

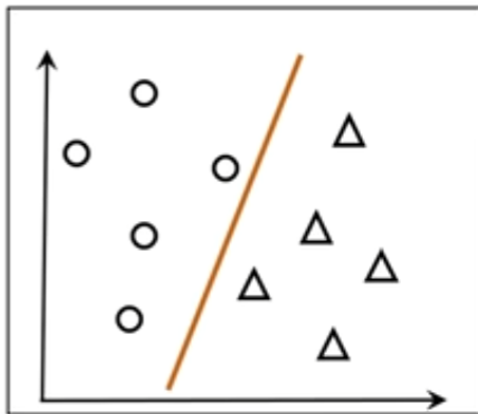
Lecture 3: 线性模型

线性回归

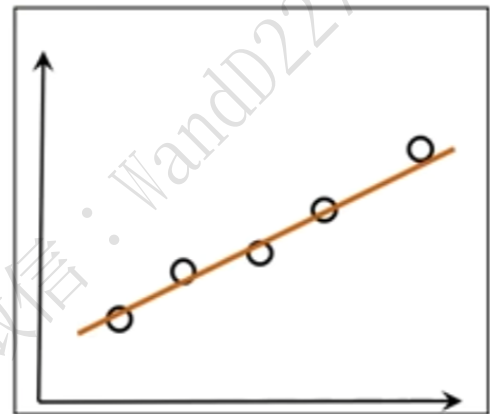
线性模型以及求解

概念：为什么需要线性模型

- 概念：
 - 人类考虑非线性问题时很困难，但是对于线性问题的考虑更加直观；
- 举例：
 - 线性模型 (linear model) 试图学得一个通过属性的线性组合来进行预测的函数



分类



回归

概念：线性回归

- 概念：
 - 找到一个方程 $f(x_i) = wx_i + b$ 使得 $f(x_i) \sim y_i$;
 - 即：能否使用一条直线，大概描述出一个事物的大致分布，再用得到的这个分布去进一步的估算未知事物的可能结果；
 - 即：线性方程描述了 x 和 y 之间的关系，即解和分布，然后利用分布、可以评估未知数据的结果；
 - 擅长处理数值属性；
- 第一步：将离散数据属性化
 - 概念：
 - 把离散的数据连续化；
 - 要注意这个事物当中是否有一个序的关系，不能一看到一个离散的数据、就将其划分为 0~1 的处理；
 - 有序：
 - 如果有序，那就连续化为 0~1 处理；
 - 如果你用离散的东西要去处理这种连续信号，那先要把离散数据连续化；
 - 无序：
 - 把属性变为一个向量，使用 0、1 构成的向量编码；
 - 假设一个事物有 3 个属性，那就变成三维的向量；

- 总结:

- 离散属性的处理: 若有“序”(order), 则连续化否则, 转化为 k 维向量;

- 第二步: 寻找线性回归

- 概念:

- 对于 $f(x_i) = w^T x + b$, 令均方误差最小化, 得到 w 和 b, 这个 w 和 b 就是要求解的参数;
- 如果整个 w 向量和 b 向量解出来了, 那这个方程就确定了, 也就是得到线性方程了;
- 可以用平方差来表示数据到直线的距离, 求它们的最小值:

$$\begin{aligned}(w^*, b^*) &= \arg \min_{(w, b)} \sum_{i=1}^m (f(x_i) - y_i)^2 \\ &= \arg \min_{(w, b)} \sum_{i=1}^m (y_i - wx_i - b)^2 \\ \text{对 } E_{(w, b)} &= \sum_{i=1}^m (y_i - wx_i - b)^2 \text{ 进行最小二乘参数估计}\end{aligned}$$

- 概念: 最小二乘估计

- 分析: 求出 w 和 b 的闭式解

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:

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

- 为什么求偏导? 因为令导数等于 0、是希望找到一个极值点, 这个极值点是确定, 要么到达了最小值、最大值;
- 对于这个线性回归的表达式而言, 没有最大值, 它可以无限大;

- 所以对于线性回归方程中的 w 和 b 求偏导时，找到的极值点一定就是零点、也就是相等的点；

多元线性回归

概念：多元线性回归

- 概念：
 - 多元就是多变量；
 - 所以这里的 x 就是一个向量；
- 图示：

$$f(x_i) = w^T x_i + b \text{ 使得 } f(x_i) \simeq y_i$$

$$x_i = (x_{i1}; x_{i2}; \dots; x_{id}) \quad y_i \in \mathbb{R}$$

概念：多元线性方程的求解 - 高斯消元法

- 这里面矩阵是简化表达；

$$f(x_i) = w^T x_i + b \text{ 使得 } f(x_i) \simeq y_i$$

$$x_i = (x_{i1}; x_{i2}; \dots; x_{id}) \quad y_i \in \mathbb{R}$$

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

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{pmatrix} \quad y = (y_1; y_2; \dots; y_m)$$

- 求解：如果要解出 w ，要对其求偏导（这里只需要求一次导，因为 w 和 b 都已经吸收到了 W 里面去）；
- 然后对其求偏导，另偏导等于 0，解出 w ：

同样采用最小二乘法求解，有

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

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

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

然而，麻烦来了：涉及矩阵求逆！

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

□ 若 $X^T X$ 不满秩，则可解出多个 $\hat{w}$

概念：正则化

- 概念：
 - 由于有时候方程解出来可能有很多解，因此需要进行额外的限制，也就是正则化；
- 使用：
 - 对于多元线性方程组，有时候如果是无穷解，此时可以进行正则化、进行限制，加入限制之后，得到我们想要的解；

广义线性模型

概念：广义线性模型

- 引入：可以用线性回归解决非线性问题
 - 对于样例希望线性模型的预测值逼近真实标记则得到线性回归模型 $f(x_i) = w^T x + b$ ；
 - 如果要求解的不是一个线性的 y 而是一个 y 的衍生物，比如 y 的对数：

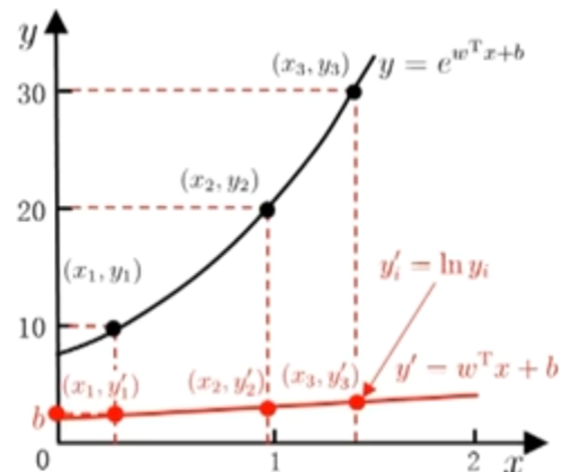
令预测值逼近 y 的衍生物？

若令 $\ln y = w^T x + b$

则得到对数线性回归

(log-linear regression)

实际是在用 $e^{w^T x + b}$ 逼近 y



- 也就是用线性回归，来接近对数的目标；
- 概念：广义线性模型的一般形式

- 图示:

一般形式: $y = g^{-1}(w^T x + b)$



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

$$\ln y = w^T x + b$$

...

- g 函数的作用:
 - 将线性回归得到的结果、和你真正想要的结果联系起来;

对数几率回归

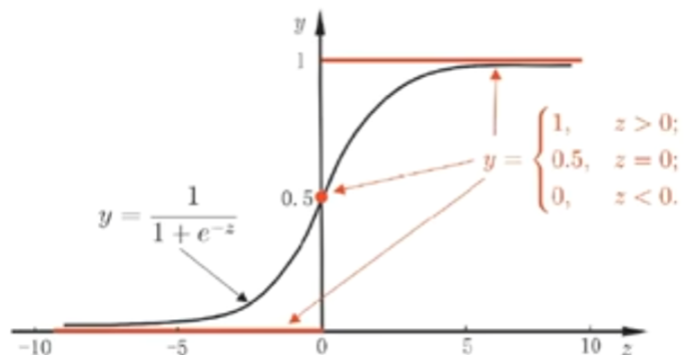
概念: 二分类任务与对率回归

- 概念: 二分类任务
 - 线性回归产生的是实值输出, 而期望的输出是离散的 0、1;
 - 所以就需要一个函数 g , 使得实值输出变成 0、1;
 - 最简单的就是直接使用“单位阶跃函数”, 但直接使用理想的“单位阶跃函数”其性能很糟糕;
 - 因为其中间有间断点, 而且也不可微;
 - 所以要找到一个替代函数, 这个替代函数就是对率函数;

线性回归模型产生的实值输出 $z = w^T x + b$
期望输出 $y \in \{0, 1\}$ } 找 z 和 y 的联系函数

理想的“单位阶跃函数”
(unit-step function)

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



性质不好,
需找“替代函数”
(surrogate function)

常用
单调可微、任意阶可导

$$y = \frac{1}{1 + e^{-z}}$$

对数几率函数
(logistic function)
简称“对率函数”

- 概念: 对率函数与逻辑回归

- 原因：
 - 用一个性质更好的、单调可微分、任意阶可导的函数，来替代单理想的“单位阶跃函数”；
 - 即：根据事物的线性分布，转化为概率进行判断，所以如果一个事物不是简单的线性分布，则其效果也不会好；
 - 逻辑回归常用在分类问题上，因为它得到的结果是一个概率，可以用来判断一个事物的好和坏；

- 概念：

以对率函数为联系函数：

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

即：

$$\ln \frac{y}{1-y} = w^T x + b$$

“对数几率”

(log odds, 亦称 logit)

几率(odds), 反映了 x 作为正例的相对可能性

“对数几率回归” (logistic regression)
简称 “对率回归”

- 实际上是在用线性回归模型的预测结果去逼近真实标记的对数几率，因此其对应的模型称为“对数几率回归”；
- 特别需注意到，虽然它的名字是“回归”，但实际却是一种**分类学习方法**；
- 这种方法有很多优点，例如它是直接对分类可能性进行建模，无需事先假设数据分布，这样就避免了假设分布不准确所带来的问题：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；
- 此外，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；
- 特点：
 - 无需事先假设数据分布；
 - 可得到“类别”的近似概率预测；
 - 可直接应用现有数值优化算法求取最优解

概念：对数回归求解

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

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

于是，可使用“**极大似然法**” $\longrightarrow$ 第7章
(maximum likelihood method)

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

最大化“对数似然”(log-likelihood)函数

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

• 求解：

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

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

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

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

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

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

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]

线性判别分析

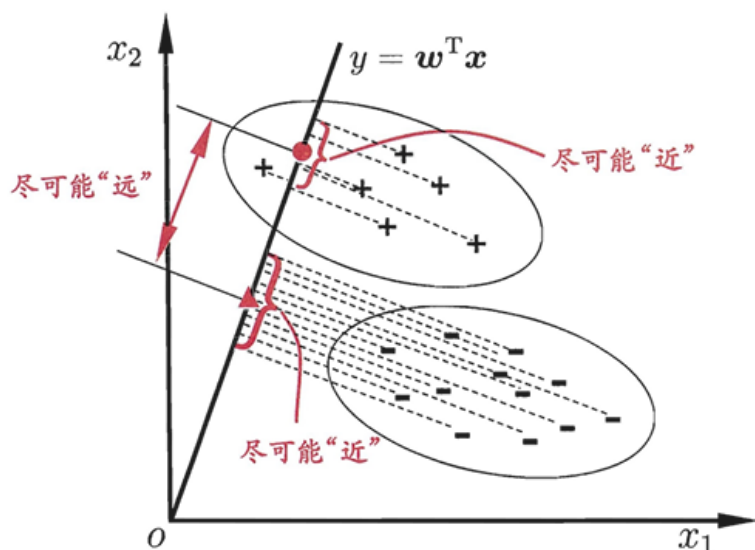
概念：线性判别分析 (Linear Discriminant Analysis)

• 概念：线性判别分析

- 线性判别分析 (Linear Discriminant Analysis，简称 LDA) 是一种经典的线性学习方法，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；

- 图示：



- 概念：LDA

给定数据集 $\{(x_i, y_i)\}_{i=1}^m$

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

于是，最大化

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

类内散度矩阵 (within-class scatter matrix)

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

类间散度矩阵 (between-class scatter matrix)

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

LDA的目标：最大化广义瑞利商 (generalized Rayleigh quotient)

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

令 $w^T S_w w = 1$ 最大化广义瑞利商等价形式为

$$\min_w -w^T S_b w$$

$$\text{s.t. } w^T S_w w = 1$$

运用拉格朗日乘子法，有 $S_b w = \lambda S_w w$

$$\text{由 } S_b \text{ 定义，有 } S_b w = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w$$

注意到 $(\mu_0 - \mu_1)^T w$ 标量，令其等于 λ

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

实践中通常是进行奇异值分解 $S_w = U \Sigma V^T$

$$\text{然后 } S_w^{-1} = V \Sigma^{-1} U^T$$

→ 附录 A

多分类学习

概念：LDA 的多分类推广

假定有 N 个类

□ 全局散度矩阵 $S_t = S_b + S_w = \sum_{i=1}^m (x_i - \mu)(x_i - \mu)^T$

□ 类内散度矩阵 $S_w = \sum_{i=1}^N S_{w_i} \quad S_{w_i} = \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T$

□ 类间散度矩阵 $S_b = S_t - S_w = \sum_{i=1}^N m_i (\mu_i - \mu)(\mu_i - \mu)^T$

多分类LDA有多种实现方法：采用 S_b , S_w , S_t 中的任何两个

例如, $\max_W \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)}$



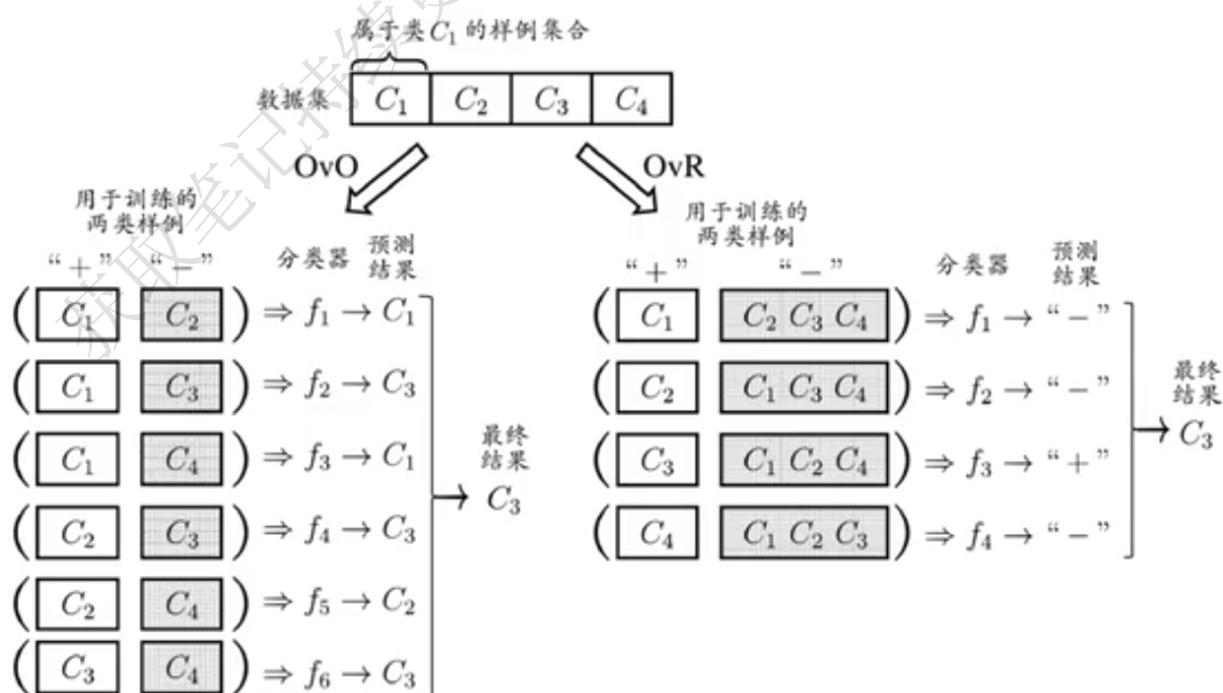
$S_b W = \lambda S_w W$

$W \in \mathbb{R}^{d \times (N-1)}$

W 的闭式解是 $S_w^{-1} S_b$ 的 d' ($\leq N-1$) 个最大非零广义特征值对应的特征向量组成的矩阵

概念：多分类学习

- 概念：
 - 如何建一个多分类问题转化为二分类问题；
- 做法：
 - 拆解法：
- 将一个多分类任务拆分为若干个二分类任务求解；



概念：类别不平衡

- 概念：

- 只有在丢掉的小类很重要时、才需要处理；
- 概念：
 - 处理思路

不同类别的样本比例相差很大；“小类”往往更重要

基本思路：



基本策略

—— “再缩放” (rescaling):

$$\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$$

然而，精确估计 m^-/m^+ 通常很困难！

常见类别不平衡学习方法：

- 过采样 (oversampling)
例如：SMOTE
- 欠采样 (undersampling)
例如：EasyEnsemble
- 阈值移动 (threshold-moving)

获取笔记持续更新可加微信