集成学习 (ensemble learning)

多个弱学习器，组合产生最终的结果，具有较好的泛化能力

考虑二分类，$y \in \{-1, 1\}$，基分类器的错误率 ε，$P(h_i(x) \neq g(x)) = \varepsilon$

投票法：$\mathrm{sign}\left(\sum_{i=1}^{T} h_i(x)\right) = H(x)$

$$P(H(x) \neq g(x)) = \sum_{k=0}^{[T/2]} \binom{T}{k} (1-\varepsilon)^k \varepsilon^{T-k} \underset{\text{Hoeffding}}{\leq} \exp\left\{-\frac{1}{2} T (1-2\varepsilon)^2\right\}$$

①与T有关

② $\varepsilon < \frac{1}{2}$

提升方法 (Boosting)

1. 提升方法与Adaboost算法

Adaboost 提高前一段被错误分类的权值，降低正确分类的样本的权值

输入：$\{(x_i, y_i)\}$，$y_i \in \{-1, 1\}$，$G(x) = \mathrm{sign}\left(\sum_m \alpha_m G_m(x)\right)$

(1) 初始的权值分布：$D_1 = (w_{11}, w_{12}, \cdots, w_{1n})^T \quad w_{1i} = \frac{1}{N}$

(2) 对 $m = 1, 2, \cdots, M$

(a) 使用有权值分布 D_m 的训练器，得到 $G_m(x)$（取值 $\{-1, 1\}$）

(b) 计算 $G_m(x)$ 在训练集上的分类误差率 $e_m = \sum_i w_{mi} I\{G_m(x_i) \neq y_i\}$

(c) $\alpha_m = \frac{1}{2} \log\left\{\frac{1-e_m}{e_m}\right\}$

(d) 更新权值分布：

$$w_{m+1,i} = \frac{w_{m,i} \exp(-\alpha_m y_i G_m(x_i))}{Z_m}$$

其中，$Z_m = \sum_i w_{m,i} \exp(-\alpha_m y_i G_m(x_i))$

注：(1) e_m 越小，α_m 越大（越准越大）

(2) w_{mi} $\quad y_i = G_m(x_i)$ 正确分类：$\exp(-\alpha_m)$

错误分类：$\exp(\alpha_m)$

2. Adaboost的解读

(1) 前向分步算法

可加模型 $f(x)=\sum_{m=1}^{M}\beta_m b(x;r_m)$

给定损失函数 $\min\limits_{\{(\beta_m,r_m)\ 1\le m\le M\}} L(y_i,\sum_{m=1}^{M}\beta_m b(x,r_m))$

前向分步算法

(1) 初始化 $f_0(x)=0$

(2) 对 $m=1,2,\cdots,M$

极小化

$$(\beta_m,r_m)=\arg\min_{\beta,r}\sum_{i=1}^{N}L(y_i,f_{m-1}(x_i)+\beta b(x_i,r))$$

$$f_m(x)=f_{m-1}(x)+\beta_m b(x,r_m)$$

定理: Adaboost是前向分步算法的特例，损失函数 → 指数损失函数

$$L(y,f(x))=\exp(-yf(x))$$

证明: 假设经过 n-1 轮迭代 已经得到 $f_{m-1}(x)$　　$f_{m-1}(x)=\sum_{i=1}^{m-1}\alpha_i G_i(x)$

$$(\alpha_m,G_m(x))=\arg\min_{\alpha,G}\sum_i\exp(-y_i(f_{m-1}(x_i)+\alpha G(x_i)))$$

$$=\arg\min_{\alpha,G}\sum_i\overline{w}_{mi}\exp(-\alpha G(x_i))$$

先求 $G_m(x)$，对于任意 $\alpha>0$，$G_m^*(x)$ 应满足

$$G_m(x)=\arg\min_{G}\sum_i\overline{w}_{mi}\,I(y_i\neq G(x_i))$$

误分类越少越好

这是因为 $\sum_i\overline{w}_{mi}\exp(-y_i\cdot\alpha\cdot G(x_i))=\sum_{y_i=G(x_i)}\overline{w}_{mi}\exp(-\alpha)+\sum_{y_i\neq G(x_i)}\overline{w}_{mi}\exp(\alpha)$

再求 α_m，

对 $\sum_i \overline{w}_{mi}\exp(-y_i\alpha G_m(x_i))$ 求导，令导数 $=0$

$$-\sum_{y_i=G_m(x_i)} \overline{w}_{mi}\exp(-\alpha)+\sum_{y_i\neq G_m(x_i)} \overline{w}_{mi}\exp(\alpha)=0$$

$$\frac{\sum w_{mi} I(y_i=G_m(x_i))}{\sum w_{mi} I(y_i\neq G_m(x_i))}=\frac{1-e_m}{e_m}=e^{2\alpha}$$

则 $\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m}$

其中 $e_m=\frac{\sum_i \overline{w}_{mi} I(y_i\neq G_m(x_i))}{\sum_i \overline{w}_{mi}}=\sum_i w_{mi} I(y_i\neq G_m(x_i))$

$\overline{w}_{m+1,i}=\overline{w}_{mi}\exp\{-y_i\alpha_m G_m(x)\}$ 与 Adaboost 更新等价

3. Adaboost 的训练误差分析

定理：Adaboost 的训练误差界

$$\frac{1}{N}\sum_i I(G(x_i)\neq y_i)\leq\frac{1}{N}\sum_i\exp(-y_i f(x_i))=\prod_{m=1}^{M} Z_m$$

$Z_m=\sum_i w_{mi}\exp(-\alpha_m y_i G_m(x_i))$，$w_{mi}=\exp(-y_i f_{m-1}(x_i))/\sum_i\exp(-y_i f_{m-1}(x_i))$

解读：可以在每一轮找 $G_m(x)$ 使 Z_m 最小，从而使训练误差下降最快

proof：(1) ① $G(x_i)\neq y_i$ y_i 与 $f(x_i)$ 不同号，$y_i f(x_i)<0$，$\exp(-y_i f(x_i))>1$

② $G(x_i)=y_i$ y_i 与 $f(x_i)$ 同号，$y_i f(x_i)>0$，$0<\exp(-y_i f(x_i))<1$

$$\frac{1}{N}\sum I(G(x_i)\neq y)\leq\frac{1}{N}\sum_{G(x_i)\neq y_i}\exp(-y_i f(x_i))$$

$$\leq\frac{1}{N}\sum\exp(-y_i f(x_i))$$

(2) Adaboost

$$\frac{w_{mi}\cdot \exp(-\alpha_m y_i G_m(x_i))}{Z_m} = w_{m+1,i}$$

$$w_{mi}\exp(-\alpha_m y_i G_m(x_i)) = w_{m+1,i} Z_m$$

$$\frac{1}{N}\sum_i \exp(-y_i f(x_i)) = \frac{1}{N}\sum_i \exp\left(-y_i \sum_m \alpha_m G_m(x_i)\right)$$

$$= \sum w_{1i} \exp\left(-y_i \sum_m \alpha_m G_m(x_i)\right)$$

$$= \sum w_{1i} \exp(-y_i \alpha_1 G_1(x_i)) \prod_{m=2}^{M} \exp(-y_i \alpha_m G_m(x_i))$$

$$= \sum_i Z_1 w_{2i} \prod_{m=2}^{M} \exp(-y_i \alpha_m G_m(x_i))$$

$$= Z_1 \sum_i w_{2i} \prod_{m=2}^{M} \exp(-y_i \alpha_m G_m(x_i))$$

$$= \prod_{m=1}^{M} Z_m$$

· 定理(二分类的 Adaboost 上界)

$$\prod_{m=1}^{M} Z_m = \prod_{m=1}^{M}\left\{2\sqrt{e_m(1-e_m)}\right\} = \prod_{m=1}^{M}\sqrt{1-4\gamma_m^2} \leq \exp\left(-2\sum_{m=1}^{M}\gamma_m^2\right),$$

其中 $\gamma_m = \frac{1}{2} - e_m$, $e_m = \sum_i w_{mi} I(y_i \neq G_m(x_i))$

解读:假设 $\gamma_m > \gamma$, $\exp\left(-2\sum_{m=1}^{M}\gamma_m^2\right) \leq \exp(-2M\gamma^2)$

proof:

$$Z_m = \sum_i w_{mi} \exp(-\alpha_m y_i G_m(x_i))$$

$$= \sum_{y_i = G(x_i)} w_{mi}\exp(-\alpha_m) + \sum_{y_i \neq G(x_i)} w_{mi}\exp(\alpha_m)$$

$$= (1-e_m)\exp(-\alpha_m) + e_m \exp(\alpha_m),$$ 其中 $\alpha_m = \frac{1}{2}\log\frac{1-e_m}{e_m}$

$$= 2\sqrt{(1-e_m)e_m} = \sqrt{1-4\gamma_m^2} \leq \exp(-2\gamma_m)$$ 比较 $\sqrt{1-2x^2}$ 与 $\exp(-x)$

提升树(Boosting Tree)

- 以决策树为基的提升方法称为提升树，提升树可以表示成决策树的加法模型 $f_M(x)=\sum_{m=1}^{M} T(x;\theta_m)$

1. 提升树算法：

第m步模型为：$f_m(x)=\sum_{m-1} T(x;\theta_m)+T(x;\theta_m)$

通过经验风险最小化求解参数 θ_m

$$\theta=\underset{\theta_m}{\arg\min}\, L(y_i, f_{m-1}(x_i)+T(x;\theta_m))$$

回归问题：平方误差 $L(y, f_{m-1}(x)+T(x;\theta_m))=(y-f_{m-1}(x)-T(x;\theta_m))^2$

相当于对m-1步的模型残差拟合

2. 梯度提升(Freidman)

对一般的损失函数，利用损失函数负梯度 $-\left\{\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right\}_{f(x)=f_{m-1}(x)}$

- 输出回归树

(1) 初始化 $f_0(x)=\underset{c}{\arg\min}\sum_{i=1}^{N} L(y_i;c)$

(2) 对 $m=1,2,\cdots,M$，

(a) 对 $i=1,2,\cdots,N$， $r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x_i)}$

(b) 对 r_{mi} 拟合一个回归树，得到第m棵树的叶节点的区域 R_{mj}，$j=1,\cdots,J$

(c) 对 $j=1,2,\cdots,J$，$c_{mj}=\underset{c}{\arg\min}\sum_{x_i\in R_{mj}} L(y_i, f_{m-1}(x_i)+c)$

(3) $\hat{f}(x)=\sum_{m=1}^{M}\sum_{j=1}^{J} c_{mj} I(x\in R_{mj})$

Bagging算法 (Breiman, 1996)

- 希望：个体学习器相对独立
- 解决方法：重新采样数据 (Bootstrap sampling)
- 步骤：① 对数据进行T次Bootstrap重采样，每次采样n个训练样本
  ② 分类问题 → 投票法，回归问题 → 求平均
- 泛化误差："包外估计" (out-of-bag estimate) 对于每棵树，在"包内"训练在"包外"预测，泛化误差使用"包外"误差率

随机森林 (Random Forest, Breiman 2001)

- 对特征也随机选择，每次选L个特征.

评估变量重要性

- 方法1 (用训练数据)：对每一棵树，变量重要性可用在该变量分裂前后的评价指标 (e.g. 基本指标) 对所有树取平均
- 方法2 (用包外数据)：评价 X_j 的重要性，对包外数据的第 X_j 列进行干扰 (随机变换顺序)，计算预测率的降低。对所有树取平均