PAC probably approximately correct

$$P(|f(x)-y| \le \varepsilon) \ge 1-\delta$$

独立同分布

归纳偏好 → 奥卡姆剃刀

NFL $\sum_f E_{ote}(\mathfrak{L}_a|X,f) = \sum_f \sum_h \sum_{x\in\mathcal{X}-X} P(x)\,\mathbb{I}(h(x)\neq f(x))\,P(h|X,\mathfrak{L}_a)$

分析不能脱离具体的问题 $= \sum_x P(x) \sum_h P(h|X,\mathfrak{L}_a) \sum_f \mathbb{I}(h(x)\neq f(x))$

$= \sum_x P(x) \sum_h P(h|X,\mathfrak{L}_a)\,\frac{1}{2}2^{|\mathcal{X}|}$

$= \frac{1}{2}2^{|\mathcal{X}|} \sum_x P(x)\cdot 1$

模型评估

1、留出法：$D=S\cup T,\ S\cap T=\emptyset$　　S：训练集　T：测试集

- 数据分布一致
- 多次重复划分
- $|T|$不能太大，也不能太小

2、k折交叉验证法：将D划分为k个相同大小的子集 $D_1 \cdots D_k$，每次用k-1个训练，1个测试

当$k=m$时，退化为留一法，计算开销大

3、自助法：采样相同数量(可重复采样)，未采样的作测试

- 训练集同规模
- 数据分布有所改变

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m \to \frac{1}{e}\approx 0.368$$

回归

均方误差 $E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$

错误率 $E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)\neq y_i)$

精度 $acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)=y_i)=1-E(f;D)$

分类

查准率 $P=\frac{TP}{TP+FP}$

查全率 $R=\frac{TP}{TP+FN}$

$\frac{1}{F1}=\frac{1}{2}\left(\frac{1}{P}+\frac{1}{R}\right) \Rightarrow F1=\frac{2PR}{P+R}=\frac{2TP}{total+TP-TN}$

PR图、BEP

宏 macro　把P、R算出来再平均

微 micro　先平均TP等再算P、R

ROC. AUC，右下面积↑

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

比较检验

偏差-方差　$E(f;D)=bias^2(x)+var(x)+\varepsilon^2 \rightarrow E_D[(y_D-y)^2]$

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

一、线性模型　$f(x)=w^Tx+b$

1、线性回归

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

求导 $2X^T(X\hat{w}-y)=0 \Rightarrow$ 若X^TX满秩或正定 $\hat{w}^*=(X^TX)^{-1}X^Ty$；否则 有多个解？(正则化引入归纳偏好)

2、广义线性模型　$y=g^{-1}(w^Tx+b)$　（单调可微）

$y=\frac{e^z}{1+e^z}$，$\ln y^{y_i}(1-y)^{1-y_i}=\ln\frac{e^{zy_i}}{1+e^z}=zy_i-\ln(1+e^z)$

对率回归　$y=\frac{1}{1+e^{-z}}$，$\ln\frac{y}{1-y}=w^Tx+b$

最大对数似然　$\ell(w,b)=\sum_{i=1}^{m}\ln p(y_i|x_i;w,b)=\sum_{i=1}^{m}\left(y_i\beta^T\hat{x}_i-\ln(1+e^{\beta^T\hat{x}_i})\right)$

3. LDA (Linear Discriminant Analysis)　分类、监督降维

类内散度矩阵 $S_w=\Sigma_0+\Sigma_1=\sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^T+\sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^T$

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

$$J=\frac{w^TS_bw}{w^TS_ww}$$

求解：取$w^TS_ww=1$，$\min_w -w^TS_bw$　s.t. $w^TS_ww=1$

拉格朗日乘子法　$S_bw=\lambda S_ww \Rightarrow (\mu_0-\mu_1)(\mu_0-\mu_1)^Tw=\lambda S_ww$　令$(\mu_0-\mu_1)^Tw=\lambda$　$\Rightarrow w=S_w^{-1}(\mu_0-\mu_1)$

多分类LDA

多分类学习
- OvO 拆为 C_n^2 个二分类，计票数
- OvR 拆为 n 个 正例(1个)-反例(n-1个) 问题
- MvM 编码 正例-反例问题

类别不平衡
- 再缩放（阈值移动）：$\frac{y}{1-y} > 1 \longrightarrow \frac{y}{1-y} > \frac{m^+}{m^-}$
- 过采样
- 欠采样

二、PCA Principal Component Analysis

降维到超平面上
- 最大可分
- 最近重构

① 首先对样本进行中心化，考虑投影后样本点的方差最大（最大可分）

$$\max_W \ tr(W^T X X^T W) \quad s.t.\ W^T W = I$$

拉格朗日乘子法 $XX^T W = \lambda W$，W由 XX^T 特征向量组成

② 假设投影后新坐标系为 $\{w_1, w_2, \cdots, w_{d'}\}$，是标准正交基向量，$\hat{x}_i = \sum_{j=1}^{d'} z_{ij} w_j$

$$\sum_{i=1}^{m} \left\| \sum_{j=1}^{d'} z_{ij} w_j - x_i \right\|_2^2 = \sum_{i=1}^{m} z_i^T z_i - 2\sum_{i=1}^{m} z_i^T W^T x_i + const \propto -tr\left(W^T \left(\sum_{i=1}^{m} x_i x_i^T\right) W\right) \Rightarrow \min_W -tr(W^T X X^T W) \quad s.t.\ W^T W = I \text{（标准正交）}$$

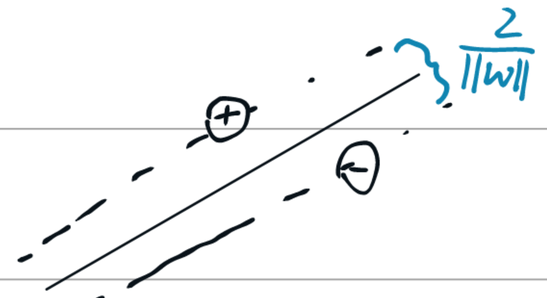
设置 d'　重构误差判断 $\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq 95\%$，交叉验证

拓展1：自编码：$\min_{W: W^T W = I} \sum_{i=1}^{m} \|x_i - W z_i\|_2^2 = \sum_{i=1}^{m} \|x_i - W W^T x_i\|_2^2$（decoder / encoder）

拓展2：Robust PCA：$\min_{rank(\hat{x})=d'} \|x - \hat{x}\|_2^2 \rightarrow \min_{\hat{x}} \|x - \hat{x}\|_0 + rank(\hat{x}) \rightarrow \min_{\hat{x}} \|x - \hat{x}\|_1 + \|\hat{x}\|_*$

拓展3：函数空间傅立叶级数重构

三、SVM



超平面 $W^T x + b = 0$ 分割数据　$\max_{w,b} \frac{2}{\|w\|}$　$s.t.\ y_i(W^T x_i + b) \geq 1 \quad i = 1, 2, \cdots, m \Rightarrow \min \frac{1}{2}\|w\|^2 \quad s.t.\ y_i(W^T x_i + b) \geq 1$

拉格朗日乘子法 $L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(W^T x_i + b))$

偏导 $\begin{cases} w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \\ \sum_{i=1}^{m} \alpha_i y_i = 0 \end{cases}$

回代 $\frac{1}{2}\left(\sum_i \alpha_i y_i x_i\right)^T \left(\sum_i \alpha_i y_i x_i\right) + \sum \alpha_i - \sum \alpha_i y_i W^T x_i - \sum \alpha_i y_i b$

$= \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_i \alpha_i y_i \left(\sum_j \alpha_j y_j x_j^T x_i\right) + \sum \alpha_i$

$= -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum \alpha_i$

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \quad s.t. \sum_{i=1}^{m}\alpha_i y_i = 0,\ \alpha_i \geq 0$$

$$f(x) = w^T x + b = \sum_{i=1}^{m}\alpha_i y_i x_i^T x + b \qquad \text{KKT条件} \begin{cases} \alpha_i \geq 0 \\ 1 - y_i f(x_i) \leq 0 \\ \alpha_i(1 - y_i f(x_i)) = 0 \end{cases} \Rightarrow \forall i \quad \alpha_i = 0 \text{ 或 } y_i f(x_i) = 1$$

SMO：每次针对一对 α_i, α_j 闭式解优化

特征空间映射 $f(x) = w^T\phi(x) + b$，$k(x_i, x_j) = \phi(x_i)^T\phi(x_j)$

软间隔 结构风险 经验风险 正则化，贝叶斯先验，归纳偏好

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} loss(y_i(w^T x_i + b) - 1)$$

$$l_{0/1}(z) = \begin{cases} 1, & z < 0 \\ 0, & else \end{cases}$$

使用 hinge 损失，$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\max(0, 1 - y_i(w^T x_i + b))$

$$l_{hinge}(z) = \max(0, 1-z)$$

$$l_{log}(z) = \log(1 + \exp(-z))$$

引入松弛变量 $\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \quad s.t.\ y_i(w^T x_i + b) \geq 1 - \xi_i$

$$l_{exp}(z) = \exp(-z)$$

对偶问题 $\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \quad s.t. \sum_{i=1}^{m}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C$

四、神经网络

多层前馈神经网络

激活函数：$sgn(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$

$sigmoid(x) = \frac{1}{1+e^{-x}} \rightarrow y' = y(1-y)$

标准/累积

BP 误差逆传播算法 $o = W_2 h + b_2$，$h = sigmoid(W_1 x + b_1)$ 维度：$l\times 1$，$l\times q$，$q\times 1$，$l\times 1$，$q\times 1$，$q\times d$，$d\times 1$，$q\times 1$ 李航

缓解过拟合 {早停 / 正则化}

tricks：预训练(逐层监督)、权共享、Dropout、ReLU、交叉熵 $-\frac{1}{m}\sum_{i=1}^{m} y_i \log \hat{y}_i$

RBF 径向基函数 单隐层前馈，径向基函数激活，如 $\rho(x, c_i) = e^{-\beta_i\|x - c_i\|^2}$

SOM 自组织特征映射 竞争型无监督 高维数据映射到低维空间，神经元竞争，表示一个样本

CC 级联相关网络 构造性，学习结构

Elman 一种递归神经网络

五、贝叶斯分类器 西瓜书

条件风险 $R(c_i|x)=\sum_{j=1}^{N}\lambda_{ij}P(c_j|x)$ ，λ_{ij}表示将第j类分类到第i类的损失

贝叶斯判定准则 $h^*(x)=\underset{c\in\mathcal{Y}}{\arg\min}\ R(c|x)$

- 判别式模型 直接建模$P(c|x)$ 决策树、神经网络、SVM
- 生成式模型 建模联合概率$P(x,c)$，求$P(c|x)=\frac{P(x,c)}{P(x)}$ 贝叶斯分类器

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

连乘容易下溢

极大似然估计 $\prod_x P(x|c_x)\longrightarrow$ 对数似然 $\sum_x \log P(x|c_x)$

朴素贝叶斯分类器：假定属性相互独立

离散：$P(x_i|c)=\frac{|D_{c,x_i}|}{|D_c|}$

连续：$p(x_i|c)\sim N(\mu_{c,i},\sigma_{c,i}^2)$ ，$p(x_i|c)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

拉普拉斯修正 $\hat{P}(c)=\frac{|D_c|+1}{|D|+N}$ $\hat{P}(x_i|c)=\frac{|D_{c,x_i}|+1}{|D_c|+N_i}$

N表示类别数 N_i表示第i个属性的取值数

usage
- 预计算 → 查表
- 懒惰学习
- 增量学习

半朴素贝叶斯分类器

- ODE 独依赖估计 $p(x|c)=\sum_{i=1}^{d}P(x_i|c,pa_i)$ （pa_i：parent 父属性）
  - SPODE 一个是大家的父亲，称为超父
  - TAN 最强依赖
  - AODE 每个属性都做一次超父，加权平均（集成 SPODE）
- KDE 高阶依赖

贝叶斯网（有向无环图）→ 结构学习：评分函数基于信息论准则评估贝叶斯网与训练数据的契合度

最小描述长度 $s(B|D)=f(\theta)|B|-LL(B|D)$

参数比特数，参数个数，$\sum\log P_B(x_i)$

AIC $f(\theta)=1$

BIC $f(\theta)=\frac{1}{2}\log m$

推断 基于已知属性（证据），推测其他属性
- 吉布斯采样
- 变分推断

EM算法

$LL(\Theta|X,Z)=\ln P(X,Z|\Theta)$ ，Z是隐变量无法直接求解

E步：基于Θ^t对Z算期望Z^t M步：用X和Z^t更新Θ^t

六、集成学习　个体准确且不同

1、Boosting　一类算法，先训练一个learner，再将其做错的加大比重训练下一个learner，最后加权结合

AdaBoost　加性模型　　有效降低偏差，针对泛化性能弱的分类器

2、Bagging　多次自助采样并分别训练，最后投票　　有效降低方差，对易受样本干扰的学习上效用更明显

3、随机森林　bagging + 决策树 + 随机属性选择（先随机再最优）

结合方法
- 简单平均、加权平均
- 绝对多数投票（不过半拒绝预测）
- 相对多数投票、加权投票
- 用学习来集成（初级learner —— 次级learner）

多样性 → 度量

↳ 增强
- 数据样本扰动
- 输入属性扰动
- 输出属性扰动
- 算法参数扰动

七、聚类

性能度量
- 外部指标　与"参考模型"进行比较
- 内部指标　直接考察聚类结果

距离度量
- 性质：非负、同一、对称、（直递）　⟶ 闵可夫斯基距离
- 无序属性：VDM、MinkovDM

原型聚类
- k-means：随机簇中心，划分，更新中心，划分…（EM）
- 高斯混合聚类（GMM）：$P_M(x)=\sum_{i=1}^{k}\alpha_i\, p(x|\mu_i,\Sigma_i)$　随机 α,μ,Σ，算后验，最大似然更新，后验…　E　M　E…
- LVQ原型向量划分

密度聚类　DBSCAN　核心对象、密度直达、密度可达、密度相连

层次聚类　AGNES　一个点一个簇，逐渐合并，可以获得不同级别的聚类结果

八、决策树

每个中间结点寻找一个属性划分

信息熵度量样本集合纯度 $Ent(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k$ ent↓纯度↑

信息增益：对离散属性 $a=\{a^1,a^2,\cdots,a^V\}$ 进行划分 $Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v)$ 对可取值数目多的属性有偏好

增益率：$Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)}$，$IV(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$ 随a取值多而大

C4.5：先选 Gain 较高和一起比，再使用 Gain_ratio

基尼指数 $Gini(D)=\sum_{k=1}^{|\mathcal{Y}|}\sum_{k'\neq k}p_k p_{k'}$ 反映从D随机抽2个，label不一样的概率，Gini↓纯度↑

$$Gini\_index(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}Gini(D^v)$$

结束条件：
- D中类别相同，无需划分
- 属性集为空（D中属性取值相同），无法划分
- D为空，不能划分

输入： 训练集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$；
属性集 $A=\{a_1,a_2,\ldots,a_d\}$.
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** 结束(1)
3: 将 node 标记为 C 类叶结点; **return**
4: **end if**
5: **if** $A=\varnothing$ **OR** D 中样本在 A 上取值相同 **then** 结束(2)
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_*; 划分选取
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** 结束(3)
12: 将分支结点标记为叶结点, 其类别标记为 D 中样本最多的类; **return**
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

分支划分

划分选择对尺寸有较大影响，但对泛化性能影响有限

剪枝是决策树应对过拟合的关键手段

预剪枝 测试时间开销↓、训练时间开销↓、过拟合风险↓、欠拟合↑

后剪枝 测试时间开销↓、训练时间开销↑、过拟合风险↓、欠拟合 __ 、泛化性能一般比预剪枝好

连续值：离散化（二分…）

缺失值：权重划分 { 计算信息增益 / 划分分支

轴平行划分 ⟶ 多变量决策树