机 器 学 习

第3章　线性模型

1、基本形式

线性模型试图学得一个通过属性的线性组合来进行预测的函数，即：

$$\mathrm{f}(x) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$

$$f(x) = w^T x + b \text{（向量形式）} \quad w = (w_1; w_2; \ldots; w_d)$$

线性模型形式简单、易于建模，而且具有很好的可解释性

2、线性回归

“线性回归”试图学得一个线性模型以尽可能准确地预测实值输出标记。即学习公式 $f(x) = w^T x + b$，使得 $f(x_i) \simeq y_i$（多元线性回归）

最小二乘法“参数估计”

$$(w^*, b^*) = \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} (f(x_i) - y_i)^2 = \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

$\Rightarrow$

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b)x_i\right)$$

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m} (y_i - wx_i)\right)$$

- 最小二乘法：基于均方误差最小化来进行模型求解的方法
- 闭式解：也称为解析解，与数值解相对应，是指一些严格的公式，给出任意的自变量就可以求出其因变量，是一种包含公式、三角函数、指数、对数甚至无限级数等基本函数的解的形式。

3、对数几率回归（逻辑斯谛回归）

针对分类问题，考虑二分类任务，其输出标记 $y \in \{0,1\}$，线性回归模型产生的预测值是实值，于是将实值转换为0/1值，最理想的是“单位阶函数”

- 单位阶函数：自变量取值大于0时，判为正例，函数值为1；自变量取值小于0时，判为反例，函数值为0；等于0时可判别为任意值、**因此不连续**
- 对数几率函数 $y = \dfrac{1}{1+e^{-z}}$，将其带入广义线性模型公式中：$\ln\dfrac{y}{1-y} = w^T x + b$

若将y视为样本x作为正例的可能性，则1-y是其反例可能性，两者的比值称为“几率”，反映了x作为正例的相对可能性，对几率取对数则得到“对数几率”。$\ln\dfrac{y}{1-y}$

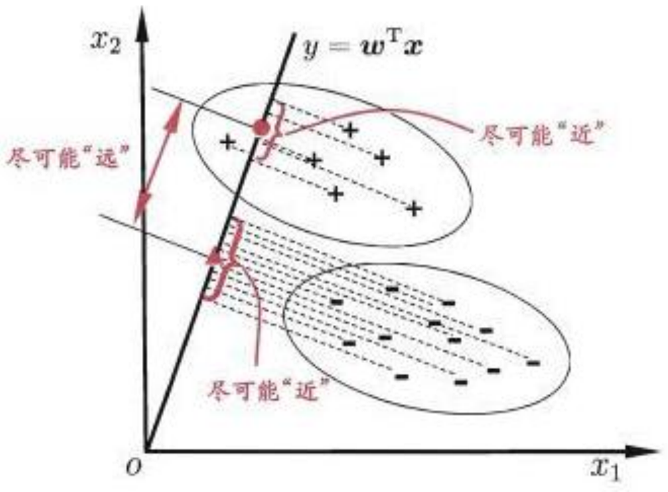
逻辑回归模型的优点有:
1.它是直接对分类可能性进行建模，无需事先假设数据分布，这样避免了假设分布不准确所带来的问题;
2.它不是仅预测出“类别”，而是可得到近似概率预测，这对许多需利用概率辅助决策的任务很有用;
3.对率函数是任意阶可导的凸函数，有很好的数学性质，现有的许多数值优化算法都可直接用于求取最优解。

“极大似然法”参数估计

$$l(w,b)=\sum_{i=1}^{m}\ln p(y_i\mid x_i;w,b) \Longrightarrow \ell(\beta)=\sum_{i=1}^{m}\left(-y_i\beta^T\hat{x}_i+\ln(1+e^{\beta^T\hat{x}_i})\right) \xrightarrow[\text{梯度下降法}]{\text{牛顿法}} \beta^*=\arg\min \ell(\beta)$$

4、线性判别分析（LDA）

LDA的思想：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。



- 均值向量：$\mathrm{w}^T\mu_0$　　$\mathrm{w}^T\mu_1$
- 协方差矩阵：$\mathrm{w}^T\Sigma_0 w$　　$w^T\Sigma_1 w$

$$J = \frac{\| \mathrm{w}^T\mu_0 - \mathrm{w}^T\mu_1 \|_2^2}{\mathrm{w}^T\Sigma_0 w + w^T\Sigma_1 w}$$

- 类内散度矩阵：$S_b = \sum_{x \in X_0} (x-\mu_0)(x-\mu_0)^T + \sum_{x \in X_1} (x-\mu_1)(x-\mu_1)^T$
- 类间散度矩阵：$S_b = (\mu_0-\mu_1)(\mu_0-\mu_1)^T$

- 广义瑞丽商：它的最大值等于矩阵A最大的特征值，而最小值等于矩阵A的最小的特征值

“拉格朗日乘子法”**参数估计**

$$\min_{w} \quad -w^T S_b w$$

$$s.t. \quad w^T S_w w = 1$$

$$S_b w = \lambda S_w w$$

$$S_b w = \lambda(\mu_0 - \mu_1)$$

$$w = S_w^{-1}(\mu_0 - \mu_1)$$

在分类器的理论中，贝叶斯分类器是最优的分类器，而为了得到最优的分类器，我们就需要知道类别的后验概率$P(C_k|x)$。

这里假设$f_k(x)$是类别C_k的类条件概率密度函数，π_k 是类别C_k的先验概率，毫无疑问有$\sum_k \pi_k = 1$。根据贝叶斯理论有：

$$P(C_k|x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

由于π_k 几乎是已知的，所以对于贝叶斯公式而言，最重要的就是这个类条件概率密度函数$f_k(x)$，

LDA假设f(x)是均值不同，方差相同的高斯分布

$$S_B = S_T - S_W = \sum_{x \in X}(x-\bar{x})(x-\bar{x})^T - \sum_{i=1}^{c} S_i = \sum_{x \in X}(x-\bar{x})(x-\bar{x})^T - \sum_{i=1}^{c}\sum_{x \in X_i}(x-\bar{x}_i)(x-\bar{x}_i)^T$$

$$= \sum_{i=1}^{c}\sum_{x \in X_i}(x-\bar{x})(x-\bar{x})^T - \sum_{i=1}^{c}\sum_{x \in X_i}(x-\bar{x}_i)(x-\bar{x}_i)^T \qquad // \sum_{x \in X} = \sum_{i=1}^{c}\sum_{x \in X_i}$$

$$= \sum_{i=1}^{c}\sum_{x \in X_i}[(x-\bar{x})(x-\bar{x})^T - (x-\bar{x}_i)(x-\bar{x}_i)^T]$$

$$= \sum_{i=1}^{c}\sum_{x \in X_i}[xx^T - x\bar{x}^T - \bar{x}x^T + \bar{x}\bar{x}^T - (xx^T - x\bar{x}_i^T - \bar{x}_i x^T + \bar{x}_i\bar{x}_i^T)]$$

$$= \sum_{i=1}^{c}\sum_{x \in X_i}[-x\bar{x}^T - \bar{x}x^T + \bar{x}\bar{x}^T + x\bar{x}_i^T + \bar{x}_i x^T - \bar{x}_i\bar{x}_i^T]$$

$$\sum_{x \in X_i}(-x\bar{x}^T - \bar{x}x^T + \bar{x}\bar{x}^T + x\bar{x}_i^T + \bar{x}_i x^T - \bar{x}_i\bar{x}_i^T)$$

$$= \sum_{x \in X_i} - x\bar{x}^T - \sum_{x \in X_i}\bar{x}x^T + \sum_{x \in X_i}\bar{x}\bar{x}^T + \sum_{x \in X_i}x\bar{x}_i^T + \sum_{x \in X_i}\bar{x}_i x^T - \sum_{x \in X_i}\bar{x}_i\bar{x}_i^T$$

$$= -N_i\bar{x}_i\bar{x}^T - N_i\bar{x}\bar{x}_i^T + N_i\bar{x}\bar{x}^T + N_i\bar{x}_i\bar{x}_i^T + N_i\bar{x}_i\bar{x}_i^T - N_i\bar{x}_i\bar{x}_i^T \qquad // \sum_{x \in X_i} x = N_i\bar{x}_i, \quad \sum_{x \in X_i} x^T = N_i\bar{x}_i^T$$

$$= N_i(-\bar{x}_i\bar{x}^T - \bar{x}\bar{x}_i^T + \bar{x}\bar{x}^T + \bar{x}_i\bar{x}_i^T)$$

$$= N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

5、多分类学习

拆解法：将多分类任务拆为若干个二分类任务求解

- 一对一：将N个类别两两配对，从而产生N(N-1)/2个二分类任务，最终结果可通过投票产生：即把被预测得最多的类别作为最终分类结果。
- 一对其余：每次将一个类的样例作为正例、其他所有类的样例作为反例来训练N个分类器。在测试时如果有一个分类器预测为正，则对应的类别标记作为最终分类结果。
- 多对多：每次选若干个正例，若干个反例。纠错输出码技术ECOC，纠错输出码对分类器的错误有一定的容忍和修正能力。

ECOC（纠错输出码）

工作过程：

- 编码：对N个类别做M次划分，每次划分将一部分划为正类，一部分划为反类，从而形成一个二分类训练集；这样一共产生M个训练集，训练出M个分类器。
- 解码：M个分类器分别对测试样本进行预测，这些预测标记组成一个编码，将这个预测编码与每个类别各自的编码进行比较，返回其中距离最小的类别作为最终预测结果。

类别划分：主要通过“编码矩阵”确定

二元码：将每个类别分别只定位正类和反类

三元码：在正反类之外，还可指定“停用类”

海明距离：两个合法代码对应位上编码不同的位数称为码距。

欧式距离：$P(A,B)=\sqrt{\sum(a_i-b_i)^2}$

6、类别不平衡问题

类别不平衡：是指分类任务中不同类别的训练样例数目差别很大的情况。

基本策略：

- 再缩放：$\frac{y'}{1-y'} = \frac{y}{1-y} * \frac{m^-}{m^+}$
- 欠采样：去除一些反例使得正、反例数目接近，然后再进行学习；
- 过采样：增加一些正例，使得正、反例数目接近，然后再进行学习；
- 阈值移动：直接基于原始训练集进行学习，但在用训练好的分类器进行预测时，将上式嵌入到其决策过程中；

谢 谢 大 家