第四章

本章目的：

分类的目的是将输入变量 x 对应到K个离散的类别中的一个。一般这些类别都是不相交的，因此一个输入只对应到一个类别，则输入空间被根据对应的类别划分为各个决策区域，决策区域的边界为决策边界或决策面。 本章考虑用线性模型进行分类，即对应的决策面是在D-1维上的超平面。能够被这些线性决策面完全分开的数据称为线性可分。target的表示方法，对于概率模型来说，二分类问题可以用一个二值变量表示，多分类问题可以用长度为K的向量表示 1 of K编码方式；对于非概率模型，有其它表示方法。

第一章定义了3种方式来解决分类问题：

- 1.直接构造判别函数，对于输入的x对应输出类别
- 2.在推理阶段 对$p(C_k|x)$ 建模，基于分布做出最优决策。分离推理和决策
  - 2.1直接建模$p(C_k|x)$为带参数的模型，基于训练集优化参数
  - 2.2使用生成式的方法，先建模类条件密度 $p(x|C_k)$ 和类先验$p(C_k)$ ，然后基于贝叶斯定理求出 $p(C_k|x)$

对于第三章的回归问题，预测模型 $y(x,w)$ 是关于参数 w 的线性函数，最简单的形式是$y(x)=w^Tx+w_0$ ，即它既是关于w的线性函数，也是关于输入变量x的线性函数。

对于分类问题，我们预测的是离散的类标签，或更一般的，是类后验概率。考虑线性函数的推广，通过一个非线性函数对 参数w 的线性函数进行变换，即 $y(x)=f(w^Tx+w_0)$ ，其中$f(\cdot)$ 称作激活函数，它的反函数称作连接函数 link function 。决策面则等价于$y(x)=constant$，因此 $w^Tx+w0=constant$ 。

4.1 判别函数

二分类

最简单的线性判别函数 $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ 若 $y(\mathbf{x}) \geq 0$ 则 $\mathbf{x}$ 归到C_1 类；反之归到C_2 因此决策面为 $y(\mathbf{x}) = 0$

假设在决策面上的两个点为 $\mathbf{x}_A, \mathbf{x}_B$ ，则 $y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0,\ \mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) = 0$ ，即 $\mathbf{w}$ 为决策面的法向量。

- 若在决策面上的一点记为 $\mathbf{x}$， 则原点到决策面的垂直距离

$$\frac{\mathbf{w}^T\mathbf{x}}{||\mathbf{w}||} = \frac{||\mathbf{x}|| \cdot ||\mathbf{w}|| \cdot cos < \mathbf{x}, \mathbf{w} >}{||\mathbf{w}||}$$

$$\frac{\mathbf{w}^T\mathbf{x}}{||\mathbf{w}||} = \frac{-w_0}{||\mathbf{w}||}$$

- 对任意一点 $\mathbf{x}$ ，其投影在决策面上的点为 $\mathbf{x}_\perp$ 记点 $\mathbf{x}$ 到决策面的垂直距离为 r ， 则 $\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{||\mathbf{w}||}$

  两边乘 $\mathbf{w}^T$ 并加上 w_0 ，得到 $y(\mathbf{x}) = 0 + r||\mathbf{w}||,\quad r = \frac{y(\mathbf{x})}{||\mathbf{w}||}$

同样，可以引入附加的下标，令 $x_0 = 1, \widetilde{\mathbf{w}} = (w_0, \mathbf{w}),\ \widetilde{\mathbf{x}} = (x_0, \mathbf{x}),$ 则

$$y(\mathbf{x}) = \widetilde{\mathbf{w}}^T\widetilde{\mathbf{x}}$$

决策平面变为 D+1维输入空间中的D维超平面

多分类

- 直观的方法是引入K/2 个二分类的线性判别函数(**one-versus-rest classifier**)，但存在歧义。
  - 比如对四个类别的问题，若引入2个二分类的线性判别函数，则可分为 C1, C2, not C1 and not C2，只能明显区分三类；
- 另一个方法是引入 K(K-1)/2个二分类的线性判别函数(**one-versus-one classifier**)，每个二分类器对应于类别可能的组合形式(即从K类里面选2类 组合数$C_k^2 = K(K-1)/2$，位于分界线两侧的分别属于对应的类)，每个输入样本点分到其所在区域中vote(得票)最多的类别，但也存在歧义。

- 比如三条分界线分三类，但是在三条分界线相交的内部无法区分到底属于哪一类，因为每一类的得票都相等。

考虑一个K类的判别函数，由K个线性函数组成，

$y_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0}$ ，则 $\mathbf{x}$ 的类别为 $k^* = argmax_k\ y_k(\mathbf{x})$

因此类之间的决策边界为 $y_k(\mathbf{x}) = y_j(\mathbf{x})$ ，$(\mathbf{w}_k - \mathbf{w}_j)^T\mathbf{x} + (w_{k0} - w_{j0}) = 0$

决策区域是连续且convex 凸的? 区域里任意两点的连线仍在区域内部。

线性判别函数的参数学习方法

1. 最小二乘

参数矩阵$\widetilde{\mathbf{W}}$ 第k列对应着第k类的参数$\widetilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k^T)^T$。

给定n个样本，$\widetilde{\mathbf{X}}$ 中每一行是$\widetilde{\mathbf{x}_n}^T$, $\mathbf{T}$中的每一行是对应的K维(1 of K) 向量 $\mathbf{t}_n^T$

均方误差 $E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}Tr\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\}$

对参数求导，$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{T} = \widetilde{\mathbf{X}}^\dagger\mathbf{T}$

为线性判别函数提供了闭式解，然而缺乏鲁棒性

2.Fisher's 线性判别

从维度降低的角度考察。考虑二分类情形，若有一个D维输入向量$\mathbf{x}$，将其投影到一维 $y = \mathbf{w}^T\mathbf{x}$, 若在y 上设置一个阈值，将$y \geq w_0$ 的样本分为C_1，其余分为 C_2 ，则得到之前的标准线性分类器。然而向一维投影会造成许多信息丢失，因此在原始D维空间中能够完美分离开的样本可能在一维空间中互相重叠。通过调整权向量 $\mathbf{w}$，可选择使类别之间分开最大的一个投影。

对于二分类问题:

- 假设 C_1类N_1个点，C_2类N_2个，两类的均值向量 $\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in C_1}\mathbf{x}_n$，$\mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in C_2}\mathbf{x}_n$
- 最简单的度量类别之间分开程度的方式是: 类别均值投影之后的距离，

即最大化$m_2 - m_1 = \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)$

- 来自类别C_k的数据点经过投影到一维空间后，类内方差 $s_k^2 = \sum_{n \in C_k}(y_n - m_k)^2$
- 可以定义整个数据集总的类内方差为 $s_1^2 + s_2^2$

Fisher准则根据类间方差和类内方差的比值定义， 即$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$

$\mathbf{S}_B$是协方差矩阵，表示类间方差，$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$

$\mathbf{S}_W$是整体类内方差，
$\mathbf{S}_W = \sum_{n \in C_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in C_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$

对$\mathbf{w}$求导，令导数为零，得到$J(\mathbf{w})$取最大值时，
$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_W \cdot \mathbf{w} = (\mathbf{w}^T\mathbf{S}_W\mathbf{w})\mathbf{S}_B \cdot \mathbf{w}$

观察$\mathbf{S}_B \cdot \mathbf{w}$，可发现其与$(\mathbf{m}_2 - \mathbf{m}_1)$同向 (因为后面两项相乘为常数)， 则可以丢弃上面求导后的等式左右两边的括号里的常数项，并两边同乘 $\mathbf{S}_W^{-1}$ ，得到 $\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$ ，即**Fisher判别函数**

3.Fisher判别与最小二乘(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接近，而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准则可看成最小平方的一个特例。

一般将目标值表示为K维的向量 "1 of K"的表示方法，这里换用另一种方法，让属于类别C_1的目标值为$\frac{N}{N_1}$, 其中N_1, N 分别为属于类别1的样本个数和总体样本个数，而属于类别C_2的目标值为$-\frac{N}{N_2}$

平方误差函数可写成$E = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n + w_0 - t_n)^2$

对$\mathbf{w}$, w_0 求导，导数为零，则 $\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n + w_0 - t_n) = 0$ 且 $\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n + w_0 - t_n)\mathbf{x}_n = 0$

根据我们对目标值的定义，有$\sum_{n=1}^{N} t_n = N_1\frac{N}{N_1} + N_2\frac{-N}{N_2} = 0$

则 $w_0 = -\mathbf{w}^T\mathbf{m}$，其中$\mathbf{m} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}(N_1\mathbf{m}_1 + N_2\mathbf{m}_2)$

对求导后令导数为零得到的第二个等式，同样使用对t_n的定义代入，

得到 $(\mathbf{S}_W + \frac{N_1 N_2}{N}\mathbf{S}_B)\mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2)$

同样可以推出 $\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$

- 推广Fisher判别到多分类，构建类间方差和类内方差

4. 感知机算法

对应于二分类问题，$y(\mathbf{x}) = f(\mathbf{w}^T\phi(\mathbf{x}))$, 其中非线性激活函数$f(\cdot)$ 是一个阶梯函数，$f(a) = +1, a \geq 0.\ else -1$

另外定义一种误差函数，称作感知机准则 perceptron criterion，为
$E_p(\mathbf{w}) = -\sum_{n\in M}\mathbf{w}^T\phi(\mathbf{x}_n)t_n$

M为所有可能的误分类模式。*当目标与线性函数同号时(相乘大于0)，分类正确。因此为了最小化误差，在定义误差时要加个负号。* 某个特定的误分类模式对于误差函数的贡献是 $\mathbf{w}$空间中，模式被误分类的区域中$\mathbf{w}$的线性函数，总的误差函数是分段线性的。

由此参数的更新可写成 $\mathbf{w}^{(\tau+1)} = w^{(\tau)} - \eta\nabla E_p(\mathbf{w}) = w^{(\tau)} + \eta\,\phi(\mathbf{x}_n)t_n$

感知机的类型可分为 批处理(一次处理一批样本) 或 固定增量单样本(如上式)

缺点： 感知机算法无法提供概率形式的输出，也无法直接推广到多分类的情形。

4.2 概率生成模型

引入sigmoid函数：

考虑二分类情形，$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1)+p(\mathbf{x}|C_2)p(C_2)}$,

记 $a = \ln\frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$ (对数后验比，log odds)， 则 $p(C_1|\mathbf{x}) = \frac{1}{1+exp(-a)} = \sigma(a)$,

$\sigma(a)$为logistic sigmoid函数，定义为$\sigma(a) = \frac{1}{1+exp(-a)}$

- sigmoid函数的性质：

- 对称性 $\sigma(-a)=1-\sigma(a)$
- 反函数为$\sigma^{-1}=a=\ln\frac{\sigma}{1-\sigma}$
- 导数为 $\sigma'=\sigma(a)\cdot(1-\sigma(a))$

进一步，对于多分类问题，$p(C_k|\mathbf{x})=\frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_j p(\mathbf{x}|C_j)p(C_j)}$，定义
$a_k=\ln(p(\mathbf{x}|C_k)p(C_k))$，

则$p(C_k|\mathbf{x})=\frac{exp(a_k)}{\sum_j exp(a_j)}$，称作归一化指数，是logistic sigmoid函数的推广，也被称为softmax函数

softmax理解：可看作是max函数的平滑版本，若 $a_k\gg a_j$，则
$p(C_k|\mathbf{x})\approx 1, p(C_k|\mathbf{x})\approx 0$

输入变量为连续

假设类条件密度为高斯分布，所有的类都用同一个协方差矩阵

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

由上述有，类后验概率为 $P(C_1|\mathbf{x})=\sigma(\mathbf{w}^T\mathbf{x}+w_0)$

定义 $\mathbf{w}=\Sigma^{-1}(\mu_1-\mu_2),\ w_0=-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1+\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2+\ln\frac{p(C_1)}{p(C_2)}$

从而，高斯概率密度的指数项中 $\mathbf{x}$ 的二次项消失了，(因为假设类条件密度的协方差矩阵相同)，则相应的后验分布为 参数为$\mathbf{x}$的线性函数的logistic sigmoid函数

如果不假设各个类别的协方差矩阵相同，那么二次项就不会消去，得到 $\mathbf{x}$的二次函数，即二次判别函数，相应的决策边界也不再是线性的而是二次的。

最大似然求解参数

一旦具体化了类条件概率密度的参数化函数形式，就能够使用最大似然确定参数值。

二分类，每个类别都有一个高斯类条件概率密度，且协方差矩阵相同。

- 假设给定的数据集是$\{\mathbf{x}_n,t_n\}$, $t_n=1$ 对应于类别C_1, $t_n=0$对应于C_2，则

$$p(t_n)=p(t_n=1)^{t_n}\cdot p(t_n=0)^{1-t_n}$$

- 记类先验概率$p(C_1)=\pi$

$$p(t_n=1)=p(\mathbf{x}_n,C_1)=p(C_1)p(\mathbf{x}_n|C_1)=\pi\mathcal{N}(\mathbf{x}_n|\mu_1,\Sigma),$$
$$p(t_n=0)=p(\mathbf{x}_n,C_2)=p(C_1)p(\mathbf{x}_n|C_1)=(1-\pi)\mathcal{N}(\mathbf{x}_n|\mu_2,\Sigma)$$

- 极大似然估计中，似然函数

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

  - 取对数，对 π 求导，得到$\pi=\frac{1}{N}\sum_{n=1}^{N}t_n=\frac{N_1}{N}=\frac{N_1}{N_1+N_2}$
  - 取对数，得到 $\sum_{n=1}^{N}t_n\ln N(\mathbf{x}_n|\mu_1,\Sigma)$，从而求和项里只有高斯分布中位于指数项的部分，对$\mu_1$求导，得到 $\mu_1=\frac{1}{N_1}\sum_{n=1}^{N}t_n\mathbf{x}_n$ ，即属于类别 C_1 的输入向量的均值

    对μ_2求导，同理可得到 $\mu_1=\frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)\mathbf{x}_n$， 即属于类别 C_2 的输入向量的均值
  - 考虑协方差矩阵的最大似然解。。过于复杂不想打公式

离散特征

假设输入向量有D个特征，每个特征取离散值{0,1}。利用朴素贝叶斯的假设，特征看作互相独立

则类条件分布 $p(\mathbf{x}|C_k)=\prod_{i=1}^{D}\mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$，模仿之前对sigmoid函数的应用

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

可看出这是输入变量的线性函数

指数分布

4.3 概率判别模型

通过选择某一种类条件概率密度[比如连续输入时，假设类条件概率密度为高斯分布]，对于二分类问题，类后验概率 $p(C_1|x)$ 可以写成logistics sigmoid函数的形式，其变量是(关于 x 的线性函数) [即$\sigma(a)$, $a = g(x)$ 是关于x的线性函数]；对于多分类可写成softmax函数。而对于类条件概率密度具体的参数求解[如假设为高斯分布，则需求解均值和协方差]，可以使用极大似然估计求解，并计算出类先验概率，最后使用贝叶斯定理求解类后验概率。

另一种方法是，对于类条件概率密度，直接显式地使用更一般的线性函数(generalized linear model)的形式，并使用极大似然估计进行参数求解。

- 参数求解的直接方法，对似然函数 $p(C_k|x)$ 进行极大似然估计 ，这是判别式训练的一种方式，优点是未知参数较少，当类条件概率密度的假设没能很好地近似真实分布时，使用该方法能提升预测时的表现。
- 参数求解的间接方法，分别寻找类条件概率密度和类别先验，然后使用贝叶斯定理，这属于生成式建模

固定基函数

在使用各种分类算法前，首先使用一个基函数向量$\phi(\mathbf{x})$，对输入变量 $\mathbf{x}$ 进行固定的非线性变换。最终的决策边界，存在较多局限性

Logistic regression

虽然名字里有回归，但这是用于分类的模型，而不是用于回归的模型。

二分类问题，类后验概率可以写成
$p(C_1|\phi) = y(\phi) = \sigma(w^T x),\ p(C_2|\phi) = 1 - p(C_1|\phi)$

*注意logistic sigmoid函数的性质: * $[\sigma(a)]' = \sigma(1-\sigma)$

对于给定数据集$\{\phi_n, t_n\}$, $t_n \in \{0, 1\},\ \phi_n = \phi(x_n)$

记 $y_n = p(C_1|\phi_n)$ 似然函数可以写成

$$p(\mathbf{t}|w) = \prod_{n=1}^{N} p(t_n|w) = \prod_{n=1}^{N} p(t_n = 1)^{t_n} \cdot p(t_n = 0)^{1-t_n}$$

$$= \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n} \cdot p(\phi_n)$$

$$\propto \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n}$$

取似然函数的负对数，定义为误差函数

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

对w求导 (代 $y_n = p(C_1|\phi_n) = \sigma(w^T x_n)$)，得到$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n)\phi_n$

1. 迭代重加权最小二乘

对参数进行迭代优化，使用Newton-Raphson方法

$$w^{new} = w^{old} - H^{-1} \nabla E(w)$$

其中，H是一个Hessian矩阵，元素是E(w)关于w的二阶导组成

- 验证这个方法是否有效：将其应用到回归模型，比如对于误差函数为均方误差的情况：
  - $y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x)$ ，
    $E(w) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2$
  - $\nabla E(w) = \sum_{n=1}^{N} (w^T \phi_n - t_n)\phi_n = \phi^T \phi w - \phi^T t$
  - 二阶导 $H = \nabla \nabla E(w) = \phi^T \phi$
  - 代入参数更新的公式，$w^{new} = w^{old} - (\phi^T \phi)^{-1} \nabla E(w) = (\phi^T \phi)^{-1} \phi^T t$

    此即标准的最小二乘解，故该方法给出的解有效

应用到 logistic 回归（即本问题），求出一阶导数和二阶导数，代入

$$w^{new} = (\phi^T R \phi)^{-1} \{\phi^T R \phi \, w^{old} - \phi^T (y - t)\}$$

记后面一部分为 $z = \phi w^{old} - R^{-1}(y-t)$， 整个式子的形式可化为与最小二乘法类似，即iteratie reweighted least squares 迭代重加权最小二乘

2.应用到多分类的logistic regression

与二分类相似，只不过将 sigmoid函数换成softmax函数。

Probit回归

probit模型的重要特性是对离群点的敏感性

4.4拉普拉斯近似

拉普拉斯近似的目标是 对于定义在一组连续变量上的概率密度，寻找其高斯近似

假设一个连续变量的概率密度 $p(z) = \frac{1}{Z}f(z)$, Z 是让概率归一化的因子

拉普拉斯方法中，目标是寻找高斯近似 $q(z)$, 其中心位于$p(z)$ 的众数z_0，[众数即满足[p'(z)=0]

将$f(z)$ 取对数，并在z_0处泰勒展开， $\ln f(z) \approx \ln f(z_0) - \frac{1}{2}A(z-z_0)^2$，
$A = -\frac{d^2}{dz^2}\ln f(z)|_{z=z_0}$

再取指数，$f(z) \approx f(z_0)exp\{-\frac{A}{2}(z-z_0)^2\}$ 从而得到一个指数项，且指数部分为二次型

即高斯近似 $q(z) = (\frac{A}{2\pi})^{1/2}exp\{-\frac{A}{2}(z-z_0)^2\}$

4.5 贝叶斯logistic回归

1.拉普拉斯近似

参数的后验 正比于 先验*似然，为了得到后验的高斯近似，先由最大后验求出对应参数，其就是对应高斯近似中的均值，则协方差就是负对数似然函数的二阶导的逆矩阵 $S_N^{-1} = -\nabla\nabla \ln p(w|t)$

故得到的高斯近似形式为 $q(z)=\mathcal{N}(w|w_{MAP},S_N)$

2. 预测分布

$$p(C_1|\phi,t)=\int p(C_1|\phi,w)p(w|t)dw\approx\int\sigma(w^T\phi)q(w)dw$$