CART

优点	特点	衡量指标
可用于分类和回归	构建的二叉树	Gini

算法总结

算法	支持模型	树结构	特征选择	连续值处理	缺失值处理	剪枝	特征属性多次使用
ID3	分类	多叉树	信息增益	✗	✗	✗	✗
C4.5	分类	多叉树	信息增益率	✓	✓	✓	✗
CART	分类、回归	二叉树	Gini、均方差	✓	✓	✓	✓

4. 多变量决策树

主要算法：OC1 —贪心地寻找每个属性的最优权值，在局部优化的基础上再对分类边界进行随机扰动以试图找到更好的边界

第i个神经元的连接权重

w_1 w_2 w_n → θ → y

n个输入信号

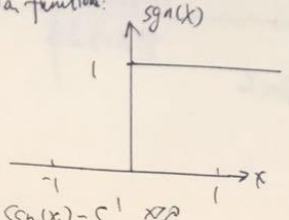
§5 神经网络

输出 $y=f(\sum_{i=1}^{n} w_i x_i - \theta)$

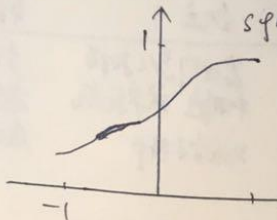
threshold logic unit

1. 感知机(perceptron)：输入层(接收信号) → 输出层(M-P神经元，也称阈值逻辑单元)

activation function:



$$sgn(x)=\begin{cases}1 & x\geq 0\\ 0 & x<0\end{cases}$$



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

$w_i \leftarrow w_i + \Delta w_i$ $\Delta w_i = \eta(y-\hat{y})x_i$

$\eta \in (0,1)$ — learning rate.

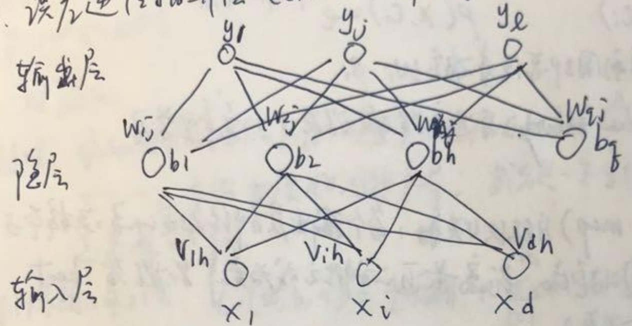
由于只有一层输出神经元，只可对线性可分（与、或、非） 异或（非线性可分）

多层功能神经元：输入层 — hidden layer — 输出层

更一般的：层级结构、神经元之间不存在同层连接，也不存在跨层连接。

→ 多层前馈神经网络 (multi-layer feedforward neural networks)

2. 误差逆传播算法 (error backpropagation, BP)



$\beta_j = \sum_{h=1}^{q} w_{hj} b_h$

$\alpha_h = \sum_{i=1}^{d} v_{ih} x_i$

均方误差 $E_k = \frac{1}{2} \sum_{j=1}^{l} (\hat{y}_j^k - y_j^k)^2$

即基于梯度下降 $\Delta w_{hj} = -\eta \frac{\partial E_k}{\partial w_{hj}}$ η — 学习率

目标：最小化训练集D上的累积误差 $E = \frac{1}{m} \sum_{k=1}^{m} E_k$

- 标准BP：每次更新只针对单例，需更多次迭代，训练集D大时有较好解
- 累积BP：直接针对累计误差最小化，读取D后才更新，更新频率低

防止过拟合：
- early stopping：若D误差降低但验证集V误差升高则停止
- regularization：$E = \lambda \frac{1}{m} \sum_{k=1}^{m} E_k + (1-\lambda) \sum_i w_i^2$

3. 全局最小与局部极小

跳出局部极小、进一步接近全局最小（三种）

1) 多组不同参数值初始化多个神经网络. 按标准方法训练后. 取其中误差最小的网作为最终参数

2) 模拟退火. 每一步以一定概率接受比当前更差的结果.

3) 随机梯度下降

4. 其他常见神经网络

1) RBF (Radial Basis Function) 径向基函数

$$\varphi(x)=\sum_{i=1}^{p} w_i \rho(x, c_i) \qquad \rho(x, c_i)=e^{-\beta_i \|x-c_i\|^2}$$

确定神经元中心 c_i. 再利用BP算法来确定 w_i, β_i

2) ART (adaptive resonance theory) 自适应谐振理论) 竞争型学习

winner-take-all

3) SOM (self-organizing map) 自组织映射: 每个输出层神经元会计算该样本与自身携带的权向量之间的距离. 距离最近的神经元成为竞争获胜者 best matching unit 最佳匹配单元

§6 支持向量机

1. 感知器模型

$$y=\operatorname{sgn}(\theta \cdot x)=\begin{cases} 1, & \theta x>0 \\ -1, & \theta x<0 \end{cases}$$

$y\theta x>0$ 正确分类

$y\theta x<0$ 错误分类

Loss function: 期望使所有错分样本到超平面之距离之和最小

$$L=\sum_{i=1}^{k} \frac{-y^i \theta \cdot x^i}{\|\theta\|_2} \rightarrow L=-\sum_{i=1}^{k} y^i \theta x^i \qquad \frac{\partial L(\theta)}{\partial \theta}=-\sum_{i=1}^{k} y^i x^i$$

< 线性可分 SVM: 让离超平面较近的点尽可能远离这个点 超平面 (separating hyperplane)
非线性可分 SVM

支持向量 (support vector) 离分割超平面最近的那些点

间隔 (margin): 支持向量表现点到超平面的距离