机器学习

1. 任务：预测（回归）、分类。
2. 数据情况：有监督、无监督、半监督
3. 三要素：模型 + 策略 + 算法

感知机：（线性分类模型）

1. 二分类：$y \in \{+1, -1\}$，划分超平面类超平面
2. 思想：导入基于误分类的损失函数，利用随机梯度下降（SGD）极小化损失函数

若 $y_i(w \cdot x_i + b) \le 0$（误分类）
$w \leftarrow w + \eta y_i x_i$
$b \leftarrow b + \eta y_i$

3. $f(x) = sign(wx + b)$.　　$sign(\cdot) = \begin{cases} +1, & x \ge 0 \\ -1, & x < 0 \end{cases}$

4. 损失函数　$\min\limits_{w,b} L = \min\limits_{w,b} -\sum\limits_{x_i \in M} y_i (w \cdot x_i + b)$.

▷ 误分类点到分离超平面的距离：$-\frac{1}{\|w\|} y_i (w \cdot x_i + b) > 0$ 恒成立.

线性可分支持向量机（SVM）

1. 原理：在特征空间上间隔最大的分类器.

目标：间隔最大化

2. $f(x) = sign(wx + b)$.

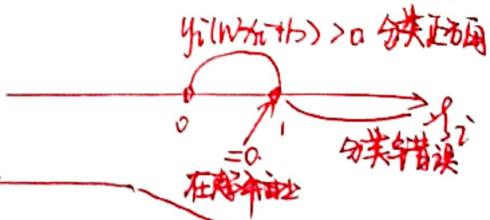
3. 最优化问题目标函数：

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t. \quad y_i(wx_i + b) - 1 \ge 0, \quad i = 1, 2, \cdots, N$$

▷ 唯一存在

当数据存在噪声（特异点：偏离很远的点）.

引入软间隔 $\xi_i \ge 0$.　$y_i(wx_i + b) \ge 1 - \xi_i$



目标函数：

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$s.t. \quad y_i(wx_i + b) \ge 1 - \xi_i, \quad i = 1, 2, \cdots, N, \quad \xi_i > 0$$

4. SVM 通过对偶算法求解（更容易），解决非线性可分问题：引入核函数.

决策树

1. 熵 $H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|}$

结果有 K 类. $|C_k|$ 为属于 C_k 的样本个数.

2. 条件熵 $H(D|A) = \sum_{i=1}^{n} P(A=a_i) H(D|A=a_i)$.

$$= -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}$$

($\frac{|D_i|}{|D|}$: 属于 a_i 的概率; $\frac{|D_{ik}|}{|D_i|}$: 在 a_i 条件下, 属于 C_k 的概率.)

根据 A 的取值 $\{a_1, \cdots, a_n\}$ 将 D 划分成 n 个子集 $D_1, \cdots, D_n$.

对于 D_i 中. $|D_{ik}|$ 为属于类 C_k 的样本个数.

3. 信息增益. $g(D,A) = H(D) - H(D|A)$

ID3 算法: 选择 $g(D,A)$ 最大的作为划分的特征, 对余下的训练集 D_i 继续计算信息增益. (若小于阈值 ε 返回)

4. △ 增益比: $g_A(D,A) = \frac{g(D,A)}{H_A(D)}$.

C4.5 算法: 选择增益比高于平均水平的特征中, 增益比最高的.

其中 $H_A(D) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$

根据 A 的取值将 D 划分成 n 个子集 $D_1, \cdots, D_n$.

5. △ 缺失问题: ① 仅使用无缺失的样例 → 泛数.

② 对于每一个划分特征, 计算 $H(D|A)$ 时仅使用在该特征上无缺失值的样例.

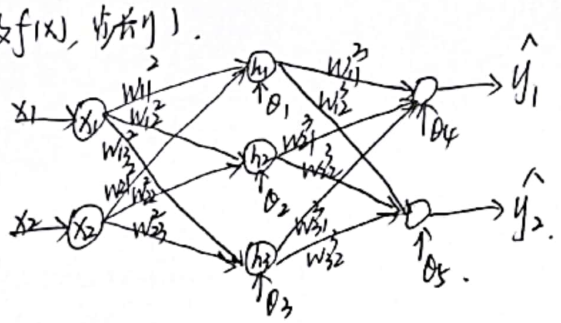
BP算法：　输入层、隐藏层、输出层。

1. 训练参数：权重。（需学习的参数）。
   超参数：网络层数、连接方式、每层节点数、学习步长 η。

2. 激活函数　① sigmoid：$\sigma(z)=\frac{1}{1+e^{-z}}$　$\sigma'(z)=\sigma(z)(1-\sigma(z))$。

   ② ReLU：$f(z)=\max(0,x)$。　$f'(z)=\begin{cases}0, & x<0\\ 1, & x\geq 0\end{cases}$

3. 训练过程。（激活函数 $f(x)$，步长 η）。



(1) 前向传播。

$z_1 = w_{11}^2 x_1 + w_{21}^2 x_2 - \theta_1$　$h_1 = f(z_1)$。

$z_2 = w_{12}^2 x_1 + w_{22}^2 x_2 - \theta_2$　$h_2 = f(z_2)$。

$z_3 = w_{13}^2 x_1 + w_{23}^2 x_2 - \theta_3$　$h_3 = f(z_3)$。

$z_4 = w_{11}^3 h_1 + w_{21}^3 h_2 + w_{31}^3 h_3 - \theta_4$：　$\hat{y}_1 = f(z_4)$。

$z_5 = w_{12}^3 h_1 + w_{22}^3 h_2 + w_{32}^3 h_3 - \theta_5$。　$\hat{y}_2 = f(z_5)$。

(2) 误差反向传播。（链式法则）。

$\hat{w_i} \leftarrow w_i - \eta \frac{\partial E}{\partial w_i}$。（梯度下降）。

均方误差 $E = \frac{1}{2}\sum_{j=1}^{l}(\hat{y}_j - y_j)^2$。（$l$ 为输出层节点数）。

$\frac{\partial E}{\partial w_{11}^3} = \frac{\partial E}{\partial \hat{y}_1}\cdot\frac{\partial \hat{y}_1}{\partial f}\cdot\frac{df}{dz_4}\cdot\frac{\partial z_4}{\partial w_{11}^3}$。　$\hat{w}_{11}^3 = w_{11}^3 - \eta\frac{\partial E}{\partial w_{11}^3}$　（w^3 同理）。

$\frac{\partial E}{\partial w_{11}^2} = \frac{\partial E}{\partial \hat{y}_1}\cdot\frac{\partial \hat{y}_1}{\partial f}\cdot\frac{df}{dz_4}\cdot\frac{\partial z_4}{\partial h_1}\cdot\frac{\partial h_1}{\partial f}\cdot\frac{df}{dz_1}\cdot\frac{\partial z_1}{\partial w_{11}^2}$

（两部分）：

$+\ \frac{\partial E}{\partial \hat{y}_2}\cdot\frac{\partial \hat{y}_2}{\partial f}\cdot\frac{df}{dz_5}\cdot\frac{\partial z_5}{\partial h_1}\cdot\frac{\partial h_1}{\partial f}\cdot\frac{df}{dz_1}\cdot\frac{\partial z_1}{\partial w_{11}^2}$

$\hat{w}_{11}^2 = w_{11}^2 - \eta\frac{\partial E}{\partial w_{11}^2}$。

（w^2 同理）。