

少量点无法确定高维模型，需要引入归纳偏好

数据太多用最小二乘

处理离散属性，有序就连续化，没有就转为k维向量

一、回归问题

最小二乘

- 均方误差（欧氏距离）最小化

$$\begin{aligned}(w^*, b^*) &= \arg \min_{(w, b)} \sum_{i=1}^m (f(x_i) - y_i)^2 \\ &= \arg \min_{(w, b)} \sum_{i=1}^m (y_i - wx_i - b)^2.\end{aligned}$$

- 求导得闭式解

分别对 w 和 b 求导：

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

令导数为 0，得到闭式(closed-form)解：

$$w = \frac{\sum_{i=1}^m y_i (x_i - \bar{x})}{\sum_{i=1}^m x_i^2 - \frac{1}{m} \left(\sum_{i=1}^m x_i \right)^2} \quad b = \frac{1}{m} \sum_{i=1}^m (y_i - wx_i)$$

多元表示

把 w 和 b 吸收入向量形式 $\hat{w} = (w; b)$ ，数据集表示为

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T & 1 \\ \mathbf{x}_2^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^T & 1 \end{pmatrix} \quad \mathbf{y} = (y_1; y_2; \dots; y_m)$$

- 矩阵表示

$$\hat{\mathbf{w}}^* = \arg \min_{\hat{\mathbf{w}}} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})$$

$E_{\hat{\mathbf{w}}} = (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})$, 对 $\hat{\mathbf{w}}$ 求导:

$$\frac{\partial E_{\hat{\mathbf{w}}}}{\partial \hat{\mathbf{w}}} = 2\mathbf{X}^T (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) \quad \text{令其为零可得 } \hat{\mathbf{w}}$$

- 求逆讨论

□ 若 $\mathbf{X}^T\mathbf{X}$ 满秩或正定, 则 $\hat{\mathbf{w}}^* = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T\mathbf{y}$

□ 若 $\mathbf{X}^T\mathbf{X}$ 不满秩, 则可解出多个 $\hat{\mathbf{w}}$

此时需求助于归纳偏好, 或引入 **正则化** (regularization) → 第6、11章

- 把高维当作多元: 如通过多项式变换进行增广, 等价于高阶多项式的线性模型

模型推广

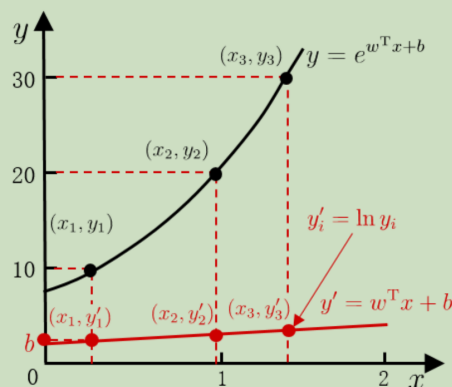
令预测值逼近 y 的衍生物?

若令 $\ln y = \mathbf{w}^T \mathbf{x} + b$

则得到对数线性回归

(log-linear regression)

实际是在用 $e^{\mathbf{w}^T \mathbf{x} + b}$ 逼近 y



- 广义形式

一般形式: $y = g^{-1}(\mathbf{w}^T \mathbf{x} + b)$



单调可微的 **联系函数** (link function)

令 $g(\cdot) = \ln(\cdot)$ 则得到 对数线性回归

$$\ln y = \mathbf{w}^T \mathbf{x} + b$$

正则化项

对比岭回归和原始线性回归的解, 能够发现这两个模型权重 \mathbf{w} 有所不同, 偏移项 b 的形式是一致的, 但会基于对应的 \mathbf{w} 的最优解进行计算. 在岭回归的最优解中, 主要的区别在于公式 (23) 的第一项在矩阵求逆的过程中增加了 $2\lambda\mathbf{I}_d$. 新增的一项能够避免矩阵的特征值趋于 0, 使得 $\mathbf{X}^\top \mathbf{H} \mathbf{X} + 2\lambda\mathbf{I}_d$ 矩阵的特征值至少大于 2λ , 从而方便矩阵的求逆操作. 岭回归方法也能够看做具有高斯先验的线性回归模型, 在第 7 章中将进一步讨论.

实际应用中, 正则化 $\Omega(\cdot)$ 一般用于对模型的复杂度进行约束, 防止过拟合. 除了本例所示的 $\Omega = \|\mathbf{w}\|_2^2$ 外, 也可以设置为其他的范数, 例如 L_1 范数 $\Omega = \|\mathbf{w}\|_1 = \sum_{i=1}^d |w_i|$, 即权重 \mathbf{w} 各元素绝对值之和. 使用 L_1 范数正则化能够使权重 \mathbf{w} 的元素稀疏 (有大量元素为 0).

在线性回归模型中, 正则化一般只用于权重 \mathbf{w} 而不施加于偏移项 b . 观察线性回归模型对于训练样例和测试样例的预测结果, 偏移项 b 刻画了残差的均值, 当对样例进行中心化之后, 目标函数中将不包含 b . 因此, 一般保留 b 的实际语义, 使 b 能够直接刻画一种“截距”, 而不对 b 进行大小的约束. 从这一处理也能够看出, 在求解岭回归模型中, 将偏移项并入权重项合并求解和单独求解得到的结果有所不同.

二、分类问题--对数几率回归

从建模到优化, 是标准过程, 麻雀五脏俱全

- 名字叫回归, 实际是分类, 是分类学习算法

单位阶跃函数:

- 把实质转化为0/1

预测值大于0判为正例, 为0随意

- 缺点: 不可导, 不连续

$$y = \begin{cases} 0, & z < 0; \\ 0.5, & z = 0; \\ 1, & z > 0, \end{cases}$$

对数几率函数 (Sigmoid) :

- 平滑可导

替代函数surrogate function, 简称为对率函数logistic function

以对率函数为联系函数:

$$y = \frac{1}{1 + e^{-z}} \quad \text{变为} \quad y = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

y : 样本 x 正例的概率

$y/1-y$: 几率, 反映 x 是正例的相对可能性

$\ln(y/1-y)$: 对数几率

优点

- 无需假设数据分布, 直接对分类可能性建模
(避免假设不准确的误差)
- 得到近似概率预测
(不仅仅预测类别, 适用于概率辅助决策任务)
- 凸函数, 任意阶可导
好算 (数值优化算法求解最优解)

和一般回归的区别

- 可以预测事件可能性
- 度量模型拟合程度
- 估计回归系数

求解思路

1、把样本正例概率用后验概率描述

若将 y 看作类后验概率估计 $p(y = 1 | \mathbf{x})$, 则

$$\ln \frac{y}{1-y} = \mathbf{w}^T \mathbf{x} + b \quad \text{可写为} \quad \ln \frac{p(y = 1 | \mathbf{x})}{p(y = 0 | \mathbf{x})} = \mathbf{w}^T \mathbf{x} + b$$

2、极大似然法

给定数据集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$

最大化“对数似然”(log-likelihood)函数

$$\ell(\mathbf{w}, b) = \sum_{i=1}^m \ln p(y_i | \mathbf{x}_i; \mathbf{w}, b)$$

3、似然项重写

令 $\beta = (w; b)$, $\hat{x} = (x; 1)$, 则 $w^T x + b$ 可简写为 $\beta^T \hat{x}$

再令 $p_1(\hat{x}_i; \beta) = p(y = 1 | \hat{x}_i; \beta) = \frac{e^{w^T x + b}}{1 + e^{w^T x + b}}$

$$p_0(\hat{x}_i; \beta) = p(y = 0 | \hat{x}_i; \beta) = 1 - p_1(\hat{x}_i; \beta) = \frac{1}{1 + e^{w^T x + b}}$$

则似然项可重写为 $p(y_i | x_i; w, b) = y_i p_1(\hat{x}_i; \beta) + (1 - y_i) p_0(\hat{x}_i; \beta)$

- 似然项整合形式(可重写为指数形式, 把 y_i 和 $1-y_i$ 拿上去, 便于计算)
- 似然函数高阶可导凸

4、等价问题

于是, 最大化似然函数 $\ell(w, b) = \sum_{i=1}^m \ln p(y_i | x_i; w, b)$

$$\text{等价于最小化 } \ell(\beta) = \sum_{i=1}^m \left(-y_i \beta^T \hat{x}_i + \ln \left(1 + e^{\beta^T \hat{x}_i} \right) \right)$$

高阶可导连续凸函数, 可用经典的数值优化方法
如梯度下降法/牛顿法 [Boyd and Vandenberghe, 2004]

梯度下降

- 步长 η 设置: 可用二阶导的倒数

牛顿法

- 需要求二阶导的逆, 高维不好求, 求逆 $O(n^3)$
- 没有高阶法不能用

三、线性判别分析LDA

Linear Discriminant Analysis

别名Fisher判别分析

一大类方法

思想: 将所有样例投影到一条直线上, 使同类的尽量近, 异类的尽量远

- 需要在直线上找到点, 距离同类最短。画阈值, 投影结果大于则正
- 降维可以反映性质

给定数据集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$

第 i 类示例的集合 X_i

第 i 类示例的均值向量 μ_i

第 i 类示例的协方差矩阵 Σ_i

两类样本的中心在直线上的投影: $w^T \mu_0$ 和 $w^T \mu_1$

两类样本的协方差: $w^T \Sigma_0 w$ 和 $w^T \Sigma_1 w$

同类样例的投影点尽可能接近 $\rightarrow w^T \Sigma_0 w + w^T \Sigma_1 w$ 尽可能小

异类样例的投影点尽可能远离 $\rightarrow \|w^T \mu_0 - w^T \mu_1\|_2^2$ 尽可能大

用均值投影表示一堆样本

同类近: 协方差刻画散度, 越小越好

异类远: 均值远

于是, 最大化

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

- 最大化广义瑞利商, 求 w 最优解, 先变形转化:

定义 “类内散度矩阵” (within-class scatter matrix)

$$\begin{aligned} S_w &= \Sigma_0 + \Sigma_1 \\ &= \sum_{x \in X_0} (x - \mu_0)(x - \mu_0)^T + \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^T \end{aligned}$$

以及 “类间散度矩阵” (between-class scatter matrix)

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T,$$

则式(3.32)可重写为

$$J = \frac{w^T S_b w}{w^T S_w w}.$$

可以规定 w 长度, 因为 w 长度无所谓, 只和方向有关

拉格朗日乘子可用: 并不严格凸, 但问题和限制都是二次型

拉格朗日得出广义特征值

令 $\mathbf{w}^T \mathbf{S}_w \mathbf{w} = 1$, 最大化广义瑞利商等价形式为

$$\begin{cases} \min_{\mathbf{w}} & -\mathbf{w}^T \mathbf{S}_b \mathbf{w} \\ \text{s.t.} & \mathbf{w}^T \mathbf{S}_w \mathbf{w} = 1 \end{cases}$$

运用拉格朗日乘子法, 有 $\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$

由 \mathbf{S}_b 定义, 有 $\mathbf{S}_b \mathbf{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \mathbf{w}$

注意到 $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \mathbf{w}$ 是标量, 令其等于 λ

于是 $\mathbf{w} = \mathbf{S}_w^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

实践中通常是进行奇异值分解 $\mathbf{S}_w = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T$

→ 附录 A

然后 $\mathbf{S}_w^{-1} = \mathbf{V} \boldsymbol{\Sigma}^{-1} \mathbf{U}^T$

- tr: 把矩阵信息转化为标量

推广到多类:

- 不能无穷维度, \mathbf{S}_b 的秩有限制
- 多个 \mathbf{w} , 是矩阵, 投影到低维空间中
- 还是用散度评估近和远

可以将 LDA 推广到多分类任务中. 假定存在 N 个类, 且第 i 类示例数为 m_i . 我们先定义 “全局散度矩阵”

$$\begin{aligned} \mathbf{S}_t &= \mathbf{S}_b + \mathbf{S}_w \\ &= \sum_{i=1}^m (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T, \end{aligned} \quad (3.40)$$

其中 $\boldsymbol{\mu}$ 是所有示例的均值向量. 将类内散度矩阵 \mathbf{S}_w 重定义为每个类别的散度矩阵之和, 即

$$\mathbf{S}_w = \sum_{i=1}^N \mathbf{S}_{w_i}, \quad (3.41)$$

其中

$$\mathbf{S}_{w_i} = \sum_{\mathbf{x} \in X_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T. \quad (3.42)$$

由式(3.40)~(3.42)可得

$$\begin{aligned} \mathbf{S}_b &= \mathbf{S}_t - \mathbf{S}_w \\ &= \sum_{i=1}^N m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T. \end{aligned} \quad (3.43)$$

显然, 多分类 LDA 可以有多种实现方法: 使用 \mathbf{S}_b , \mathbf{S}_w , \mathbf{S}_t 三者中的任何两个即可. 常见的一种实现是采用优化目标

63

$$\max_{\mathbf{W}} \frac{\text{tr}(\mathbf{W}^T \mathbf{S}_b \mathbf{W})}{\text{tr}(\mathbf{W}^T \mathbf{S}_w \mathbf{W})}, \quad (3.44)$$

其中 $\mathbf{W} \in \mathbb{R}^{d \times (N-1)}$, $\text{tr}(\cdot)$ 表示矩阵的迹(trace). 式(3.44)可通过如下广义特征值问题求解:

$$\mathbf{S}_b \mathbf{W} = \lambda \mathbf{S}_w \mathbf{W}. \quad (3.45)$$

\mathbf{W} 的闭式解则是 $\mathbf{S}_w^{-1} \mathbf{S}_b$ 的 $N-1$ 个最大广义特征值所对应的特征向量组成的矩阵.

若将 \mathbf{W} 视为一个投影矩阵, 则多分类 LDA 将样本投影到 $N-1$ 维空间, $N-1$ 通常远小于数据原有的属性数. 于是, 可通过这个投影来减小样本点的维数, 且投影过程中使用了类别信息, 因此 LDA 也常被视为一种经典的监督降维技术.

当假设各类样例的协方差矩阵相同时, FDA 退化为线性判别分析 LDA。

四、多分类学习

使用策略, 让2分类解决多分类

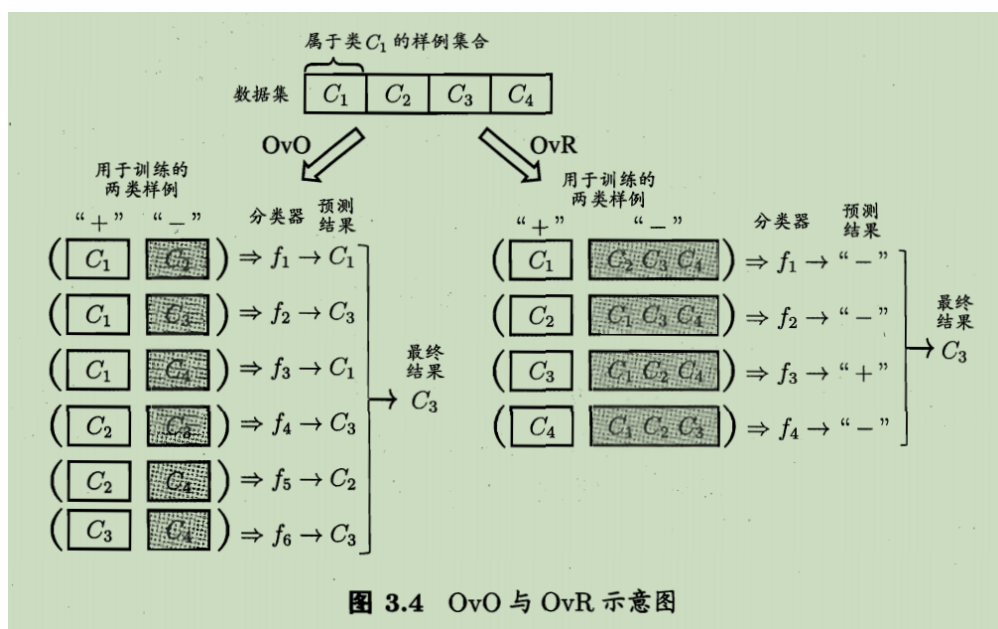
- 拆解法：拆成多个二分类

OvO—对一

- 两两配对产生 $N(N-1)/2$ 个二分类任务
- 得到 $N(N-1)/2$ 个分类结果，最终结果可通过投票产生

OvR—对其余

- 每次将一个类的样例作为正例、所有其他类的样例作为反例来训练 N 个分类器
- N 次中选结果为正里面置信度最大的



MvM多对多

OvO和OvR是 MvM的特例

- 每次将若干个类作为正类，若干个其他类作为反类

一种常见方法：纠错输出码 (Error Correcting Output Code)

编码：对 N 个类别做 M 次划分，每次将一部分类别划为正类，一部分划为反类



M 个二类任务；
(原)每类对应一个长为 M 的编码

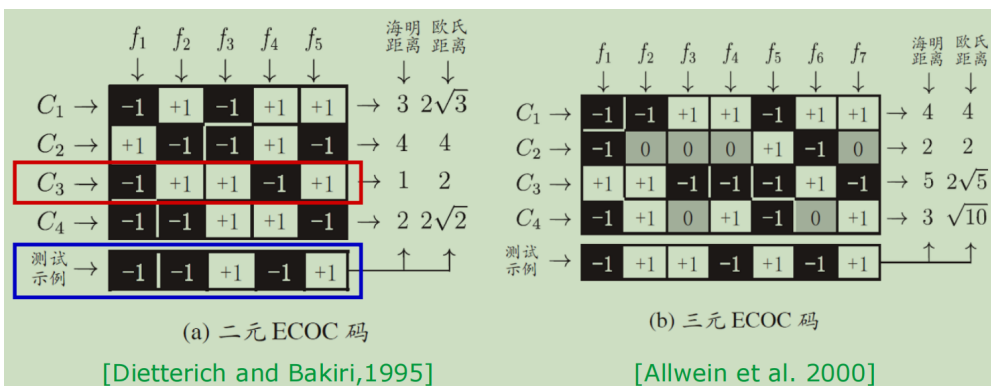
距离最小的类为
最终结果



解码：测试样本交给 M 个分类器预测



长为 M 的预测结果编码



- ECOC编码对分类器错误有一定容忍和修正能力，编码越长、纠错能力越强
- 对同等长度的编码，理论上来说，任意两个类别之间的编码距离越远，则纠错能力越强

优缺点

- OvO的存储开销和测试时间开销通常比 OvR 更大
- 类别很多时，OvO的训练时间开销通常比OvR 更小

OvR的每个分类器均使用全部训练样例、而 QvO 的每个分类器仅用到两个类的样例

- 预测性能则取决于具体的数据分布，在多数情形下两者差不多

难以处理的情况

- OvO 中, 可能存在多个分类器的投票相等
- OVR 中, 可能存在所有分类器均判断为负类, 或多个分类器预测为正类