活函数的一般的线性模型被称为probit回归。

注意： 关于离群点，由于它们在错误的一侧距离理想的决策边界相当远，所以会干扰分类器，而logistic 回归模型和probit回归模型的表现是不同的，因为对于$x\to\infty$,前者会像$\exp(-x)$那样渐进衰减，而probit激活函数则会像$\exp(-x)$那样衰减，因此会更加敏感。
二者都是假设数据点是正确标记的，错误标记的影响会合并到概率模型当中，我们引入目标值t被翻转到错误值的概率ϵ,这时数据点x的目标值分布为：

$$\begin{aligned}p(t|x)&=(1-\epsilon)\sigma(x)+\epsilon(1-\sigma(x))\\&=\epsilon+(1-2\epsilon)\sigma(x)\end{aligned}$$

其中$\sigma(x)$是输入向量x的激活函数这里，ϵ可以事先设定，也可以从数据中推断。

4.3.6 标准链接函数

对于高斯噪声分布的线性回归模型，误差函数，对应于负对数似然函数。如果我们对数据点n对误差函数的贡献关于参数向量w求导数，那么导数的形式为“误差”y_n-t_n与特征向量φ_n的乘积，其中$y_n=w^T\varphi_n$。类似地，对于logistic sigmoid激活函数与交叉熵误差函数的组合，以及多类交叉熵误差函数的softmax激活函数，我们再次得到了同样的简单形式。现在我们证明，如果假设目标变量的条件分布来自于指数族分布，对应的激活函数选为标准链接函数（canonical link function），那么这个结果是一个一般的结果。
我们再次使用指数族分布的限制形式。注意，这里我们把指数族分布的假设应用于目标变量t，而不是应用于输入向量x。于是，我们考虑目标变量的条件分布。

$$p(t|\eta,s)=\frac{1}{s}h(\frac{t}{s})g(\eta)\exp(\frac{\eta t}{s})$$

t的条件均值

$$y=\mathbb{E}[t|\eta]=-s\frac{\partial\ln g(\eta)}{\partial\eta}$$

我们将这个关系记作：

$$\eta=\varphi(y)$$

我 们 将 一 般 线 性 模 型定义为这样的模型：y是输入变量（或者特征变量）的线性组合的非线性函数，即

$$y=f(w^T\phi)$$

其中f为激活函数，f^{-1}为链接函数。
对该模型的对数似然：

$$\ln p(\mathbf{t}|\eta,s)=\sum_{n=1}^N\ln p(t_n|\eta,s)=\sum_{n=1}^N\{\ln g(\eta_n)+\frac{\eta_n t_n}{s}\}+\text{常数}$$

其中，我们假设所有观测有相同的缩放参数，因此s与n无关，对数似然函数关于模型参数w的导数为：

$$\nabla_w \ln p(\mathbf{t}|\eta, s) = sum_{n=1}^{N}\{\frac{\partial \ln g(\eta_n)}{\partial \eta_n} + \frac{t_n}{s}\}\frac{\partial \eta_n}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n$$
$$= \sum_{n=1}^{N}\frac{1}{s}\{t_n - y_n\}\varphi'(y_n)f'(a_n)\phi_n$$

其中$a_n = w^T\phi_n$,使用链接函数:

$$f^{-1}(y) = \varphi(y)$$

上式表明$f(\varphi(y)) = y$,$f'(\varphi)\varphi'(y) = 1$,由于$a = f^{-1}(y)$,所以$a = \varphi$,所以$f'(a)\varphi(y) = 1$,误差函数梯度化简为:

$$\nabla E(w) = \frac{1}{s}\sum_{n=1}^{N}\{y_n - t_n\}\phi_n$$

对于高斯分布，$s = \beta^{-1}$,对于logistic模型，s=1.

4.4 拉普拉斯近似

当讨论logistic回归的贝叶斯观点时，由于后验概率分布不再是高斯分布，所以不能精确地关于参数向量x求积分，所以我们使用拉普拉斯近似。
首先考虑单一连续变量z，假设分布$p(z)$的定义为

$$p(z) = \frac{1}{Z}f(z)$$

其中$Z = \int f(z)\mathrm{d}z$是归一化系数，我们假定它是未知的。拉普拉斯近似中要寻找一个高斯近似$q(z)$，其中心在$p(z)$的众数的位置。
首先，我们寻找$p(z)$的众数，也就是寻找一个点z_0使得$p'(z_0) = 0$,或者等价的

$$\frac{\partial f(z)}{\partial z}|_{z=z_0} = 0$$

高斯分布有一个性质，就是它的对数是变量的二次函数，于是我们考虑$\ln f(z)$以众数z_0为中心的泰勒展开，即:

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2}A(z - z_0)^2$$

这里由于z_0处是最大值，所以一阶项消失，两侧同时取指数，可以得到:

$$f(z) = f(z_0)\exp(-\frac{A}{2}(z - z_0)^2)$$

使用归一化后高斯分布的标准形式，可以得到:

$$q(z) = (\frac{A}{2\pi})^{\frac{1}{2}}\exp(-\frac{A}{2}(z - z_0)^2)$$

注意: 高斯近似只在精度A>0时有良好的定义，所以驻点z_0一定是一个局部最大值，使得$f(z)$在z_0处二阶导数为负。

将拉普拉斯方法推广到M维空间z上的概率分布$p(z) = \frac{f(z)}{z}$.在驻点z_0处，梯度会消失，在驻点处展开，我们可以得到:

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2}(z-z_0)^T A(z-z_0)$$

其中M*M的Hessian矩阵A的定义为：

$$A = -\nabla\nabla \ln f(z)|_{z=z_0}$$

两边同时取指数，可以得到：

$$f(z) \simeq f(z_0)\exp\{-\frac{1}{2}(z-z_0)^T A(z-z_0)\}$$

分布$q(z)$是正比于$f(z)$,归一化系数可以通过观察归一化的多元高斯分布的标准形式得到：

$$q(z) = \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}}\exp(-\frac{1}{2}(z-z_0)^T A(z-z_0)) = \mathcal{N}(z|z_0, A^{-1})$$

其中$|A|$是A的行列式，高斯分布有良好定义的前提是，精度矩阵是正定的。表明驻点z_0是局部最大值，而非局部最小值或者是鞍点。
使用拉普拉斯近似首先要寻找众数，然后计算众数位置上的Hessian矩阵。实际使用时我们要近似的许多概率分布都是多峰的，因此考虑的众数不同，拉普拉斯近似也不同。
要注意的是，真实概率分布的归一化常数Z不需要事先知道。根据中心极限定理，可以预见的是随着观测数据点的增加，模型的后验概率会越来越近似于高斯分布，因此，数据点越多，拉普拉斯近似越有用。

- 拉普拉斯近似的一个缺点：由于它以高斯分布为基础，所以它只能直接用于实值变量，在其他情况下，它可以将拉普拉斯近似应用于变换之后的变量上。例如，如果$0 < \tau < +\infty$,那么我们可以考虑$\ln \tau$的拉普拉斯近似。但是拉普拉斯近似最严重的局限性在于完全依赖于真实概率分布在变量某个位置上的性质，因此无法描述一些重要的全局属性。

4.4.1 模型比较和BIC

除了近似概率分布$p(z)$,我们也可以获得对归一化常数Z的一个近似，我们可以得到：

$$\begin{aligned} Z &= \int f(z)\mathrm{d}z \\ &\simeq f(z_0)\int \exp(-\frac{1}{2}(z-z_0)^T A(z-z_0))\mathrm{d}z \\ &= f(z_0)\frac{(2\pi)^{\frac{M}{2}}}{|A|^{\frac{1}{2}}} \end{aligned}$$

可以使用上式来获得对于模型证据的一个近似。
对于一个数据集D以及一组模型$\{M_i\}$以及模型参数$\{\theta_i\}$.对于每一个模型，我们定义一个似然函数$p(D|\theta_i, M_i)$,模型证据为：

$$p(D|M_i) = \int p(D|\theta, M_i)p(\theta)\mathrm{d}\theta$$

令$f(\theta) = p(D|\theta, M_i), Z = p(D|M_i)$,我们可以得到：

$$\ln p(D|M_i) \simeq \ln p(D|\theta_{MAP}, M_i) + \ln p(\theta_{MAP}) + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|A|$$

第一项表示使用最优参数计算的对数似然值，其后三项由“Occam因子”组成，它对模型的复杂度进行惩罚。
其中θ_{MAP}是在后验概率分布众数上的位置的θ值，A是负对数后验概率的二阶导数组成的Hessian矩阵：

$$A = -\nabla\nabla \ln p(D|\theta_{MAP}, M_I)p(\theta_{MAP}) = -\nabla\nabla \ln p(\theta_{MAP}|D, M_i)$$

如果我们假设参数的高斯先验分布较宽，而且Hessian矩阵满秩，那么：

$$\ln p(D) \simeq \ln p(D|\theta_{MAP}) - \frac{1}{2}M \ln N$$

其中N为数据的总数，M为θ中参数的数目，省略了一些额外的常数，这被称为贝叶斯信息准则（BIC），与AIC相比，该信息准则对于模型复杂度的惩罚更严重。
但是对于AIC和BIC这样的复杂性度量很容易计算，但也会有误导性结果，因为Hessian矩阵满秩的假设通常是不成立的，因此许多参数不能良好的确定。我们可以使用拉普拉斯近似来获得对于模型证据的一个更准确的估计。

4.5 贝叶斯logistic回归

对于**logistic**回归的贝叶斯观点。对于logistic回归，精确的贝叶斯推断是无法处理的，计算后验概率分布需要对先验概率分布与似然函数的乘积进行归一化，而似然函数本身有一系列 **logistic sigmod** 函数的乘积组成，每一个数据点都有一个 **logistic sigmod** 函数。对于预测分布的计算类似的也是无法处理的。我们采用拉普拉斯近似来处理贝叶斯 **logistic** 回归的问题。

4.5.1 拉普拉斯近似

拉普拉斯近似：首先寻找后验概率分布的众数，然后调节一个以众数为中心的高斯分布，这需要计算对数后验概率的二阶导数，这等价于寻找Hessian矩阵。
由于我们寻找后验概率分布的高斯表示，因此我们在一开始选择一个高斯分布的先验，我们将先验写成一般的形式：

$$p(w) = \mathcal{N}(w|m_0, S_0)$$

m_0, S_0是固定的超参数,w的后验概率分布为:

$$p(w|\mathbf{t}) \propto p(w)p(\mathbf{t}|w)$$

其中$\mathbf{t} = \{t_1, \cdots, t_N\}^T$.两侧取对数后，可以得到：

$$\ln p(\mathbf{t}|w) = -\frac{1}{2}(w - m_0)^T S_)^{-1}(w - m_0) + \sum_{n=1}^{N}\{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\} + 常数$$

其中$y_n = \sigma(w^T\phi_n)$,为了得到后验概率的高斯近似，我们首先最大化后验概率分布，得到MAP（最大后验）解w_{MAP},它定义了高斯分布的均值，协方差就是负对数似然函数的二阶倒数矩阵的逆矩阵，形式

$$S_N^{-1} = -\nabla\nabla \ln p(w|\mathbf{t}) = S_0^{-1} + \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^T$$

于是后验概率分布的高斯近似为：

$$q(w) = \mathcal{N}(w|w_{MAP}, S_N)$$

得到了后验概率分布的高斯近似之后，接下来对这个概率分布求积分来进行预测。

4.5.2 预测分布

给定一个新的特征向量$\phi(x)$,类别C_1的预测分布可以通过对后验概率$p(w|\mathbf{t})$进行积分，后验概率本身由高斯分布$q(w)$近似，也就是：

$$p(C_1|\phi, \mathbf{t}) = \int p(C_1|\phi, w)p(w|\mathbf{t})\mathrm{d}w \simeq \int \sigma(w^T\phi)q(w)\mathrm{d}w$$

且类别C_2对应的概率是

$$p(C_2|\phi, \mathbf{t}) = 1 - p(C_1|\phi, \mathbf{t})$$

为了预测分布，我们首先注意到$\sigma(w^T\phi)$对于w的依赖只通过它在ϕ上的投影实现。记$a = w^T\phi$,我们有

$$\sigma(w^T\phi)=\int\delta(a-w^T\phi)\sigma(a)\mathrm{d}a$$

其中$\delta(.)$是狄拉克Delta函数，因此：

$$\int\sigma(w^T\phi)q(w)\mathrm{d}w=\int\sigma(a)p(a)\mathrm{d}a$$

其中

$$p(a)=\int\delta(a-w^T\phi)q(w)\mathrm{d}w$$

注意到Delta函数给w施加了一个线性限制，因此在所有与ϕ正交的方向上积分，就可以得到联合概率分布$q(w)$的边缘分布。由于$q(w)$是高斯分布，所以边缘概率分布也是高斯分布。我们可以计算各阶矩然后交换a和w的积分顺序来计算均值和协方差，即：

$$\mu_a=\mathbb{E}[a]=\int p[a]a\mathrm{d}a=\int q(w^T\phi)\mathrm{d}w=w_{MAP}^T\phi$$

同样地，

$$\begin{aligned}\sigma_a^2=\mathbf{var}[a]&=\int p(a)\{a^2-\mathbb{E}[a]^2\}\mathrm{d}a\\&=\int q(w)\{(w^t\phi)^2-(m_N^T\phi)^2\}\mathrm{d}w=\phi^TS_N\phi\end{aligned}$$

a的分布的函数形式与线性回归模型的预测分布相同，其中噪声方差被设为0，因此对于预测分布的近似变成了

$$p(C_1|\mathbf{t})=\int\sigma(a)p(a)\mathrm{d}a=\int\sigma(a)\mathcal{N}(a|\mu_a,\sigma_a^2)\mathrm{d}a$$

关于a的积分表示一个高斯分布和一个 **logistic sigmod** 函数的卷积，不能解析的求值。然而$\sigma(a)=\frac{1}{1+\exp(-a)}$,和逆probit函数$\phi(a)$的高度相似性来获得一个好的近似。为了获得对于logistic函数的最好近似，需要重新为横轴定义一个标度，使得我们可以用$\Phi(\lambda a)$来近似$\sigma(a)$,通过令两个函数在原点处有相同的斜率，找到$\lambda^2=\frac{\pi}{8}$.
使用逆probit函数的一个优势是他与高斯的卷积可以用另一个逆probit函数解析的表达出来，可以知道：

$$\int\Phi(\lambda a)\mathcal{N}(a|\mu,\sigma^2)\mathrm{d}a=\Phi(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}})$$

将$\sigma(a)\simeq\Phi(\lambda a)$应用于两侧，可以得到 logistic sigmod 函数与高斯的卷积近似

$$\int\sigma(a)\mathcal{N}(a|\mu,\sigma^2)\mathrm{d}a\simeq\sigma(\kappa(sigma^2)\mu)$$

其中

$$\kappa(\sigma^2)=(1+\frac{\pi\sigma^2}{8})^{-\frac{1}{2}}$$

用到：

$$p(C_1|\mathbf{t})=\int\sigma(a)p(a)\mathrm{d}a=\int\sigma(a)\mathcal{N}(a|\mu_a,\sigma_a^2)\mathrm{d}a$$

就可以得到：

$$p(C_1|\phi,\mathbf{t})=\sigma(\kappa(\sigma_a^2)\mu_a)$$

其中

$$\mu_a = \mathbb{E}[a] = \int p[a]a\mathrm{d}a = \int q(w^T\phi)\mathrm{d}w = w_{MAP}^T\phi$$

$$\begin{aligned}\sigma_a^2 = \mathbf{var}[a] &= \int p(a)\{a^2 - \mathbb{E}[a]^2\}\mathrm{d}a \\ &= \int q(w)\{(w^t\phi)^2 - (m_N^T\phi)^2\}\mathrm{d}w = \phi^T S_N\phi\end{aligned}$$

注意，对应于$p(C_1|\phi,\mathbf{t}) = 0.5$的决策边界由$\mu_a = 0$给出，这与使用w的MAP值得到的结果相同。因此，如果决策准则是基于最小错误分类率，且先验概率相同，那么对w积分没有效果。然而对于更复杂的决策准则，这个积分就起着很重要的作用了。