

# 1 数学基础

<https://www.cnblogs.com/pinard/archive/2004/01/13/10791506.html>

## 1.1 矩阵

### 1.1.1 迹相关性质

- 转置不变性:  $tr(A^T) = tr(A)$
- 迹的循环相等性:  $tr(ABC) = tr(BCA) = tr(CAB)$
- 加减分配:  $tr(X \pm Y) = tr(X) \pm tr(Y)$
- 乘法和迹交换:  $tr((A \odot B)^T C) = tr(A^T (B \odot C))$

## 1.2 矩阵导数

### 1.2.1 导数布局

行头对列头求导	标量 $y$	$m$ 维向量 $\mathbf{y}$	$p \times q$ 维矩阵 $\mathbf{Y}$
标量 $x$	$1 \times 1$	$m \times 1$	$\frac{\partial \mathbf{Y}}{\partial x}, p \times q$
$n$ 维向量 $\mathbf{x}$	$\frac{\partial y}{\partial \mathbf{x}}, n \times 1$	$n \times m$	
$r \times s$ 维矩阵 $\mathbf{X}$	$\frac{\partial y}{\partial \mathbf{X}}, r \times s$		

### 1.2.2 标量对张量求导

标量值函数的矩阵微分定义为:

$$df = \sum_{i,j} \frac{\partial f}{\partial X_{ij}} dX_{ij} = tr((\frac{\partial f}{\partial \mathbf{X}})^T d\mathbf{X}) = tr((\frac{\partial f}{\partial \mathbf{X}}) d\mathbf{X}^T)$$

并且具有如下常用性质:

1.  $d(X \pm Y) = dX \pm dY$ ,  $d(XY) = (dX)Y + X(dY)$ ,  $d(X \odot Y) = (dX) \odot Y + X \odot (dY)$
2.  $d(X^T) = (dX)^T$ ,  $d(tr(X)) = tr(dX)$
3.  $dX^{-1} = -X^{-1}(dX)X^{-1}$
4.  $d|X| = |X|tr(X^{-1}dX)$
5. 逐元素求导:  $d\sigma(X) = \sigma'(X) \odot dX$

例题1. 求  $\frac{\partial tr(ABA^T)}{\partial A}$

$$d(tr(ABA^T)) = tr((dA)BA^T + AB(dA^T)) = tr(AB^T(dA)^T) + tr(AB(dA)^T) = tr(A(B^T + B)(dA)^T)$$

故有  $\frac{\partial tr(ABA^T)}{\partial A} = A(B^T + B)$

例题2.  $y = \mathbf{a}^T \exp(X\mathbf{b})$ , 求  $\frac{\partial y}{\partial X}$

$$dy = tr(dy) = tr(\mathbf{a}^T d\exp(X\mathbf{b})) = tr(\mathbf{a}^T \exp(X\mathbf{b}) \odot d(X\mathbf{b})) = tr((\mathbf{a} \odot \exp(X\mathbf{b}))^T d(X\mathbf{b})) = tr(b(\mathbf{a} \odot \exp(X\mathbf{b}))^T dX)$$

故  $\frac{\partial y}{\partial X} = (\mathbf{a} \odot \exp(X\mathbf{b}))\mathbf{b}^T$

### 1.2.3 链式法则

- 比如  $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ , 本质上就是对中间变量的每个分量对  $\mathbf{x}$  求导乘上  $\mathbf{y}$  对中间变量的每个分量求导, 然后求和.

### 1.2.4 对向量求导链式法则

- 向量对向量的链式法则:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$$

- 标量对向量的链式法则:

$$\frac{\partial z}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial z}{\partial \mathbf{y}}, \quad ((\text{len}(\mathbf{x}), \text{len}(\mathbf{y})) \cdot (\text{len}(\mathbf{y}), 1))$$
$$\frac{\partial z}{\partial \mathbf{x}} = \left( \frac{\partial \mathbf{y}_n}{\partial \mathbf{y}_{n-1}} \frac{\partial \mathbf{y}_{n-1}}{\partial \mathbf{y}_{n-2}} \cdots \frac{\partial \mathbf{y}_2}{\partial \mathbf{y}_1} \right) \frac{\partial z}{\partial \mathbf{y}_n}, \quad ((\text{len}(\mathbf{y}_n), \text{len}(\mathbf{y}_1)) \cdot (\text{len}(\mathbf{y}), 1))$$

例题1. 最小二乘法损失函数的求导, 即求  $\frac{\partial l}{\partial \boldsymbol{\theta}}$ , 这里  $l = (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$

令  $\mathbf{z} = \mathbf{X}\boldsymbol{\theta} - \mathbf{y}$

则  $\frac{\partial l}{\partial \boldsymbol{\theta}} = \left( \frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}} \right) \frac{\partial l}{\partial \mathbf{z}} = \mathbf{X}^T \frac{\partial \mathbf{z}^T \mathbf{z}}{\partial \mathbf{z}}$

而  $\text{tr}(d(\mathbf{z}^T \mathbf{z})) = \text{tr}(\mathbf{z}^T(d\mathbf{z}) + (d\mathbf{z}^T)\mathbf{z}) = \text{tr}(\mathbf{z}^T(d\mathbf{z}) + (\mathbf{z}d\mathbf{z}^T)) = \text{tr}(2\mathbf{z})$

所以  $\frac{\partial l}{\partial \boldsymbol{\theta}} = \mathbf{X}^T 2\mathbf{z} = 2\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$

### 1.2.5 特殊求导

- 梯度:  $\nabla f(\mathbf{x})_i = \frac{\partial f(\mathbf{x})}{\partial x_i}$
- 海森矩阵:  $\nabla^2 f(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$
- Frobenius范数求导:  $\frac{\partial \|\mathbf{A}\|_F^2}{\partial \mathbf{A}} = \frac{\partial \mathbf{A}^T \mathbf{A}}{\partial \mathbf{A}} = 2\mathbf{A}$

## 1.3 矩阵分解

### 1.3.1 特征值分解

- 特征方程:  $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$
- 特征值分解:  $\mathbf{A} = \mathbf{V}\text{diag}(\boldsymbol{\lambda})\mathbf{V}^{-1}$ ,  $\mathbf{V}$ 的每一列为特征向量 $\mathbf{v}_i$
- 特征值为正(非负)的矩阵为正定(半正定)矩阵
- 可特征值分解(可对角化)的充分条件: 实对称矩阵
- 手算分解方法: 求出特征向量就行了.

### 1.3.2 奇异值分解

- 设矩阵  $\mathbf{A} \in \mathbb{C}^{m \times n}$ ,  $\mathbf{A}^T \mathbf{A} \mathbf{v} = \lambda \mathbf{v}$ ,  
令  $\mathbf{A} \mathbf{v} = \sigma \mathbf{u}$ , 则
  - 两边同乘  $\mathbf{A} \mathbf{A}^T$  得到,  $\mathbf{A} \mathbf{A}^T \mathbf{u} = \frac{\lambda}{\sigma} \mathbf{A} \mathbf{v} = \lambda \mathbf{u}$ ,  $\mathbf{u}$  对应  $\mathbf{A} \mathbf{A}^T$  的特征值为  $\lambda$  的特征向量
  - 两边同左乘  $\mathbf{u}^T$  得到,  $\mathbf{u}^T \mathbf{A} \mathbf{v} = \sigma$
  - 两边同左乘  $\mathbf{v}^T \mathbf{A}^T$  得到,  $\frac{\lambda}{\sigma} = \mathbf{v}^T \mathbf{A}^T \mathbf{u} = \mathbf{u}^T \mathbf{A} \mathbf{v}$

因此  $\lambda = \sigma^2$

写成矩阵形式, 有  $AV = U\Sigma$  即  $A = U\Sigma V^T$ , 其中  $U \in C^{m \times m}$ , 为  $AA^T$  的特征向量构成矩阵;  $\Sigma \in C^{m \times n}$ ;  $V \in C^{n \times n}$ , 为  $A^T A$  的特征向量构成矩阵

- 任意矩阵都可以作奇异值分解
- 截断奇异值分解:  $A \approx U_k \Sigma_k V_k^T$ ,  $U_k \in C^{m \times k}$  即取前  $k$  列;  $\Sigma_k \in C^{k \times k}$ ;  $V_k \in C^{n \times k}$ , 即取前  $k$  行

## 1.4 矩阵的秩

- Sylvester 不等式.  $A \in \mathbf{R}^{m \times n}$ ,  $B \in \mathbf{R}^{n \times s}$ , 那么有

$$\text{rank}(A) + \text{rank}(B) - n \leq \text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$$

•

$$\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$$

## 1.5 概率分布

- 多元正态分布

$$- \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

$$- \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\beta}(\mathbf{x} - \boldsymbol{\mu})\right)$$

- 贝塔分布

– 概率密度函数:

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

$$- E[\mu] = \frac{a}{a+b}$$

$$- \text{Var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

- 狄利克雷分布(是贝塔分布的多元扩展)

– 多个连续变量  $\mu_i \in [0, 1]$  的分布, 满足  $\sum \mu_i = 1$

$$- \text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \alpha_i} \prod_i \mu_i^{\alpha_i - 1}$$

$$- E[\mu_i] = \frac{\alpha_i}{\sum_i \alpha_i}$$

- 伽马分布( $\tau > 0$ )

$$- \text{Gam}(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}$$

$$- E[\tau] = \frac{a}{b}$$

$$- \text{Var}[\tau] = \frac{a}{b^2}$$

## 1.6 K-L散度

- 衡量两个分布的差异

– K-L散度定义:

$$D_{KL}(P\|Q) = E_{X \sim P} \left[ \log \frac{P(x)}{Q(x)} \right]$$

即

$$D_{KL}(P\|Q) = E_{X \sim P} \log P(x) - E_{X \sim P} Q(x)$$

$E_{X \sim P} \log P(x)$  为P的熵,  $E_{X \sim P} Q(x)$  为P和Q的交叉熵

- 非负, P=Q时为0
- $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$ , 但理论上最小值都是P=Q

## 1.7 优化

- 标准优化问题:

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } \begin{cases} g_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m \\ h_j(\mathbf{x}) = 0, j = 1, 2, \dots, n \end{cases}$$

- 凸函数

- 定义(零阶条件): 对于  $f: \mathbb{R}^n \mapsto \mathbb{R}$ , 若  $\forall t \in [0, 1], f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$ , 则  $f$  为凸函数
- 一阶条件:  $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$
- 二阶条件:  $\nabla^2 f(\mathbf{x}) \geq 0$ , 即海森矩阵半正定

- 凸优化问题:

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } \begin{cases} g_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m \\ \mathbf{a}_i^T \mathbf{x} = b, j = 1, 2, \dots, n \end{cases}$$

- 无约束优化: 梯度下降:  $x_{t+1} \leftarrow x_t - \alpha \nabla f(x_t)$

- 有约束优化: 拉格朗日乘子法

- 拉格朗日函数:  $L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_i u_i g_i(\mathbf{x}) + \sum_j v_j h_j(\mathbf{x})$ , 其中  $u_i \geq 0$
- $\forall \mathbf{u} \geq 0, \mathbf{v}$  和可行解  $\mathbf{x}$ , 有

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \leq f(\mathbf{x})$$

- 对偶问题:

$$\max_{\mathbf{u}, \mathbf{v}} g(\mathbf{u}, \mathbf{v}), \text{ s.t. } \mathbf{u} \geq 0$$

其中  $g(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \mathbf{v})$

- 弱对偶性: 设可行解集  $C$ , 原问题最优解  $f^*$ , 则

$$f^* \geq \min_{\mathbf{x} \in C} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \geq \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = g(\mathbf{u}, \mathbf{v})$$

从而得到

$$f^* \geq g^* = \max_{\mathbf{u}, \mathbf{v}} g(\mathbf{u}, \mathbf{v})$$

## 2 模型评估与选择

### 2.1 误差与过/欠拟合

- 误差: 预测与实际之差异
- 训练误差/经验误差: 在训练集上的误差
- 泛化误差: 在新样本上的误差
- 过拟合/欠拟合

### 2.2 评估方法

#### 2.2.1 数据集划分

- 留出法:
  - 注意分层采样以保留比例
  - 一般采用若干次随即划分取平均值作为留出法的评估结果.
  - 一般将2/3 ~ 4/5的样本用于训练
- k折交叉验证法
  - 最常用k=10
  - 当k=m(即数据集D中样本数)时, 称为留一法. 准确(未必)但开销大.



- 自助法: 采样到  $D'$  后放回, 则约有  $\lim_{m \rightarrow \infty} \left(1 - \frac{1}{m}\right)^m = \frac{1}{e} \approx 0.368$  的样本未出现在采样的数据集  $D'$  中, 可将其 ( $D/D'$ ) 作为测试集
  - 数据集小难以划分时有用, 但改变了初始数据集的分布会引入估计偏差

#### 2.2.2 调参

- 网格法/随机法
- 验证集: 从训练集中选出用来验证超参数

### 2.3 性能度量

- 均方误差(MSE)/根均方误差(RMSE)
- 绝对误差(MAE)

### 2.3.1 分类任务

- 基本定义:

- 错误率: 分错样本数/样本总数
- 精度(accuracy): 分对样本数/样本总数

– 混淆矩阵:

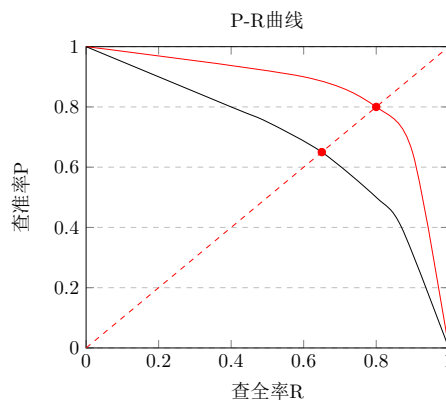
真实情况	预测结果	
	正例	反例
正例	$TP$ (真正例)	$FN$ (假反例)
反例	$FP$ (假正例)	$TN$ (真反例)

- 查准率  $P = \frac{TP}{TP+FP}$
- 查全率  $R = \frac{TP}{TP+FN}$
- 真正例率  $TPR = \frac{TP}{TP+FN}$
- 假正例率  $FPR = \frac{FP}{FP+TN}$

- 度量方法

- P-R图法: 查全率为横轴, 查准率为纵轴, 作出曲线.

- \* 用P-R曲线下面积大小度量
- \* 用平衡点度量(平衡点在直线P=R上)



- F1度量:

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

- $F_\beta$ 度量:

$$F_\beta = \frac{1 + \beta^2}{\frac{1}{P} + \frac{\beta^2}{R}}$$

$\beta > 1$ 时, 查全率影响更大;  $\beta < 1$ 时, 查准率影响更大

- 宏F1: 由于多次重复训练测试, 有多个混淆矩阵. 可用:

$$\begin{aligned} macro_P &= \frac{1}{n} \sum_{i=1}^n P_i \\ macro_R &= \frac{1}{n} \sum_{i=1}^n R_i \\ macro_{F1} &= \frac{2 \times macro_P \times macro_R}{macro_P + macro_R} \end{aligned}$$

- 微F1:

$$\begin{aligned} micro_P &= \frac{\bar{TP}}{TP+FP} \\ micro_R &= \frac{\bar{TP}}{TP+FN} \\ micro_{F1} &= \frac{2 \times micro_P \times micro_R}{micro_P + micro_R} \end{aligned}$$

– **ROC与AUC**: 衡量排序(分类任务常先预测出一个数值)质量好坏

- \* ROC曲线: 给定不同阈值, 会有一个假正例率FPR和一个真正例率TPR, 以这些点画图. 得到以假正例率为横轴, 真正例率为纵轴的图
- \* AUC: Area Under ROC Curve. 有公式:

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1})$$

- \* 排序损失:

$$\begin{aligned} l_{rank} &= \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left( \mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right) \\ &= 1 - AUC \end{aligned}$$

– 代价敏感错误率与代价曲线

- \* 代价敏感错误率(cost-sensitive):

$$E(f; D; cost) = \frac{1}{m} \left( \sum_{x_i \in D^+} \mathbb{I}(f(x_i) \neq y_i) \times cost_{01} + \sum_{x_i \in D^-} \mathbb{I}(f(x_i) \neq y_i) \times cost_{10} \right)$$

- \* 代价曲线: 正例概率代价为横轴, 归一化代价为纵轴. ROC上每一点对应代价曲线上一条线段. 可以在非均等代价下衡量期望总体代价(ROC曲线不行)

· 正例概率代价

$$P(+)\text{cost} = \frac{p \times cost_{01}}{p \times cost_{01} + (1 - p) \times cost_{10}}$$

· 归一化代价

$$cost_{norm} = \frac{FNR \times p \times cost_{-1} + FPR \times (1 - p) \times cost_{10}}{p \times cost_{01} + (1 - p) \times cost_{10}}$$

## 2.4 比较检验

### 2.4.1 假设检验

- 二项检验: 若测试错误率 $\hat{\epsilon}$ 小于临界值 $\bar{\epsilon}$ , 则可认为在 $\alpha$ 显著度下, 假设" $\epsilon \leq \epsilon_0$ 不能被拒绝.

$$\bar{\epsilon} = \min \epsilon \text{ s.t. } \sum_{i=\epsilon \times m+1}^m \binom{m}{i} \epsilon_0^i (1 - \epsilon_0)^{m-i} < \alpha$$

- t检验: 多次重复留出法/交叉验证法, 得到多个测试错误率 $\hat{\epsilon}_1 \cdots \hat{\epsilon}_k$ . 则( $\mu$ 和 $s^2$ 是测试错误率的样本均值和样本方差)

$$\tau_t = \frac{\sqrt{k}(\mu - \epsilon_0)}{s} \sim t_{k-1}$$

- 成对交叉验证t检验: 两两同折作差得到 $\Delta_1 \cdots \Delta_k$ , 其样本均值和样本方差为 $\mu, \sigma^2$ , 则

$$\tau_t = \left| \frac{\sqrt{k}\mu}{\sigma} \right| \sim t_{k-1}$$

- 问题: 采样可能不独立, 可以用" $5 \times 2$ 交叉验证"(书41)

## 2.4.2 McNemar检验

算法B	算法A	
	正确	错误
正确	$e_{00}$	$e_{01}$
错误	$e_{10}$	$e_{11}$

假设性能相同应该有 $e_{01} = e_{10}$ ,  $|e_{01} - e_{10}|$ 服从只有大于0部分的类正态分布, 从而

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \sim \chi_1^2$$

## 2.4.3 Friedman检验与Nemenyi后续检验

- 定义序值,对每个算法在不同数据集序值求均值, 则有

$$\tau_{\chi^2} = \frac{k-1}{k} \frac{12N}{k^2-1} \sum_{i=1}^k (r_i - \frac{k+1}{2})^2 \sim \chi_{k-1}^2$$

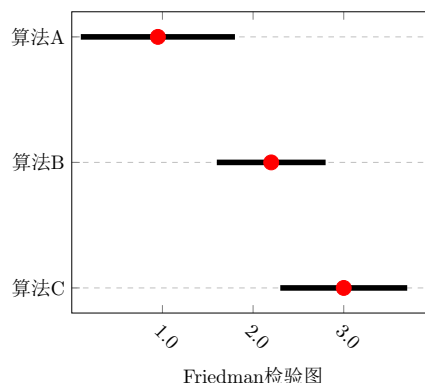
- 另有

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1) - \tau_{\chi^2}} \sim F_{k-1, (k-1)(N-1)}$$

- 后续检验(拒绝了算法性能相同后的后续比较)

— 临界值域:  $CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}}$

- 有交叠就无明显差异, 否则有显著优性.(如图中算法A显著优于算法C)



## 2.5 偏差与方差

- 学习算法的期望预测:

$$\bar{f}(x) = E_D[f(x; D)]$$

- 方差:

$$var(x) = E_D [(f(x; D) - \bar{f}(x))^2]$$

- 噪声(数据标签就标错了):

$$\epsilon^2 = E_D [(y_D - y)^2]$$

- 偏差(期望输出与真实标记的差别):

$$bias^2(x) = (\bar{f}(x) - y)^2$$

- 偏差-方差分解:

$$E(f; D) = bias^2(x) + var(x) + \epsilon^2$$

- 通常来说偏差与方差不可兼得. 例如用二次函数和三次函数, 二次函数偏差大(拟合能力差)而方差小(阶数小), 三次函数则相反.



## 3 线性模型

### 3.1 线性回归

- 对 $d$ 个属性,  $m$ 个样本的情况, 尝试学习 $f(\mathbf{x}_i) = \boldsymbol{\omega}^T \mathbf{x}_i + b$ 使得 $f(\mathbf{x}_i) \approx y_i$
- 则有

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

只需要对 $\hat{\boldsymbol{\omega}}$ 求导得到 $2\mathbf{X}^T(\mathbf{X}\hat{\boldsymbol{\omega}} - \mathbf{y})$ , 从而

$$\hat{\boldsymbol{\omega}}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

写成熟悉的形式: $\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\omega}}^* = \mathbf{X}^T \mathbf{y}$ ,

- $\mathbf{X}^T \mathbf{X}$ 是半正定矩阵( $\mathbf{x}^T \mathbf{X}^T \mathbf{X} \mathbf{x} = (\mathbf{X} \mathbf{x})^T (\mathbf{X} \mathbf{x}) = \mathbf{y}^T \mathbf{y} \geq 0$ ).
- 如果 $\mathbf{X}^T \mathbf{X}$ 满秩, 那么可以直接这样做;
- 如果不满秩( $\mathbf{X}$ 列数多余行数, 可以解出多个 $\hat{\boldsymbol{\omega}}$ ), 需要引入正则化项

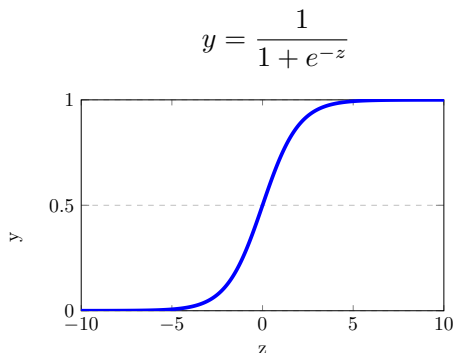
- 考虑可微函数 $g(\cdot)$ , 则

$$y = g^{-1}(\boldsymbol{\omega}^T \mathbf{x} + b)$$

称为广义线性模型(generalized linear model),  $g(\cdot)$ 称为联系函数(link function)

### 3.2 对数几率回归(二分类任务)

- 理想情况是用阶跃函数, 但是它不可微, 故用对数几率函数(logistic function)(这样的函数称为替代函数(surrogate func))



它是一种Sigmoid函数(S形函数)

- 以此则有

$$y = \frac{1}{1 + e^{-(\boldsymbol{\omega}^T \mathbf{x} + b)}}$$
$$\ln \frac{y}{1-y} = \boldsymbol{\omega}^T \mathbf{x} + b$$

其中 $\frac{y}{1-y}$ 称为几率(odds), 反映了正例的相对可能性

- 求极大似然:

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

- 这里令

$$p(y=1|x; \omega, b) = \frac{1}{1 + e^{-\omega^T x + b}} = \frac{e^{\omega^T x + b}}{1 + e^{\omega^T x + b}}$$

$$p(y=0|x; \omega, b) = \frac{e^{-\omega^T x + b}}{1 + e^{-\omega^T x + b}} = \frac{1}{1 + e^{\omega^T x + b}}$$

- 令  $\beta = (\omega; b)$ ,  $\hat{x} = (x; 1)$ , 则  $\omega^T x + b$  可简写为  $\beta^T \hat{x}$ . 进而有

$$l(\beta) = \sum_{i=1}^m (-y_i \beta^T \hat{x}_i + \ln(1 + e^{\beta^T \hat{x}_i}))$$

- 这里抽象出一个函数

$$f(x) = -ax + \ln(1 + e^x)$$

其一阶, 二阶导数为:

$$f'(x) = -a + \frac{1}{1 + e^x} = -a + \sigma(x)$$

$$f''(x) = \sigma(x)(1 - \sigma(x)) > 0 \quad (\sigma(x) \in (0, 1))$$

因此  $l(\hat{\omega})$  是关于  $\hat{\omega}$  的凸函数

- 对负对数似然求导数.

$$\nabla l(\hat{\omega}) = \sum_i f'(z_i) \frac{\partial z_i}{\partial \hat{\omega}} = \sum_i \left( -y_i + \frac{1}{1 + e^{-z_i}} \right) x_i = - \sum_i (y_i - P(y=1|\hat{x}_i; \hat{\omega})) \hat{x}_i$$

$$\nabla^2 l(\hat{\omega}) = \frac{\partial \nabla l(\hat{\omega})}{\partial \hat{\omega}^T} = \sum_i p_i(1 - p_i) \hat{x}_i \hat{x}_i^T$$

- 计算方法:

- 梯度下降法:

$$\hat{\omega}_{t+1} \leftarrow \hat{\omega}_t - \alpha \nabla l(\hat{\omega})$$

- 牛顿法:

$$\hat{\omega}_{t+1} \leftarrow \hat{\omega}_t - (\nabla^2 l(\hat{\omega}))^{-1} \nabla l(\hat{\omega})$$

### 3.3 线性判别分析(LDA, 分类任务)

#### 3.3.1 二分类任务

- 基本思想: 投影, 同类点尽可能近, 不同类点尽可能远. 是一种监督降维技术.
- 一些量:

- 第  $c$  类示例集合:  $X_c$
- 第  $c$  类示例集合的均值向量:  $\mu_c = \frac{1}{|X_c|} \sum_{x \in X_c} x$
- 第  $c$  类示例的协方差矩阵:  $\Sigma_c = \sum_{x \in X_c} (x - \mu_c)(x - \mu_c)^T$
- 两样本中心在直线上的投影:  $\frac{1}{|X_c|} \sum_{x \in X_c} \omega^T x = \omega^T \mu_c$
- 两类样本的协方差:  $\sum_{x \in X_c} (\omega^T x - \omega^T \mu_c)^2 = \omega^T \Sigma_c \omega$

- 最大化目标:

$$\begin{aligned}
 J &= \frac{\|\omega^T \mu_0 - \omega^T \mu_1\|_2^2}{\omega^T \Sigma_0 \omega + \omega^T \Sigma_1 \omega} \\
 &= \frac{\omega^T (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T \omega}{\omega^T (\Sigma_0 + \Sigma_1) \omega} \\
 &= \frac{\omega^T S_b \omega}{\omega^T S_\omega \omega}
 \end{aligned}$$

其中

- 类间散度:  $S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$
- 类内散度:  $S_\omega = \Sigma_0 + \Sigma_1$
- 广义瑞利商:  $\max \frac{\omega^T S_b \omega}{\omega^T S_\omega \omega}$

- 从而得到优化问题:

$$\max_{\omega} \omega^T S_b \omega \quad s.t. \quad \omega^T S_\omega \omega = 1$$

由拉格朗日乘子法解之得  $S_b \omega = \lambda S_\omega \omega$ , 即

$$\omega = \frac{1}{\lambda} S_\omega^{-1} (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T \omega \propto S_\omega^{-1} (\mu_0 - \mu_1)$$

### 3.3.2 LDA推广: 多分类任务

- 全局散度矩阵:  $S_t = S_b + S_\omega = \sum_i (x_i - \mu)(x_i - \mu)^T$
- 类内散度矩阵:  $S_\omega = \sum_c S_{\omega_c}$ , 其中  $S_{\omega_c} = \sum_{x \in X_c} (x - \mu_c)(x - \mu_c)^T$
- 类间散度矩阵:

$$\begin{aligned}
 S_b &= S_t - S_\omega \\
 &= \sum_i (x_i - \mu)(x_i - \mu)^T - \sum_c \sum_{x \in X_c} (x - \mu_c)(x - \mu_c)^T \\
 &= \sum_i x_i x_i^T - m \mu \mu^T - \sum_c \left( \sum_{x \in X_c} x x^T - m_c \mu_c \mu_c^T \right) \\
 &= \sum_c (m_c \mu_c \mu_c^T - m \mu \mu^T) \quad (\text{注意到} \sum_i \text{和} \sum_c \sum_{x \in X_c} \text{都是对所有样本点的求和}) \\
 &= \sum_c m_c (\mu_c - \mu)(\mu_c - \mu)^T
 \end{aligned}$$

- 由此得到优化目标:

$$\max_W \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_\omega W)} = \frac{\sum_i w_i^T S_b w_i}{\sum_i w_i^T S_\omega w_i}$$

- 写成

$$\max_W \text{tr}(W^T S_b W) \quad s.t. \quad \text{tr}(W^T S_\omega W) = 1$$

可得  $S_b W = \lambda S_\omega W$ ,  $W$  的闭式解是  $S_\omega^{-1} S_b$  的  $N - 1$  个最大广义特征值所对应的特征向量组成的矩阵.

- 这里  $S_b$  的秩为  $N - 1$

–  $\text{rank}(S_b) \leq N - 1$ 的证明:

$$\begin{aligned}\text{rank}(S_b) &= \text{rank}\left(\sum_c m_c(\mu_c - \mu)(\mu_c - \mu)^T\right) \\ &\leq \text{rank}\left(\sum_c m_c(\mu_c - \mu)\right) \quad (\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}) \\ &\leq N - 1 \quad \left(\sum_c m_c(\mu_c - \mu) = 0 \text{ 及 } \text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)\right)\end{aligned}$$

### 3.4 多分类学习——用二分类学习器解决多分类问题

- 拆分策略

- 一对一(One vs. One, OvO):  $N$ 个类别两两配对,  $N(N - 1)/2$ 个二分类任务, 投票法得到结果. 分类器多, 但每次训练量小.
- 一对其余(One vs. Rest, OvR): 一类为正例, 其余类为反例,  $N$ 个分类任务. 比较置信度得到类别. 分类器少, 但每次训练量大.
- 多对多(Many vs. Many, MvM): 对 $N$ 个类做 $M$ 次划分(每次分出正类反类或不使用), 成为二分类训练集. 然后将海明距离或欧氏距离最小的作为预测类. 具有纠错能力, 即使某个分类器有错也可能产生正确结果. 任意两个类别之间码距越小, 纠错能力越强.

- \* 海明码的性质:

- 最小码距为 $d$ 的海明编码能够纠错 $\lfloor \frac{d-1}{2} \rfloor$ 个错误. 这是因为即使错了这么多, 也能够通过选取最接近的正确编码来给出正确预测.

### 3.5 类别不平衡问题

- 欠采样: 去除一些反例使正反数目相近
- 过采样: 增加一些正例使正反数目相近
- 阈值移动: 将分类决策条件 $\frac{y}{1-y} > 1$ 预测为正例改为: $\frac{y}{1-y} > \frac{m^+}{m^-}$ 预测为正例.