1.线性模型基本概念

试图通过学习来确定一个函数，这个函数可以用来进行预测等功能，需要学习的就是**ω**和b



向量形式为

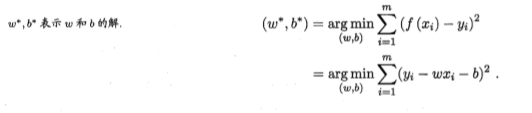


2.线性回归的实现方法

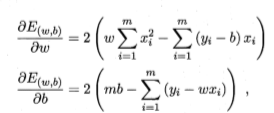
其实过程就是学得参数的过程，先讨论一个单属性的情况

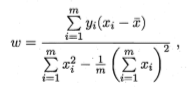


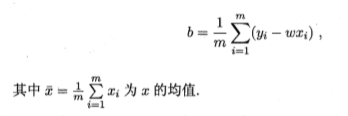
如何求利用的就是最小二乘法的思想，利用均方误差去逼近要求的参数



均方误差的思想就是让样本点与标记点之间的欧氏距离最短，也就是让均方误差值最小，求解一个凸函数（下凹函数）的最小值直接求解对**ω**和b的偏导数令其为0可以求出其值





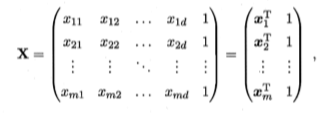


那么对于多属性的情况，我们的问题就转化为了更一般的多元线性回归

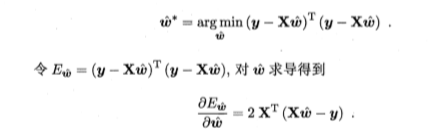
如我们的属性集为试图学得



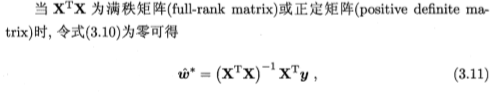
依然可以用最小二乘法但是这里写成一个向量的形式方便讨论把**ω**和b放到一个向量里面记为那我们把每个样本的属性值记为一个m×(d+1)的矩阵、







令上面的式子等于零就可以求出对应参数，但是这里求解过程我们要考虑，矩阵是否可逆的问题，若可逆



那么学得的模型为

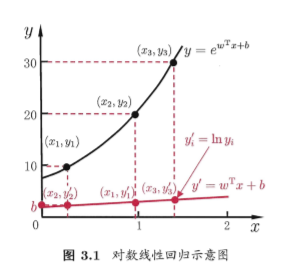


现在的问题就是在现实生活中我们所得到的矩阵不总是可逆的，大部分的情况都是样本很少但是属性值很多，那么我们可以解出多个解，如何选择合适的解就由归纳偏好决定了，常见做法是正则化（正则化关联第六章）

现在有个问题就是我们有部分问题是非线性的，但是我们想用线性的思想去解决，那么我们如何进行解决呢，我们看一个例子

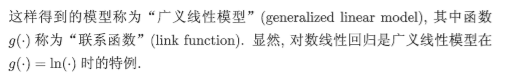
比如我们的标记是在指数尺度上变化，也就是类似一个指数函数，那么我们就无法用线性回归解非线性的方程了，那么如果我们用线性模型对标记的对数进行逼近的话，那么我们又转换为了一个线性的问题





那么对其进行泛化，我们不仅仅针对标记是指数尺度变化，可能是其它函数那么我们令是转换标记为一个线性变化的函数，是可微的那么有



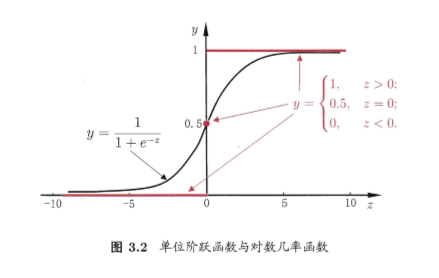


3.对数几率回归

（密度函数）：参数为时出现的概率。这里的是一个参数（parameter），是一个特定的值的（注意，可能已知也可能未知）。

对于上一节讲的任意可微函数将分线性问题转为线性回归问题的方法，若对于我们研究的问题为分类问题，那么对于离散的分类标记，我们可以用一个联系函数将其与联系函数的预测值联系起来

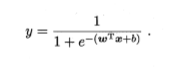
对于二分类问题，标记为0，1但是预测值是实值，那么最好是利用阶跃函数将z转化为0，1标记，但是阶跃函数不可导所以排除，那么有个函数与阶跃函数类似，那就是对率函数（sigmoid）函数如下

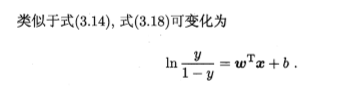


这个函数数值性质不错单调可微函数解析式为



将上函数作为g-带入之前的公式可得





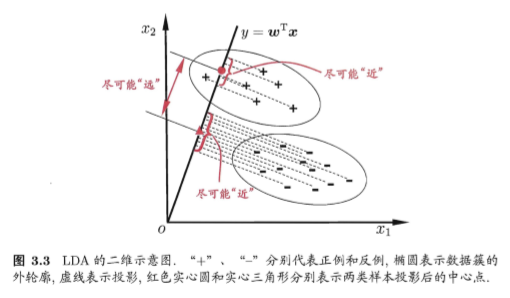
y/（1-y）叫几率所以叫这个为对数几率回归，虽然叫回归实则是分类学习方法

对率回归也是求解凸函数

通过极大似然法求解极大似然最后得出对应参数（对于高阶可导连续函数可用牛顿法，梯度下降法）

4.线性判别分析

LDA的思想非常朴素:给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近、异类样例的投影点尽可能远离;在对新样本进行分类时,将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别。图3.3给出了一个二维示意图。

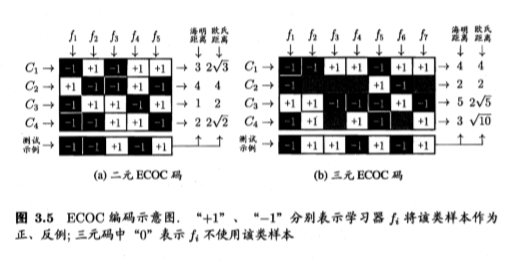


用每个簇的均值向量代表每个样例的均值，然后找到一条直线ω令均值向量其投影为其中心，每个样例的波动程度表现为一个协方差矩阵，我们希望均值投影之间的距离越大越好，协方差越小越好，那么就转为一个优化问题，LDA想要通过最大化瑞利商来找出，对应的直线满足瑞利商最大的目标，通过拉格朗日乘子法或者其它的优化方法对其进行求解。（LDA可以推广到多分类学习中）

5.多分类学习

多分类学习的思想就是基于二分类进行，对多分类问题进行分割，分割为多个二分类的问题，根据分类的方法不同可分为,OvO,OvR,MvM通过投票法则或者置信区间确定OvO和OvR，

对于MvM可见前面两个是它的特例，我们对MvM可用欧氏距离或者海明距离进行分类如下图

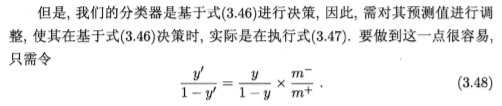


6.类别不平衡问题

在我们实际采样获得的样例中，可能存在正反样例的数目不同的情况，那么对于这种情况学习出来的分类器也许没什么用，那我们必须保证正反样例数目的平衡，如何保持在类别不均的情况下学得分类器仍然ok呢

对于线性分类器实际上我们用的y对阈值的对比来进行分类，那么阈值通常设为0.5这实际上隐含了正反例相等的条件，那么对于不相等的情况，那么由于我们一般假设训练集是总体的无偏采样，用m+代表正例，m-代表反例，m+/m-是观测几率，由于是无偏采样那么观测几率就是真实几率，那么只要只要分类器预测几率高于观测几率就应该是正例，但是我们基于下式来判断





这是不平衡学习的一个基本策略叫再缩放（代价敏感学习的基础）

这个思想很好但是并不是每次采样都是无偏采样，现在一般有3种方法解决不平衡学习，第一第二种其实思想一样那就是强行平衡，一个是过采样另一个是欠采样，使得正反例的数目相近，第三种就是用原始数据但是在决策过程中嵌入前面的式子，叫做阈值移动