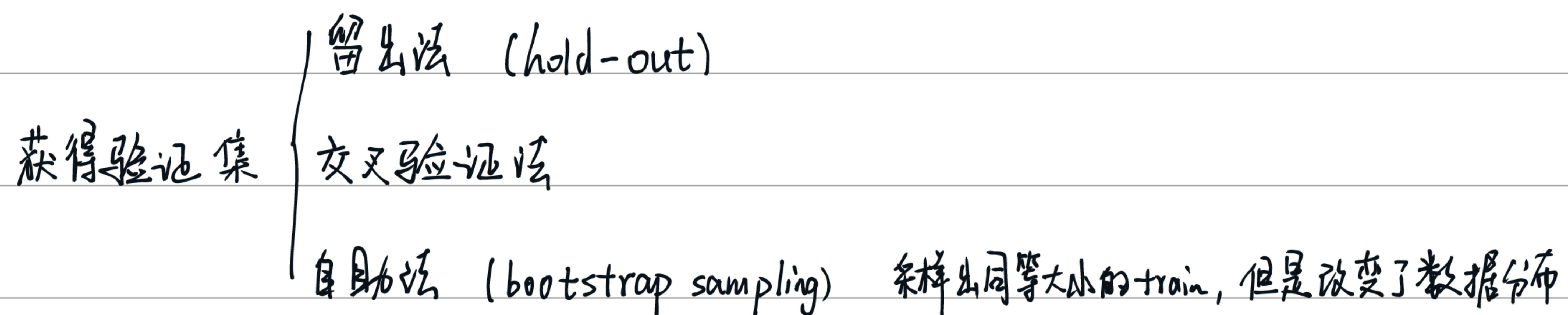


模型评估与选择

泛化能力

经验误差 \leftrightarrow 泛化误差

过拟合 \leftrightarrow 欠拟合



验证集: 本来已知的所有训练集中划分为训练集和验证集, 用于评估和选择模型 (调参)

测试集: 模型实际使用时遇到的数据, 估计模型实际使用的泛化性能

超参数 — 参数

均方误差 (回归常用)

$$\text{数据离散 } E(f; D) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2$$

$$\text{数据连续 } E(f; D) = \int_{x \sim D} (f(x) - y)^2 p(x) dx$$

分类任务

$$\text{错误率 } E(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) \neq y_i) \quad \int_{x \sim D} \mathbb{I}(f(x) \neq y) p(x) dx$$

$$\text{精度 } acc(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) = y_i) = 1 - E(f; D) \quad \int_{x \sim D} \mathbb{I}(f(x) = y) p(x) dx$$

$$\text{查准率 precision} = \frac{TP}{TP + FP}$$

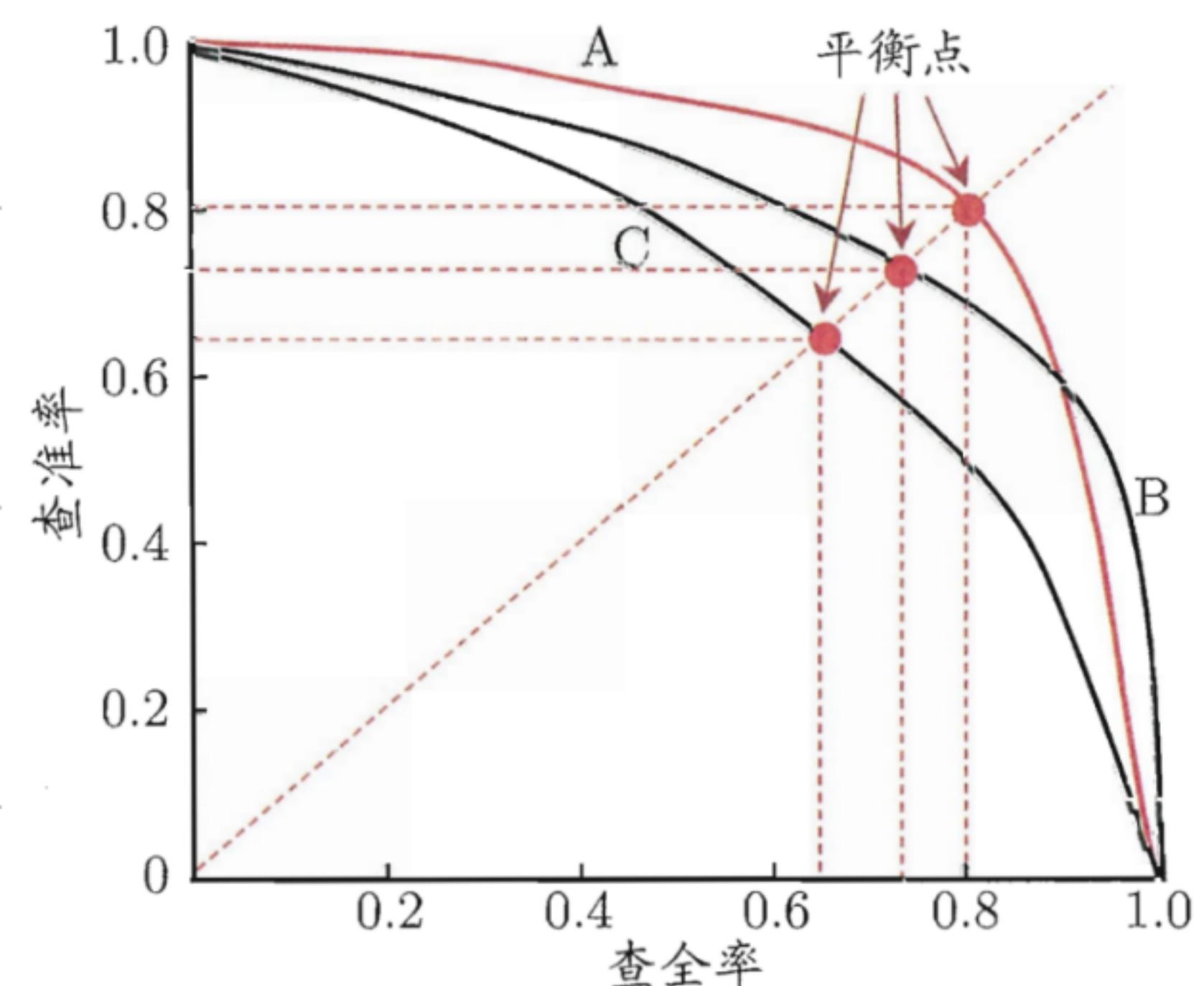
$$\text{查全率 recall} = \frac{TP}{TP + FN}$$

$$f_1 = \frac{2PR}{P+R} \quad \text{调和平均 } f_1 = \frac{1}{\frac{1}{P} + \frac{1}{R}} \quad \text{更重视较小值}$$

$$f_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R} \quad \frac{1}{f_\beta} = \frac{1}{1+\beta^2} \left(\frac{1}{P} + \frac{\beta^2}{R} \right)$$

PR图

遍历模型的所有可能阈值



BER 平衡点 : $P=R$

多个混淆矩阵

macro

micro

ROC 曲线

$$x = \text{FP rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$y = \text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

AUC : Area Under the ROC Curve 即 ROC 面积，越大越好

样本预测的排序质量

⊕⊕田田→

田田⊕⊕→

四分之三

四分之一

考虑排序损失 $\text{loss} = \frac{1}{m^+ m^-} \sum_{x \in D^+} \sum_{x \in D^-} (\mathbb{I}(f(x^+) < f(x)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x)))$

$$\text{AUC} = 1 - \text{loss}$$

非均衡代价

		actual	predict	
		0	1	
actual	0	0	cost ₀₀	
	1	cost ₁₀	0	

$$\text{代价敏感错误率: } E(f; D; \text{cost}) = \frac{1}{m} \left(\sum_{x \in D^+} \mathbb{I}(f(x) \neq y) \text{cost}_{01} + \sum_{x \in D^-} \mathbb{I}(f(x) \neq y) \text{cost}_{10} \right)$$

比较检验:

why?

交叉验证 + 检验 5×2

McNemar

回归任务，均方误差的偏差-方差分解

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

$$\text{bias}^2(x) = (\bar{f}(x) - y)^2 \quad \text{期望输出与真实输出, 模型能力}$$

$$\text{var}(x) = E_D[(f(x; D) - \bar{f}(x))^2] \quad \text{数据扰动}$$

$$\epsilon^2 = E_D[(y_0 - y)^2] \quad \text{噪声、下界}$$

线性模型

$$\hat{y} = w x + b$$

$$(w^*, b^*) = \underset{(w, b)}{\arg \max} \sum_{i=1}^m (wx_i + b - y_i)^2$$
$$w^2 x^2 + (b-y)^2 + 2wx(b-y)$$

$$\begin{cases} \frac{\partial MSE(w, b)}{\partial w} = 2(w \sum x^2 - \sum x(b-y)) = 0 \\ \frac{\partial MSE}{\partial b} \Rightarrow \text{求得 } w, b \end{cases}$$

闭式解

多维线性模型

$$y = w^T x_i + b$$

数据向量: $X = \begin{bmatrix} x_1^T & | & \\ x_2^T & | & \\ \vdots & | & \\ x_D^T & | & \end{bmatrix} \quad \hat{w} = \begin{bmatrix} w \\ b \end{bmatrix}$

$$\hat{y} = X \hat{w}$$

$$MSE = (y - X \hat{w})^T (y - X \hat{w}) \quad 2范数的平方$$

$$\frac{\partial \text{MSE}}{\partial \hat{w}} = 2X^T(X\hat{w} - y)$$

对数线性模型

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

广义线性模型

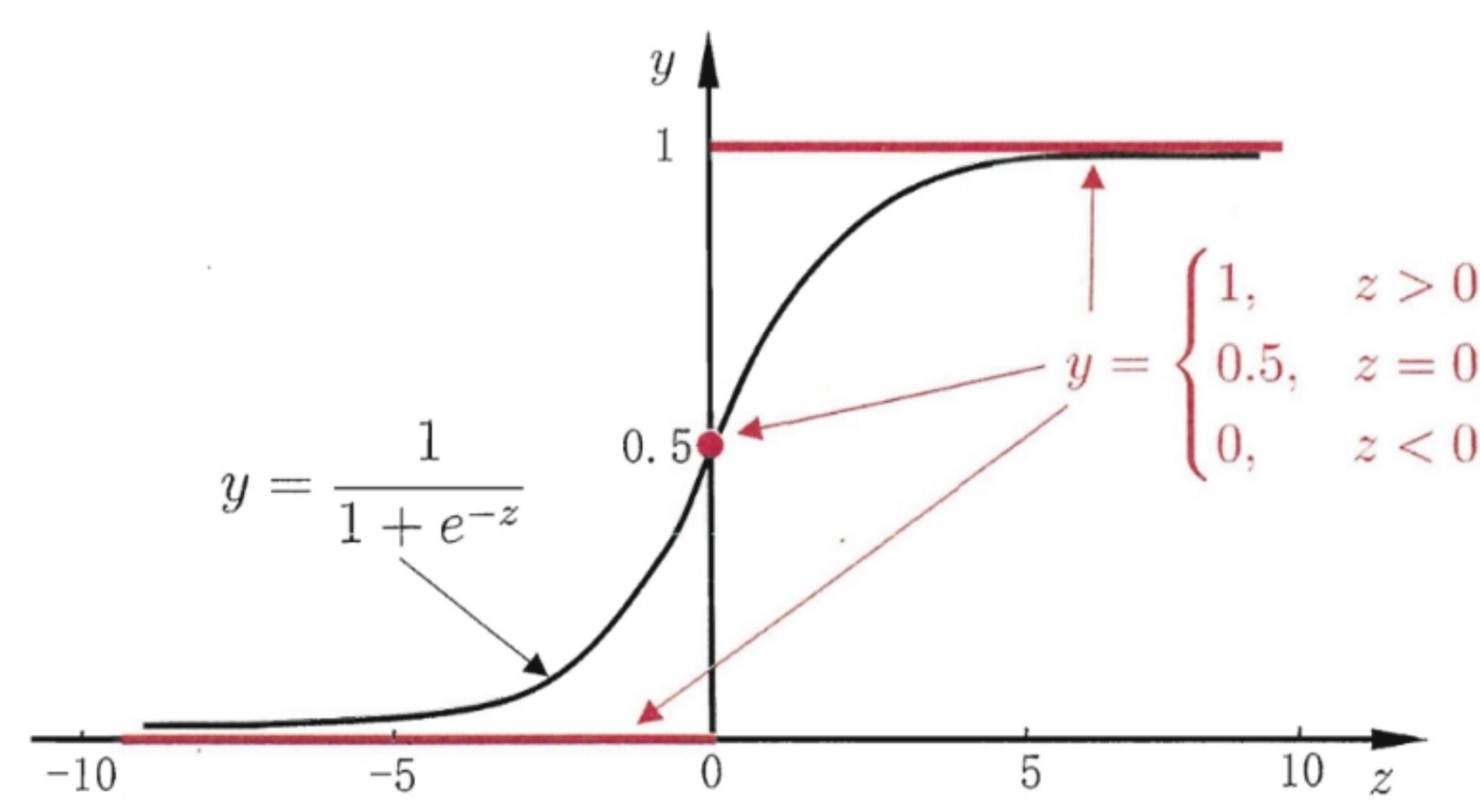
$$y = g^{-1}(w^T x + b), \text{ 其中 } g(y) = w^T x + b$$

二分类任务 $y \in \{0, 1\}$ $z = w^T x + b$

单位阶跃函数 $\begin{cases} z < 0 & y=0 \\ z=0 & y=0.5 \\ z > 0 & y=1 \end{cases}$ 不连续，性质差

key-point $(-\infty, +\infty) \rightarrow (0, 1)$

$$\text{对数几率函数 } y = \frac{1}{1+e^{-z}}$$



平滑、连续、处处可导

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

odds $\frac{y}{1-y}$ 反映正例/反例的相对概率

对数几率 $\log \text{odds} \Rightarrow \text{logistic}$

$$y = P(y=1|x)$$

$$w^T x + b = \ln \frac{P(y=1|x)}{P(y=0|x)}$$

极大似然估计

$$\text{loss}(w, b) = \sum_{i=1}^n \ln P(y_i | x_i; w, b)$$

$$(w^*, b^*) = \arg \max_{(w, b)} \text{loss}(w, b)$$

$$P(y_i | x_i; w, b) \text{ 可以写成 } y_i P_1(x_i; \beta) + (1-y_i) P_0(x_i; \beta), \text{ 还可以写成 } P_1^{y_i} P_0^{1-y_i}$$

$$\begin{aligned}
 \text{最大化} \quad & y_i \ln p_i + (1-y_i) \ln p_o = y_i \ln \frac{e^{\beta^T x}}{1+e^{\beta^T x}} + (1-y_i) \ln \frac{1}{1+e^{\beta^T x}} \\
 &= y_i \beta^T x - y_i \ln(1+e^{\beta^T x}) + y_i \ln(1+e^{\beta^T x}) - \ln(1+e^{\beta^T x}) \\
 &= y_i \beta^T x - \ln(1+e^{\beta^T x})
 \end{aligned}$$

↓

高阶可导连续凸函数

优化问题

$$\min_f \frac{1}{m} \sum_{i=1}^m l(f(x_i), y_i) + R(f)$$

高阶可导的连续凸函数可以 梯度下降法、牛顿法

线性判别分析 LDA

监督降维技术

找一个平面降维 s.t. 类内小，类间大

给定数据集 $\{(x_i, y_i)\}_{i=1}^m$

第*i*类样本集合 X_i

第*i*类均值向量 μ_i

第*i*类协方差矩阵 Σ_i

样本中 μ_i 在直线上的投影 $w^T \mu_i$

投影后样本的协方差 $w^T \Sigma_i w$

同类样本接近 $w^T \Sigma_0 w + w^T \Sigma_1 w$ 小

异类样本远离 $\|w^T \mu_0 - w^T \mu_1\|_2^2$ 大

$$\begin{aligned}
 \text{最大化} \quad J &= \frac{\|w^T \mu_0 - w^T \mu_1\|_2^2}{w^T \Sigma_0 w + w^T \Sigma_1 w} = \frac{w^T (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w}{w^T (\Sigma_0 + \Sigma_1) w}
 \end{aligned}$$

within 类内散度矩阵 $S_w = \Sigma_0 + \Sigma_1$

between 类间散度矩阵 $S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$

广义瑞利商

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

$$\text{优化 } \min -w^T S_b w$$

w 缩放不影响 J 取值, 且 $w^T S_w w = 1$

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

拉格朗日乘子法, $S_b w \rightarrow S_w w$, 由于 $S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w$ 故

$$\text{取 } \lambda S_w w = \lambda(\mu_0 - \mu_1) \Rightarrow w = S_w^{-1}(\mu_0 - \mu_1)$$

当 $S_w = I$, 即两类数据均一时, LDA 达最优分类

推广到多类 N

$$S_t = S_b + S_w$$

$w \in R^{d \times (N-1)}$ 投影到 $N-1$ 维分类问题

$$\max_w \frac{\text{tr}(w^T S_b w)}{\text{tr}(w^T S_w w)}$$

trace ratio

tr: 投影后散度和

也可以
重要的优化标量

转化

$$\max_w \text{tr} \left(\frac{w^T S_b w}{w^T S_w w} \right) \text{ ratio trace}$$

$$= \max_w \text{tr} \left((w^T S_w w)^{-1} w^T S_b w \right)$$

||

闭式解 $S_b w = \lambda S_w w$

$S_w^{-1} S_b$ 的 d 个非零特征值对应的特征向量

即

多分类学习

OvO 拆解为 C^2 个二分类问题，计票数

OvR N 个正例 - 反例 分类问题

MvM ECOC 编码

类别不平衡问题

传统对称回归 $\frac{y}{1-y} > 1$ 预测为正例

↓

$\frac{y}{1-y} > \frac{m^+}{m^-}$ 正例数
负例数 预测为正例

$$\text{等价于再缩放: } \frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$$

其它: 过采样 代价敏感 \Rightarrow 简单复杂, SMOTE

欠采样 简单丢弃, 多分类器 bagging 融合

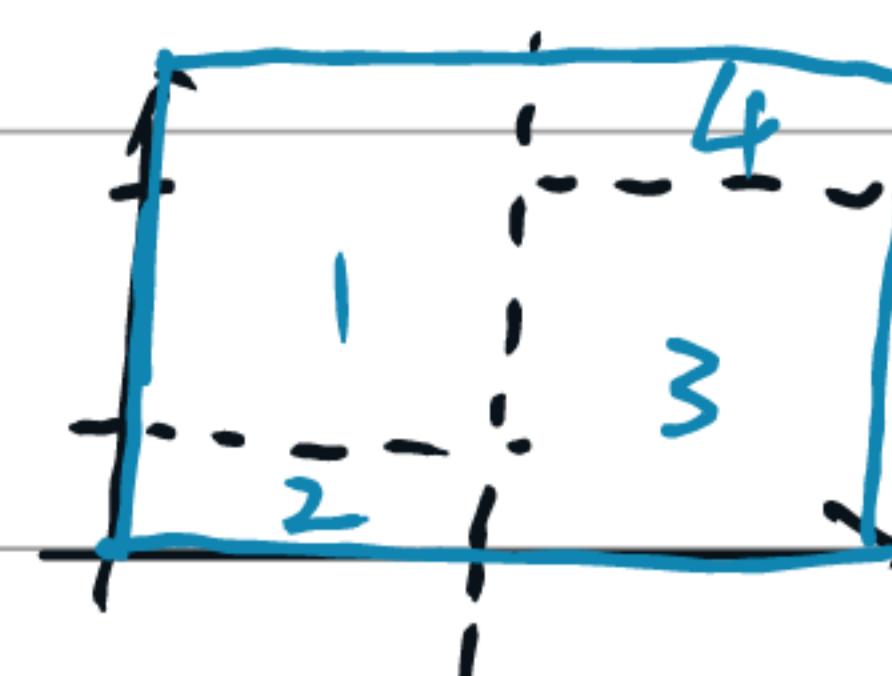
阈值移动 $\frac{y}{1-y} > 1$

决策树

内部结点：一个对属性的 test

分支：属性的一个取值

叶结点：预测结果



规则地

通过数据学到这样的树结构，将特征空间 切分成多个局部（叶结点），局部的计算简单

输入：训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$;

属性集 $A = \{a_1, a_2, \dots, a_d\}$.

过程：函数 TreeGenerate(D, A)

1: 生成结点 node;

2: if D 中样本全属于同一类别 C then

3: 将 node 标记为 C 类叶结点; return ①y相同

4: end if

5: if $A = \emptyset$ OR D 中样本在 A 上取值相同 then ②x相同

6: 将 node 标记为叶结点，其类别标记为 D 中样本数最多的类; return

7: end if

8: 从 A 中选择 最优划分属性 a_* ;

9: for a_* 的每一个值 a_*^v do

10: 为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;

11: if D_v 为空 then

12: 将分支结点标记为叶结点，其类别标记为 D 中样本数最多的类; return ③D空

13: else

14: 以 TreeGenerate($D_v, A \setminus \{a_*\}$) 为分支结点 递归

15: end if

16: end for

输出：以 node 为根结点的一棵决策树

$$\min \sum_{i=1}^m \sum_j I(x_i \in R_j) l(f_j(x_i), y_i)$$

样本集合 D 中第 k 类样本所占比例为 P_k (即 $y=k$ 的比例)， D 的信息熵定义为

$$Ent(D) = - \sum_{k=1}^{|Y|} P_k \log_2 P_k \quad \text{约定 } P_k = 0 \text{ 时 } P_k \log_2 P_k = 0$$

$Ent(D)$ 越小， D 的纯度越高

假定此时对属性 a 进行划分，取值 $\{a^1, a^2, \dots, a^V\}$ ，用 a 划分 D 的信息增益为

$$Gain(D, a) = Ent(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} Ent(D^v)$$

$Gain(D, a)$ 越大，用 a 划分的纯度提升越大

于是当使用信息增益来划分属性时，算法第 8 行 $a_* = \arg \max_{a \in A} Gain(D, a)$

ID3

“信息增益”对取值较多的属性有所偏好，可能导致泛化能力差，于是

定义 增益率

$$Gain_ratio(D, a) = \frac{Gain(D, a)}{IV(a)}$$

其中, $JV(a) = -\sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$ 称为 a 的固有值

增益率可能对取值数目较少的属性有偏好, C4.5 启发式

基尼值 $Gini(D) = \sum_{k=1}^{|Y|} \sum_{k' \neq k} P_k P_{k'} = 1 - \sum_{k=1}^{|Y|} P_k^2$, $Gini \downarrow$ 纯度↑

基尼指数 $Gini_index(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} Gini(D^v)$

$$a_* = \underset{a \in A}{\operatorname{argmin}} Gini_index(D, a)$$

CART

划分选择的各种准则对树的尺寸有较大影响, 但对泛化性能影响有限

剪枝: 为应付决策树过拟合, 提升泛化性能

预剪枝: 提前终止一些分支的生长

后剪枝: 生长完全再回头剪枝

评估剪枝前后的性能变化(验证集)

后剪枝往往保留更多分支, 久拟合风险小, 泛化性能更好; 但要等树完全生长到底向上逐一考察, 时间开销很大

连续值离散化 样本有限, 选定一个样本值作二分

缺失值 均值补全、众数补全; 决策树可以直接处理 C4.5 负责

从树到规则

轴平行划分 → 多变量决策树(斜决策树): 非叶结点是一个线性分类器

one-hot 编码 属性 $a_1(1, 0, 0, 0)$ $a_2(0, 1, 0, 0)$ $a_3(0, 0, 1, 0)$ $a_4(0, 0, 0, 1)$

主成分分析 (chapter 10) principal component analysis PCA

典型降维方法：结构特征、可视化、数据预处理

对于正交属性空间中的样本点，用一个超平面对所有点进行恰当表达

最近重构性：样本点到超平面距离足够近
两个性质等价
最大可分性：样本点在超平面上的投影尽可能分开

最大可分

样本中心化 (均值为0), 超平面 $W \in \mathbb{R}^{d \times d'}$, 投影 $W^T X \in \mathbb{R}^{d' \times 1}$, 投影后样本方差最大 $\sum_i W^T x_i x_i^T W$

于是 $\min_W -\text{tr}(W^T X X^T W)$ s.t. $W^T W = I$ 正交矩阵限制

拉格朗日乘子法: $X X^T W = \lambda W$, 对 $X X^T$ 进行特征值分解

最近重构

样本中心化, 假设降维后的坐标系为 $\{w_1, w_2, \dots, w_{d'}\}$, 一组标准正交基, 即 $\forall i, \|w_i\|_2 = 1; w_i^T w_j = 0, i \neq j$

$W = (w_1, w_2, \dots, w_{d'})$, $w_i^T x_i = (w_1^T x_i, w_2^T x_i, \dots, w_{d'}^T x_i) = z_i$, 重构 $\hat{x}_i = \sum_j z_{ij} w_j$

$\sum_i \|\hat{x}_i - x_i\|_2^2 \propto -\text{tr}(W^T (\sum_i x_i x_i^T) W)$

d' 选取: 直接指定、 $\sum_i \|\hat{x}_i - x_i\|_2^2, \frac{\sum_i \lambda_i}{\sum_i \lambda_i}$

自编码视角的PCA

Robust PCA

函数推导

存在相关性 $\xrightarrow{\text{正交变换}}$ 线性无关的新变量(主成分)

发现结构、数据降维，方差大表示蕴含更多信息。

几何直观：中心化后的坐标系旋转 方差和 + 平面距离平方 = 原点距离平方和，最大方差 \Leftrightarrow 最小距离

协方差 $\text{cov}(X, Y)$ 反映 X 与 Y 之间的相关性

标量随机变量 x, y , $\text{cov}(x, y) = E[(x - \bar{x})(y - \bar{y})]$

$$\text{对 } x_i, y_i (i=1, 2, \dots, n) \text{ 有 } \text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

特别地, $\text{cov}(x, x) = E[(x - \bar{x})^2] = D(x)$ 方差

$$\text{当 } \bar{x} = \bar{y} = 0 \text{ 时, } \text{cov}(x, y) = \frac{\sum_{i=1}^n x_i y_i}{n-1}$$

当 X 是 d 维列向量, X_i 表示第 i 个样本, X_{ij} 表示其第 j 个属性, 协方差是针对属性而言的

$\Sigma = (\Sigma_{jk})$, 其中 $\Sigma_{jk} = \text{cov}(x_{\cdot j}, x_{\cdot k}) = \frac{\sum_{i=1}^m x_{ij} x_{ik}}{m-1}$, 于是考虑数据 $X = [x_1, x_2, \dots, x_m] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1j} & x_{2j} & \cdots & x_{mj} \\ x_{1d} & x_{2d} & \cdots & x_{md} \end{bmatrix}$

$$\Sigma = \frac{1}{m-1} X X^T$$

找一组标准正交基 $\{\alpha_1, \alpha_2, \dots, \alpha_d\}$ $\left\{ \begin{array}{l} \alpha_i^T \alpha_i = 1 \\ \alpha_j^T \alpha_i = 0 \quad i \neq j \end{array} \right.$

将 X 投影到这组基上, 求系数 $X = y_1 \alpha_1 + y_2 \alpha_2 + \dots + y_d \alpha_d$, 由于 $\forall i$, $\alpha_i^T \alpha_i = 1$ 故 $y_i = \frac{\alpha_i^T X}{\alpha_i^T \alpha_i} = \alpha_i^T X$ ($\alpha_i^T \alpha_i = 1$)

于是 X 在这组基上的坐标为 $\begin{bmatrix} \alpha_1^T X \\ \alpha_2^T X \\ \vdots \\ \alpha_d^T X \end{bmatrix}$, 记正交阵 $W = [\alpha_1, \alpha_2, \dots, \alpha_d]$, 坐标可表示为 $W^T X$, 该组数据就表示为 $W^T X = \begin{bmatrix} \text{第1列} & \text{第2列} & \cdots & \text{第m列} \\ 1 & 2 & \cdots & m \end{bmatrix}$

让 X 投影后, 在 $\alpha_1, \alpha_2, \dots, \alpha_d$ 上更量协方差阵为 $\Sigma_\alpha = \frac{1}{m-1} (W^T X)(W^T X)^T = W^T \Sigma W$

总体主成分分析

$X = (x_1, x_2, \dots, x_m)^T$ 是 m 维随机变量, 均值向量 $\mu = E(X)$

协方差阵 $\Sigma = \text{cov}(X, X) = E[(X - \mu)(X - \mu)^T]$

由 X 到 $Y = (y_1, y_2, \dots, y_m)^T$ 的线性变换: $y_i = \alpha_i^T X$, $\alpha_i^T = (\alpha_{1i}, \alpha_{2i}, \dots, \alpha_{mi})$

可以导出: $E(y_i) = \alpha_i^T \mu$

$$\text{Var}(y_i) = E[(\alpha_i^T X - \alpha_i^T \mu)(\alpha_i^T X - \alpha_i^T \mu)^T] = E[\alpha_i^T (X - \mu)(X - \mu)^T \alpha_i] = \alpha_i^T \Sigma \alpha_i$$

$$\text{cov}(y_i, y_j) = \alpha_i^T \Sigma \alpha_j$$

def 总体主成分：给定上述线性变换，满足下列条件

(1) 系数向量 α_i 是单位向量，即 $\alpha_i^T \alpha_i = 1$

(2) y_i, y_j 互不相关， $\text{cov}(y_i, y_j) = 0 \quad (i \neq j)$

(3) y_1 是 x 所有线性变换方差最大的， y_2 是与 y_1 不相关的 x 所有线性变换方差最大的 ...

$$y_k = \alpha_k^T x, \text{var}(y_k) = \alpha_k^T \Sigma \alpha_k = \lambda_k$$

Pf. 拉格朗日乘子法 $\max_{\alpha_1} \alpha_1^T \Sigma \alpha_1 \text{ s.t. } \alpha_1^T \alpha_1 = 1$

$$f(\alpha_1, \lambda) = \alpha_1^T \Sigma \alpha_1 - \lambda (\alpha_1^T \alpha_1 - 1)$$

$$\frac{\partial f}{\partial \alpha_1} = \Sigma \alpha_1 - \lambda \alpha_1 = 0 \quad \text{于是 } \alpha_1 \text{ 是特征向量}$$

求解 α_2 同理，值得注意的是 $0 = \alpha_2^T \Sigma \alpha_1 = \lambda \alpha_2^T \alpha_1$ 得到正交

定义条件 $y = A^T x$ 等价于：(1) A 正交阵

(2) y 的方差阵为 $\text{cov}(y) = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m) \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$

$$\Sigma A = A \Delta, \Sigma = A \Delta A^T, \text{var}(y_k) = \alpha_k^T \Sigma \alpha_k = \lambda_k, y_k \text{ 方差} = \sum \lambda_i = \text{tr}(\Sigma) = \sum \sigma_{ii} = x \text{ 的方差和}$$

因子负荷量 (factor loading)

或者 $\text{tr}(\Sigma) = \text{tr}(A \Delta A^T) = \text{tr}(\Delta)$

$$P(y_k, x_i) = \frac{\text{cov}(y_k, x_i)}{\sqrt{\text{var}(y_k) \text{var}(x_i)}} = \frac{\text{cov}(\alpha_k^T x, e_i^T x)}{\sqrt{\lambda_k} \sqrt{\sigma_{ii}}} = \frac{\alpha_k^T \Sigma e_i}{\sqrt{\lambda_k} \sqrt{\sigma_{ii}}} = \frac{\lambda_k e_i^T \alpha_k}{\sqrt{\lambda_k} \sqrt{\sigma_{ii}}} = \sqrt{\frac{\lambda_k}{\sigma_{ii}}} \alpha_{ik}$$

$$\sum_{i=1}^m \sigma_{ii} P^2(y_k, x_i) = \lambda_k$$

$$\sum_{k=1}^m P^2(y_k, x_i) = 1, \quad y \text{ 互不相关}, P^2(x_i, (y_1, y_2, \dots, y_m)) = \sum_{k=1}^m P^2(y_k, x_i)$$

x_i 为 $y_1 \dots y_m$ 线性组合， $P^2(x_i, (y_1 \dots y_m)) = 1$

考虑整数 $q, 1 \leq q \leq m, y = B^T x, B \in \mathbb{R}^{m \times q}$ 由 q 个 m 维向量组成， $y \in \mathbb{R}^q$ 降维空间

y 的协方差阵为 $\Sigma_y = B^T \Sigma B, \text{tr}(\Sigma_y)$ 在 $B = A_q$ 时取得最大值 (A 的前 q 列)

方差贡献率 $\eta_k = \frac{\lambda_k}{\sum_{i=1}^m \lambda_i}, \text{ 用 } \sum_{i=1}^q \eta_i \text{ 来取 } q \text{ 值}$

支持向量机 SVM support vector machine

概论

定义在特征空间上间隔最大的线性可分器 . 核技巧、非线性

优化：求解凸二次规划

线性可分支持向量机 — 硬间隔最大化

二分类、输入空间一一对应 \rightarrow 特征空间、训练集在特征空间上线性可分

用一个超平面分类， w 表示其法向量，并指向正例， (w, b) 唯一表示一个超平面，用 $\text{sign}(w^T x + b)$ 作为分类决策函数

函数间隔：对于 (x_i, y_i) , $\hat{\gamma}_i = y_i(w^T x_i + b)$

y_i 的作用是 $\begin{cases} \text{①没有异常分类：若有，变成负数，但优化目标是 } \max \hat{\gamma} \\ \text{②正确分类训练数据时，为正数表示距离} \end{cases}$

函数间隔最小值： $\hat{\gamma} = \min_i \hat{\gamma}_i$

注意到成比例改变 (w, b) 时（如 $(2w, 2b)$ ），超平面不变，但是 $\hat{\gamma}$ 改变（2倍）

几何间隔： $\gamma_i = y_i \left(\frac{w^T}{\|w\|} x_i + \frac{b}{\|w\|} \right) = \frac{\hat{\gamma}_i}{\|w\|}$

几何间隔最小值： $\gamma = \min_i \gamma_i = \frac{1}{\|w\|} \hat{\gamma}$

约束优化问题： $\max_{w, b} \gamma \quad \text{s.t. } y_i \left(\frac{w^T x_i}{\|w\|} + \frac{b}{\|w\|} \right) \geq \gamma, i=1, 2, \dots, N$

$\max_{w, b} \frac{\hat{\gamma}}{\|w\|} \quad \text{s.t. } y_i(w^T x_i + b) \geq \hat{\gamma}, i=1, 2, \dots, N$

$\hat{\gamma}$ 的取值并不影响最优化问题的解，取 $\hat{\gamma}=1$ ，问题等价于

$\min_{w, b} \frac{1}{2} \|w\|^2 \quad \text{s.t. } y_i(w^T x_i + b) - 1 \geq 0, i=1, 2, \dots, N$

凸二次规划问题

支持向量 x_i 是使 $y_i(w^T x_i + b) = 1$ 成立的样本点、实例

支持向量是训练数据中很重要的样本，个数一般很少

间隔边界： $H_1: w^T x + b = 1$

$H_2: w^T x + b = -1$

对于其它样本，在间隔边界外移动，甚至去掉
都不会影响解

H_1 与 H_2 之间的距离 $\frac{2}{\|w\|}$ 称为间隔

对偶算法

定义拉格朗日函数 $L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i y_i (w^T x_i + b) + \sum_{i=1}^N \alpha_i$, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$

原始问题的对偶问题为 $\max_{\alpha} \min_{w, b} L(w, b, \alpha)$ 拉格朗日对偶性

$$\nabla_w L = w - \sum \alpha_i y_i x_i = 0$$

$$\nabla_b L = -\sum_i \alpha_i y_i = 0$$

$$\text{代入得} L = \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j) - \sum_i \alpha_i y_i ((\sum_j \alpha_j y_j x_j^T) \cdot x_i + b) + \sum_i \alpha_i$$

$$\min_{w, b} L = -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j) + \sum_i \alpha_i$$

$$\text{于是对偶问题为 } \min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j) - \sum_i \alpha_i \quad \text{s.t. } \sum_i \alpha_i y_i = 0, \alpha_i \geq 0 \quad i=1, 2, \dots, N$$

KKT条件 $\begin{cases} \alpha_i \geq 0 \\ 1 - y_i f(x_i) \leq 0 \\ \alpha_i(1 - y_i f(x_i)) = 0 \end{cases} \Rightarrow \forall i \quad \alpha_i = 0 \text{ 或 } y_i f(x_i) = 1$

支持向量
通过求解 α^* 来求 w^*, b^*

线性支持向量机——软间隔最大化

线性不可分，但去除少部分特异点后线性可分

引入松弛变量 $y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$

目标函数 $\frac{1}{2} \|w\|^2 + C \sum_i \xi_i, C > 0$ 被称为惩罚参数

引入松弛变量 $y_i(w^T x_i + b) \geq 1 - \xi_i, i=1, 2, \dots, N$

$\xi_i \geq 0, i=1, 2, \dots, N$

对偶问题 $\max_{\alpha} -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j) + \sum_i \alpha_i$

s.t. $\sum_i \alpha_i y_i = 0, i=1, 2, \dots, N$

$0 \leq \alpha_i \leq C, i=1, 2, \dots, N$

支持向量：边界、边界内、错误分类

损失函数

核技巧 (kernel trick)

输入空间 非线性可分 \rightarrow 特征空间 线性可分

非线性变换 \rightarrow 超平面

X : 输入空间: 欧氏空间 R^n 的子集或离散集合

映射 $\phi: X \rightarrow H$ H : 特征空间: 希尔伯特空间

核函数 $k(x, z) = \phi(x)^T \phi(z)$ 表示内积

实际应用中只关心 k 而不管 ϕ , 也不管 H 是什么, 这是困难的

对于一个 $k(\cdot, \cdot)$, 是否存在 ϕ , $k(x, z) = \phi(x) \cdot \phi(z)$ —— 满足的称为正定核, 一般直接称核函数

具体数学略

常用核函数

多项式核函数 $k(x, z) = (x \cdot z + 1)^p$ $p \geq 1$

高斯核函数 $k(x, z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$ $\sigma > 0$

字符串核函数

线性核 $k(x, z) = x \cdot z$

拉普拉斯核 $k(x, z) = \exp\left(-\frac{\|x-z\|}{\sigma}\right)$ $\sigma > 0$

Sigmoid 核 $k(x, z) = \tanh(\beta x \cdot z + \theta)$ $\beta > 0, \theta < 0$

组合
$$\begin{cases} \gamma_1 k_1 + \gamma_2 k_2 \\ k_1 \cdot k_2 \\ g(x) k(x, z) g(z), 任意函数 g \end{cases}$$

非线性支持向量机

$$f(x) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i K(x, x_i) + b^*\right)$$

优化 $\min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_i \alpha_i$

s.t. $\sum_i \alpha_i y_i = 0$

$0 \leq \alpha_i \leq C, i = 1, 2, \dots, N$

算法 7.5 (SMO 算法)

输入: 训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中, $x_i \in \mathcal{X} = \mathbf{R}^n$, $y_i \in \mathcal{Y} = \{-1, +1\}$, $i = 1, 2, \dots, N$, 精度 ε ;

输出: 近似解 $\hat{\alpha}$.

(1) 取初值 $\alpha^{(0)} = 0$, 令 $k = 0$;

(2) 选取优化变量 $\alpha_1^{(k)}, \alpha_2^{(k)}$, 解析求解两个变量的最优化问题 (7.101)~(7.103), 求得最优解 $\alpha_1^{(k+1)}, \alpha_2^{(k+1)}$, 更新 α 为 $\alpha^{(k+1)}$;

(3) 若在精度 ε 范围内满足停机条件

$$\sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N$$
$$y_i \cdot g(x_i) \begin{cases} \geq 1, & \{x_i | \alpha_i = 0\} \\ = 1, & \{x_i | 0 < \alpha_i < C\} \\ \leq 1, & \{x_i | \alpha_i = C\} \end{cases}$$

其中,

$$g(x_i) = \sum_{j=1}^N \alpha_j y_j K(x_j, x_i) + b$$

则转 (4); 否则令 $k = k + 1$, 转 (2);

(4) 取 $\hat{\alpha} = \alpha^{(k+1)}$.

序列最小最优化 (SMO) 算法

不断将原二次规划问题转化为两个变量的二次规划问题(闭式解), 不断迭代直到所有变量满足 KKT 条件

深灰学[——

贝叶斯分类器

- 贝叶斯决策论：

λ_{ij} 表示将第 j 类误分类为第 i 类的损失，定义基于后验概率将样本 x 分类到 i 的条件风险为

$$R(c_i|x) = \sum_j \lambda_{ij} P(c_j|x)$$

- 贝叶斯判定准则

$$h^*(x) = \arg \min_{c \in \mathcal{Y}} R(c|x)$$

h^* 称为贝叶斯最优分类器，反映学习性能的理论上限

- 判别式模型 決策树、MLP、SVM vs. 生成式模型 贝叶斯分类器（不是“贝叶斯学习”）

直接对 $P(c|x)$ 建模

对 $P(c,x)$ 建模，再计算 $P(c|x)$

$$P(c|x) = \frac{P(c) P(x|c)}{P(x)}$$

$$\text{后验} = \frac{\text{先验} \times \text{似然}}{\text{证据因子}}$$

- 确定参数：最小化 loss function (最小二乘法、梯度下降)

$$N(y_i; \omega^T x_i + b, \sigma^2)$$

$$\text{极大似然估计 } \max \prod_i P(\text{target}_i | x_i, \theta) = \prod_i P(\text{target}_i | x_i, \theta) = \dots$$

直乘容易下溢，对数似然 $\min -\log P(\cdot)$

与最小二乘（均方误差）一样

- 朴素贝叶斯分类器

证据因子相同，不管

先验 $P(c)$ 很好估计

$x \in \mathbb{R}^d$, 组合爆炸

似然 $P(x|c)$

样本稀疏

若属性相互独立, $P(x|c) = \prod_{i=1}^d P(x_i|c)$

离散 频率估计

连续 高斯分布

什么是独立同分布，为什么要独立同分布

拉普拉斯修正：新的属性值会导致 $P(a|c) = 0$ 导致连乘概率为 0 \rightarrow 每种属性值数量加 1

了解具体怎么算

trick: 预计算 增量学习
数据变化 → 懒惰计算

一半朴素贝叶斯 属性之间的依赖关系

ODE 独立依赖估计

KDE K依赖

贝叶斯网

联合概率 → 条件概率

如何化简联合概率

同父结构、V型结构、顺序结构

有向分离

有向无环 → 无向

(道德图 moral graph)

如何获得这样的结构? 结构学习, 使用评分函数(一般基于信息论)

最小描述长度: 给定数据集 D, 贝叶斯网 $B = (G, \Theta)$

$$S(B|D) = f(\theta) |B| - LL(B|D)$$

注解: $f(\theta)$ 的选择: 每个参数需要的参数个数 \downarrow $\sum_{i=1}^m \log P_B(x_i)$ 对数据做似然插进

$S \downarrow = \text{复杂程度} \downarrow - \text{对数似然} \uparrow$

模型选择 — 模型优化

贝叶斯网的推断: 给定一些属性观测值(证据), 推断其它属性变量的取值

精确推断/近似推断

吉布斯采样

变分推断 真实分布 $p(x)$ 不好算, 用一个简单的分布 $q(x)$, 优化 $D(p, q)$

EM 估计隐变量 贝叶斯 chapter 9, 13

集成学习

聚类 (无监督学习的一种, 其它: 密度估计)

没有绝对的好坏

原型聚类
用一个点来表示一个簇(中心点)

K-means

分配矩阵
 $\|X - GP\|_F^2$ 交替优化 G 和 P

GMM 高斯混合聚类
EM 算法

密度聚类

DBSCAN

层次聚类

AGNES

1.6

10:30-12:30 仙二

问答 1题10分 1章知识点 / 10题

带计算器

计算、概念、开放(实际问题
如何解决)