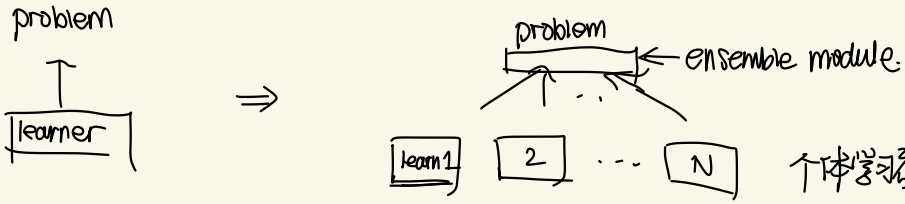
7 Boosting

集成学习(Ensemble Learning): using multiple learners to solve the problem.



个体学习器(base learner)

"弱级"

e.g. A	✓	✓	×	→	✓
B	✓	×	✓	→	✓
C	×	✓	✓	→	✓

"好而不同"

diversity

多样性度量: "holy grail problem of ensemble learning"

To Win? Ensemble!

AdaBoost

Boosting

产生 Base Learner 的方式

- Sequential (串行): $h_1 \xrightarrow{\lambda_1} h_2 \xrightarrow{\lambda_2} h_3 \to \cdots$
- Parallel: (并行) $h_1 \quad h_2 \quad \cdots \quad h_N$

Bagging

e.g. Random Forest 随机森林

def (Weak Learner) concept class $\mathcal{C}$ weakly PAC-learnable.

弱学习器 if $\exists \mathcal{A}, \gamma>0\ \forall c\in\mathcal{C}, \forall \delta\ \forall \mathcal{D}$

$$\boxed{\Pr_{S\sim\mathcal{D}^m}\left(R(h_S)\le \frac{1}{2}-\gamma\right)\ge 1-\delta}$$

if $m \ge \mathrm{poly}\left(\frac{1}{\delta}, n, \mathrm{size}(c)\right)$

ensembling

Strong learner

结合策略

- 平均 $f(x)=\sum_{i=1}^{N} w_i h_i(x)$ $\left(\sum_{i=1}^{N} w_i = 1\right)$
- 投票 $f(x)=\mathrm{sign}\left(\sum_{i=1}^{N} h_i(x)\right)$
- 学习 $f(x)=f'(h_1(x),\cdots,h_N(x))$, f' 是学习器 ("次级")

meta-learner

Stacking

西瓜书 Chapter 8.

一周完成

AdaBoost($S = ((x_1, y_1), \ldots, (x_m, y_m))$)

for $i \leftarrow 1$ **to** m **do**

  $\mathcal{D}_1(i) \leftarrow \frac{1}{m}$ (均匀分布)

for $t \leftarrow 1$ **to** T **do**

  $h_t \leftarrow$ base classifier in $\mathcal{H}$ with small error $\epsilon_t = \mathbb{P}_{i\sim\mathcal{D}_t}[h_t(x_i) \neq y_i]$

  $\alpha_t \leftarrow \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$ 权重因子 (logits)

  $Z_t \leftarrow 2[\epsilon_t(1-\epsilon_t)]^{\frac{1}{2}}$ ▷ normalization factor

  for $i \leftarrow 1$ **to** m **do**

    $\mathcal{D}_{t+1}(i) \leftarrow \frac{\mathcal{D}_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ 调整分布

$f \leftarrow \sum_{t=1}^{T}\alpha_t h_t$ (ensembling (结合))

return f

$$\boxed{h_t \in \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ \epsilon_t} \quad ☆$$

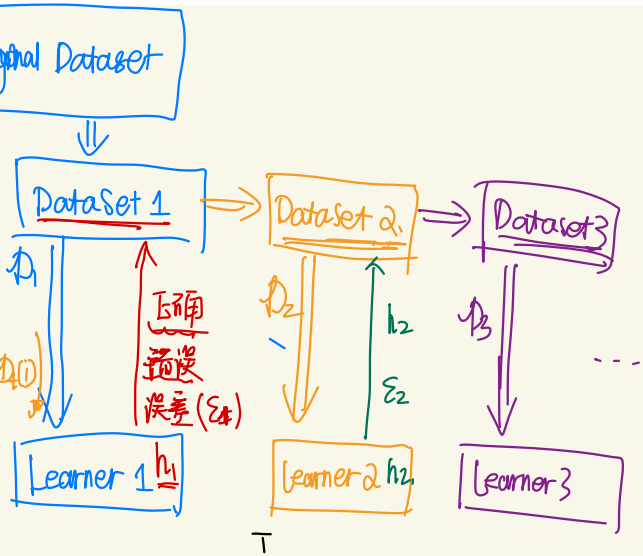
$$\epsilon_t = \mathbb{P}_{i\sim\mathcal{D}_t}[h_t(x_i)\neq y_i] = \mathbb{E}_{i\sim\mathcal{D}_t}[\mathbb{1}_{h(x_i)\neq y_i}] = \sum_{i=1}^{m}\mathcal{D}_t(i)\cdot\mathbb{1}_{h(x_i)\neq y_i}$$

核心思想：调整分布 (distribution)

$$\begin{cases}\epsilon_t > \frac{1}{2} \Rightarrow \alpha_t < 0\\ \epsilon_t < \frac{1}{2} \Rightarrow \alpha_t > 0\end{cases}$$

$$\begin{cases}\text{正确 } h(x_i)\cdot y_i = +1 \Rightarrow \text{学得不错} \Rightarrow \text{更轻松} \quad \mathcal{D}_t(i)\downarrow\\ \text{错误 } h(x_i)\cdot y_i = -1 \Rightarrow \text{学得不好} \Rightarrow \text{更重视} \quad \mathcal{D}_t(i)\uparrow\end{cases}$$

(本质上调高)



$$f(x) = \sum_{t=1}^{T}\alpha_t h_t(x)$$

AdaBoost

$$\begin{cases}\text{选择怎样的学习器？ } h \to \epsilon_t \text{ 方向}\\ \text{选择怎样的权重？ } \alpha \to \eta \text{ 步长}\end{cases}$$

(generalization (extension)：

$$\begin{cases}[-1,1], \ \mathbb{R}\text{上的闭集}\\ h_t\text{也不一定要是ERM得到的结果，只需与分布有关}\end{cases}$$

Thm (AdaBoost 降低经验误差) 随轮数 t 指数下降

$$\hat{R}_S(f) \le \exp\left(-2\sum_{t=1}^{T}\left(\tfrac{1}{2}-\varepsilon_t\right)^2\right) \quad ☆$$

$$\forall t\in[T],\ \underbrace{\gamma}_{\text{"edge"}} \le \tfrac{1}{2}-\varepsilon_t \Rightarrow \hat{R}_S(f) \le \exp(-2\gamma^2 T)$$

Proof: Proposition: $f_t = \sum_{s=1}^{t}\alpha_s h_s \quad f_T = f$

$$\frac{D_{t+1}}{D_t} = \frac{\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \Rightarrow \boxed{D_{t+1}(i) = \frac{e^{-y_i f_t(x_i)}}{m\cdot\prod_{s=1}^{t} Z_s}} \quad (*)$$

"telescopic"

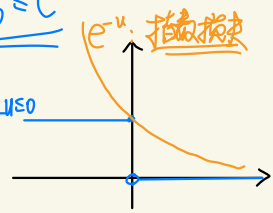
$$\hat{R}_S(f) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{y_i f(x_i)\le 0} \le \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f(x_i)} \quad ☆$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(m\cdot\prod_{t=1}^{T} Z_t\right)\cdot D_{T+1}(i)$$

$$= \prod_{t=1}^{T} Z_t \cdot \sum_{i=1}^{m} D_{T+1}(i) = \prod_{t=1}^{T} Z_t.$$

$\mathbb{1}_{u\le 0} \le e^{-u}$

0/1损失 $\mathbb{1}_{u\le 0}$



$$Z_t = \sum_{i=1}^{m} D_t(i)\cdot e^{-\alpha_t y_i h_t(x_i)} = \underbrace{\sum_{i:y_i h_t(x_i)=+1} D_t(i)\cdot e^{-\alpha_t}}_{\text{正确}} + \underbrace{\sum_{i:y_i h_t(x_i)=-1} D_t(i) e^{\alpha_t}}_{\text{错误}}$$

$$= (1-\varepsilon_t)e^{-\alpha_t} + \varepsilon_t\cdot e^{\alpha_t} = \varphi(\alpha_t)$$

$$= (1-\varepsilon_t)\cdot\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} + \varepsilon_t\cdot\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

$$\hat{R}_S(f) = \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = \prod_{t=1}^{T}\sqrt{1-4\left(\tfrac{1}{2}-\varepsilon_t\right)^2} \le \prod_{t=1}^{T}\exp\left(-2\left(\tfrac{1}{2}-\varepsilon_t\right)^2\right)$$

$$\Rightarrow \hat{R}_S(f) \le \exp\left(-2\sum_{t=1}^{T}\left(\tfrac{1}{2}-\varepsilon_t\right)^2\right) \quad \square \qquad (1-x\le e^{-x})$$

AdaBoost: Adaptive Boosting $\quad \varphi(\alpha) = (1-\varepsilon_t)e^{-\alpha} + \varepsilon_t e^{\alpha}$

① α_t minimize $\varphi(\alpha)$

$$\varphi'(\alpha) = -(1-\varepsilon_t)e^{-\alpha} + \varepsilon_t e^{\alpha} = 0 \Rightarrow \alpha = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$$

$Z_t = \varphi(\alpha_t)$ ↓ $\hat{R}_S(f) \le \prod_{t=1}^{T} Z_t$ α_t minimize the upper bound of empirical error

② $(1-\varepsilon_t)e^{-\alpha_t} = \varepsilon_t e^{\alpha_t}$

正确样本　　错误样本

AdaBoost 给正确、错误样本的权重(概率质量)相同

Why? $\varepsilon_t < \frac{1}{2}$ ⇒ 错误样本更少 ⇒ 更受重视！

ensemble accuracy > base-learner's ~

(empirical risk) <

---

a weak learning algorithm $\xrightarrow{?}$ a strong learning algorithm.

坐标下降 (coordinate descent)：非梯度优化　$f(x)$, $x \in \mathbb{R}^n$.

每次迭代中沿着一个坐标的方向进行线性搜索 $x_k \pm \eta \bar{x}_k$　直至找到 local minima

k

k: 方向

η: 步长

假设 $\mathcal{H}$ 有限　$\mathcal{H} = \{h_1, \cdots, h_N\}$

$f = \sum_{j=1}^{N} \bar{\alpha}_j h_j$　$\bar{\alpha}_j \ge 0$.　Sample Set $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$

Coordinate Descent: $F(\bar{\alpha}) = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f(x_i)} = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_j h_j(x_i)}$ Δ

objective

$\bar{\alpha} = (\bar{\alpha}_1, \cdots, \bar{\alpha}_N)$

指数损失 e^{-u} 是 0/1 损失 $\mathbb{1}_{u \le 0}$ 的上界

$\hat{R}_S(f) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y_i f(x_i) \le 0} \le \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f(x_i)} = F(\bar{\alpha})$

F: < convex, differentiable

我们下面尝试证明：F 是 AdaBoost 算法优化的目标函数.

$\bar{\alpha}_0 = (0, \cdots, 0)$

$e_k = (0, \cdots, 1, \cdots, 0)$

At each round $t \in [T]$: direction e_k ($\bar{\alpha}$ k^{th} coordinate); step size η

$\boxed{\bar{\alpha}_t = \bar{\alpha}_{t-1} + \eta \cdot e_k}$ → ensemble function $\bar{g}_t$.

$\bar{g}_t = \sum_{j=1}^{N} \bar{\alpha}_{t,j} h_j$

$= \sum_{j=1}^{N} (\bar{\alpha}_{t-1,j} + \eta \cdot \bar{e}_{k,j}) h_j = \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j + \eta \cdot h_k = \bar{g}_{t-1} + \eta \cdot h_k$

$h_k \cdot \eta$ < AdaBoost / Coordinate Descent

Proof by Induction. $\le t-1$ 时 h_k 和 η 相同. 考虑第 k 轮的情况

Coordinate Descent
- 方向 e_k : ① $|F'|$ 最大的方向 ✓
- 步长 η : ② minimize $F(\bar{\alpha}_t) = F(\bar{\alpha}_{t-1} + \eta \cdot e_k)$

① distribution: $\bar{D}_t(i) = \dfrac{e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j(x_i)}}{\bar{Z}_t} = \dfrac{e^{-y_i \bar{g}_{t-1}(x_i)}}{\bar{Z}_t} = \dfrac{e^{-y_i f_{t-1}(x_i)}}{\bar{Z}_t} = D_t(i)$

expected error: $\bar{\varepsilon}_{t,j} = \mathbb{E}_{i \sim \bar{D}_t}\left[\mathbb{1}_{y_i h_j(x_i) \le 0} \right]$

$$F'(\bar{\alpha}_{t-1})\Big|_{e_k} = \lim_{\eta \to 0} \frac{F(\bar{\alpha}_{t-1} + \eta \cdot e_k) - F(\bar{\alpha}_{t-1})}{\eta}$$

$F(\bar{\alpha}_{t-1} + \eta e_k) = \frac{1}{m} \sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j(x_i) - \eta y_i h_k(x_i)}$

$F(\bar{\alpha}_{t-1}) = \frac{1}{m} \sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j(x_i)}$

$$= \lim_{\eta \to 0} \frac{\frac{1}{m} \sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j(x_i)} \left(e^{-\eta y_i h_k(x_i)} - 1 \right)}{\eta}$$

$$= \frac{1}{m} \sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j(x_i)} \cdot \lim_{\eta \to 0} \frac{e^{-\eta y_i h_k(x_i)}}{\eta} = \lim_{\eta \to 0} -y_i h_k(x_i) \underbrace{e^{-\eta y_i h_k(x_i)}}_{1} = -y_i h_k(x_i)$$

$$= -\frac{1}{m} \sum_{i=1}^{m} y_i h_k(x_i) \cdot e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j(x_i)}$$

定义

$$= -\frac{1}{m} \sum_{i=1}^{m} y_i h_k(x_i) \cdot \bar{D}_t(i) \bar{Z}_t$$

$$= -\frac{\bar{Z}_t}{m} \cdot \left(\sum_{i: y_i h_k(x_i) = +1} \bar{D}_t(i) - \sum_{i: y_i h_k(x_i) = -1} \bar{D}_t(i) \right) = -\frac{\bar{Z}_t}{m} \left((1 - \bar{\varepsilon}_{t,k}) - \bar{\varepsilon}_{t,k} \right) = (2\bar{\varepsilon}_{t,k} - 1) \frac{\bar{Z}_t}{m}$$

与 k 无关

找的方向 e_k: $|2\bar{\varepsilon}_{t,k} - 1|$ ↑ $\bar{\varepsilon}_{t,k}$ 弱学习器 $(< \frac{1}{2})$

$\Rightarrow$ $\bar{\varepsilon}_{t,k}$ 越小越好!

第 t 步的方向

$$e_k: \min_{h_k \in H} \bar{\varepsilon}_{t,k} = \min_{h_k \in H} \mathbb{E}_{i \sim \bar{D}_t}\left[\mathbb{1}_{y_i h_k(x_i) \le 0} \right]$$

与 AdaBoost 不谋而合 !!! ☆

② 步长: $\eta = \arg\min_{\eta} F(\bar{\alpha}_{t-1} + \eta e_k)$

Convex $\Rightarrow \dfrac{\partial F(\bar{\alpha}_{t-1} + \eta e_k)}{\partial \eta} = 0$

$$\frac{\partial F(\bar{\alpha}_{t-1}+\eta e_k)}{\partial \eta} = \frac{-1}{m}\sum_{i=1}^{m} y_i h_k(x_i)\cdot \underline{\underline{e^{-y_i\sum_{j=1}^{t}\bar{\alpha}_{t-1,j}h_j(x_i)-\eta y_i h_k(x_i)}}} = 0$$

$$\Leftrightarrow \sum_{i=1}^{m} y_i h_k(x_i)\, e^{-\eta y_i h_k(x_i)}\cdot \bar{D}_t(i)\cdot \underline{\underline{\bar{Z}_t}} = 0$$

$$\Leftrightarrow \sum_{i=1}^{m} \underline{y_i h_k(x_i)}\cdot \bar{D}_t(i)\cdot e^{-\eta y_i h_k(x_i)} = 0$$

$$\Leftrightarrow \underbrace{(1-\bar{\varepsilon}_{t,k})\cdot e^{-\eta}}_{\text{正确}} - \underbrace{\bar{\varepsilon}_{t,k}\cdot e^{\eta}}_{\text{错误}} = 0 \Leftrightarrow \eta = \frac{1}{2}\log\frac{1-\bar{\varepsilon}_{t,k}}{\bar{\varepsilon}_{t,k}}$$

再一次与AdaBoost不谋而合!

$f(x)=\sum_i \alpha_i h_i(x)$ $\begin{cases} \text{学习器 } h_i \Leftrightarrow \text{方向 } e_k \\ \text{权重 } \alpha_i \Leftrightarrow \text{步长 } \eta \end{cases}$

分布函数 D

AdaBoost $\Leftrightarrow$ Coordinate Descent (坐标下降) ☆

补充: AdaBoost 扩展 loss function

0/1 损失 $\mathbb{1}_{u\le 0}$

指数损失 e^{-u} $\Leftrightarrow$ AdaBoost

hinge 损失 $\max(0, 1-x)$

△ logistic 损失 $\log_2(1+e^{-x})$ $\Leftrightarrow$ logistic regression

In Practice,

base-learner weak

决策树 (Decision Tree) Chapter 4. 西瓜书

$\Downarrow$

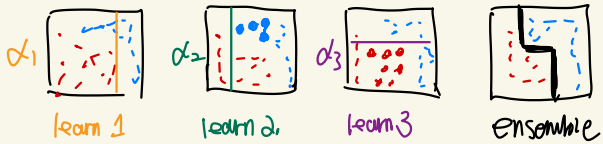
决策树桩 (stump) → decision boundary (threshold)

排序: $O(m\log m) \to O(mN\log m)$

m+1 种可能的 threshold ε_t

优化选择: $O(m) \to O(mNT)$

$\Rightarrow O(mN\log m + mNT)$



(weighted sum)

注意: stump is not a weak learner. $(\Pr(\varepsilon \le \frac{1}{2}-\gamma) \ge 1-\delta)$ "XOR"

Theoretical Results

AdaBoost $\mathcal{F}_T = \left\{ \mathrm{sgn}\left(\underbrace{\sum_{i=1}^{T} \alpha_t h_t}_{\text{linear combination}} \right) : \alpha_t \geq 0,\ h_t \in \mathcal{H},\ t \in [T] \right\}$

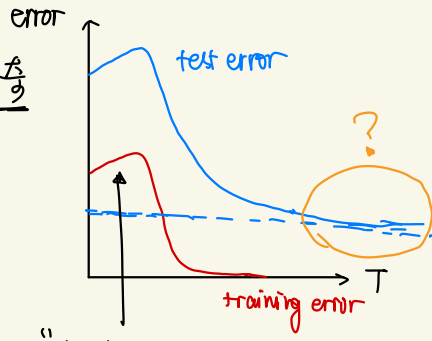
$\mathrm{VCdim}(\mathcal{H}) = d$

$\mathrm{VCdim}(\mathcal{F}_T) \leq 2(d+1)(T+1) \cdot \lg((T+1)e) = O(dT \log T)$

↳ 衡量复杂性 大：复杂程度，更有可能过拟合

Margin-Based Analysis

神奇



"走成率越大越好"

confidence margin : $y \cdot f(x)$ (↑ confidence ↑)

$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) = \vec{\alpha} \cdot \vec{h}$

weight (w) feature vector (Φ)

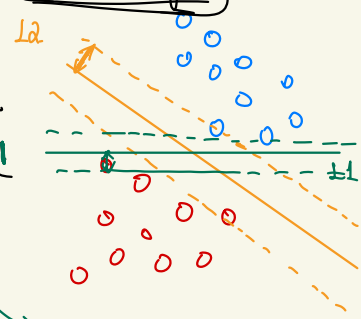
L_2-geometric margin: $\rho_2(x) = \dfrac{|w \cdot x + b|}{\|w\|_2}$

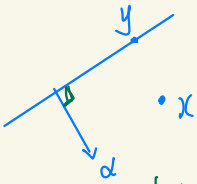
L_1-geometric margin: $\rho_f(x) = \dfrac{|f(x)|}{\|\alpha\|_1} = \dfrac{\left|\sum_{t=1}^{T} \alpha_t h_t(x)\right|}{\|\alpha\|_1} = \dfrac{|\vec{\alpha} \cdot \vec{h}|}{\|\vec{\alpha}\|_1}$

$|f(x)| \to y f(x)$

L_1-margin: $\rho_f = \min_{i \in [m]} \rho_f(x_i) = \min_{i \in [m]} \dfrac{|\vec{\alpha} \cdot \vec{h}(x_i)|}{\|\vec{\alpha}\|_1}$

maximum margin





$\vec{\alpha} \parallel \overrightarrow{xy} \Rightarrow |\vec{\alpha} \cdot \overrightarrow{xy}| = \|\vec{\alpha}\|_2 \underbrace{\|\overrightarrow{xy}\|_2}_{\rho_2(x)}$

distance: ∞-norm

$\max_i |x_i - y_i|$ 某一维度

$|\vec{\alpha} \cdot \vec{x}| = |\vec{\alpha} \cdot \overrightarrow{xy}| = \|\vec{\alpha}\|_p \cdot \|\overrightarrow{xy}\|_q$ $\left(\frac{1}{p} + \frac{1}{q} = 1 : \text{conjugate Hölder inequality}\right)$

$\left(\sum_{i=1}^{m} |x_i - y_i|^q\right)^{\frac{1}{q}}$ q-norm distance

SVM: L2-margin maximization

AdaBoost: L1-margin maximization ?

$$\bar{f} = \frac{f}{\sum_{t=1}^{T} \alpha_t} = \frac{f}{\|\alpha\|_1} \quad \text{(normalize)}$$

confidence margin: $y \cdot \bar{f}(x) = \frac{y f(x)}{\|\alpha\|_1} = \rho_f(x)$

L1-geometric-margin

$\alpha_t \geq 0$

$\rho_f(x)$ is linear combination of base learners $h_t(x)$

$h_t(x) \in [-1, +1] \Rightarrow \rho_f(x) \in [-1, +1]$

凸包

$$\text{Convex-hull}(\mathcal{H}) = \left\{ \sum_{k=1}^{p} \mu_k h_k : p \geq 1,\ \forall k \in [p],\ \mu_k \geq 0,\ h_k \in \mathcal{H},\ \sum_{k=1}^{p} \mu_k \leq 1 \right\}$$

Lemma $\forall$ Sample set S, $\hat{R}_S(\text{Convex-hull}(\mathcal{H})) = \hat{R}_S(\mathcal{H})$

$\rho_f(x)$

Proof: $\hat{R}_S(\text{Convex-hull}(\mathcal{H})) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{\substack{h_1, \cdots, h_p \in \mathcal{H} \\ \mu \geq 0,\ \|\mu\|_1 \leq 1}} \sum_{i=1}^{m} \sigma_i \sum_{k=1}^{p} \mu_k h_k(x_i) \right]$

$$= \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{h_1, \cdots, h_p \in \mathcal{H}} \underbrace{\sup_{\substack{\mu \geq 0 \\ \|\mu\|_1 \leq 1}} \sum_{k=1}^{p} \mu_k \cdot \sum_{i=1}^{m} \sigma_i h_k(x_i)}_{\perp 1} \right]$$

$$= \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{h_1, \cdots, h_p \in \mathcal{H}} \underbrace{\max_{k \in [p]} \sum_{i=1}^{m} \sigma_i h_k(x_i)}_{\perp \infty} \right]$$

placing all the weight to the largest term

$\mu_k = (0, \cdots, 1, \cdots, 0)$

$$= \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i) \right]$$

$$= \hat{R}_S(\mathcal{H}) \qquad \square$$

L1 norm
L∞ norm
> dual norm

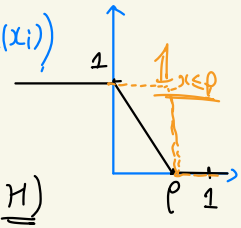
Dual Norm:

$$\|y\|_* = \sup_{\|x\| \leq 1} |\langle y, x \rangle|$$

(L_p & L_q dual if $\frac{1}{p} + \frac{1}{q} = 1$)

Cor (Ensemble Rademacher margin bound)

$$\hat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(y_i h(x_i))$$



fix $\rho>0$. $\forall \delta>0$. w.p. $\geq 1-\delta$.

$$R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2m}}$$

$h \in$ Convex-hull$(\mathcal{H})$

$$R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

ρ-margin-loss

$$\mathcal{R}_m(\mathcal{H}) \le \sqrt{\frac{2\log \Pi_{\mathcal{H}}(m)}{m}}. \quad \Pi_{\mathcal{H}}(m) \le \left(\frac{em}{d}\right)^d$$

Cor (Ensemble VC-dimension margin bound)

$$R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \qquad \text{w.p.} \geq 1-\delta. \quad \text{fixed } \rho$$

* for all ρ : add $\sqrt{\frac{\log\log_2 \frac{2}{\rho}}{m}}$

$$f = \sum_{t=1}^{T} \alpha_t h_t \quad \text{by AdaBoost}$$

$\bar{f} \in$ Convex-hull$(\mathcal{H})$

$$\mathrm{sgn}(f) = \mathrm{sgn}\left(\frac{f}{\|\alpha\|_1}\right) = \mathrm{sgn}(\bar{f})$$

$$\Rightarrow R(f) = R\left(\frac{f}{\|\alpha\|_1}\right) = R(\bar{f}).$$

$$\boxed{R(f) = R(\bar{f}) \le \hat{R}_{S,\rho}(\bar{f}) + \frac{2}{\rho}\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$$

☆ AdaBoost 算法的泛化界

与算法无关

$$\hat{R}_{S,\rho}(\bar{f}) \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y_i \bar{f}(x_i) \le \rho}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y_i \rho_f(x_i) \le \rho}$$

(L1 geometric margin: $\rho_f(x) = \frac{|f(x)|}{\|\alpha\|_1} = \frac{y f(x)}{\|\alpha\|_1}$)

$$= \frac{|\{i \in m : y_i \rho_f(x_i) \le \rho\}|}{m}.$$

fraction of points x with confidence margin $\le \rho$

"support vector"

Thm $f=\sum_{t=1}^{T}\alpha_t h_t$ by AdaBoost. (weak learner: $\varepsilon_t<\frac{1}{2}$)

$$\forall \rho>0 \quad \hat{R}_{S,\rho}(\bar{f}) \le 2^T\cdot\prod_{t=1}^{T}\sqrt{\varepsilon_t^{1-\rho}(1-\varepsilon_t)^{1+\rho}} \quad ☆$$

$$\alpha_t=\frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$$

$$Z_t=2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

by AdaBoost

Proof: $\hat{R}_{S,\rho}(\bar{f})=\frac{1}{m}\sum_{i=1}^{m}\Phi_\rho(y_i\bar{f}(x_i)) \le \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{y_i\bar{f}(x_i)\le\rho}$

$$=\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{y_i\bar{f}(x_i)-\rho\cdot\|\alpha\|_1\le 0}$$

$$\le\frac{1}{m}\sum_{i=1}^{m}e^{\rho\|\alpha\|_1}\cdot e^{-y_i f(x_i)}$$

$$=\cancel{\frac{1}{m}}\sum_{i=1}^{m}e^{\rho\|\alpha\|_1}\cdot D_{T+1}(i)\cdot\cancel{m}\prod_{t=1}^{T}Z_t$$

$$=e^{\rho\|\alpha\|_1}\prod_{t=1}^{T}Z_t\cdot\underbrace{\sum_{i=1}^{m}D_{T+1}(i)}_{\to 1}$$

(distribution)

$$=e^{\rho\sum_{t=1}^{T}\alpha_t}\prod_{t=1}^{T}Z_t$$

$$=\prod_{t=1}^{T}\left(\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}\right)^{\rho}\cdot 2\sqrt{\varepsilon_t(1-\varepsilon_t)}=2^T\cdot\prod_{t=1}^{T}\sqrt{\varepsilon_t^{1-\rho}(1-\varepsilon_t)^{1+\rho}} \quad \square$$

$$\mathbb{1}_{u\le 0}\le e^{-u}$$

$$\boxed{D_{t+1}(i)=\frac{e^{-y_i f_t(x_i)}}{m\cdot\prod_{s=1}^{t}Z_s}} \quad (*)$$

$f=f_T$

$$\hat{R}_{S,\rho}(\bar{f})\le 2^T\cdot\prod_{t=1}^{T}\sqrt{\varepsilon_t^{1-\rho}(1-\varepsilon_t)^{1+\rho}}$$

$\varepsilon_t<\frac{1}{2}$

$\forall t\in[T]$. $\frac{1}{2}-\varepsilon_t\ge\gamma$.

edge.

$$\log\left(\varepsilon_t^{1-\rho}(1-\varepsilon_t)^{1+\rho}\right)=(1-\rho)\log\varepsilon_t+(1+\rho)\log(1-\varepsilon_t)$$

求导: $\frac{1-\rho}{\varepsilon_t}-\frac{1+\rho}{1-\varepsilon_t}=\frac{1-\rho-2\varepsilon_t}{\varepsilon_t(1-\varepsilon_t)}>0 \Rightarrow \varepsilon_t\in\left(0,\frac{1}{2}-\frac{\rho}{2}\right)$

$\varepsilon_t\le\frac{1}{2}-\gamma$.

if $\rho\le 2\gamma$: $\varepsilon_t^{1-\rho}(1-\varepsilon_t)^{1+\rho}$ is maximized at $\varepsilon_t=\frac{1}{2}-\gamma$

$\Rightarrow$ upper bound of margin loss:

$$\boxed{\hat{R}_{S,\rho}(\bar{f})\le\left(\underbrace{(1-2\gamma)^{1-\rho}\cdot(1+2\gamma)^{1+\rho}}_{\le 1}\right)^{\frac{T}{2}}} \quad ☆☆$$

$$(1-2\gamma)^{1-\rho}(1+2\gamma)^{1+\rho} = (1-4\gamma^2)\underbrace{\left(\frac{1+2\gamma}{1-2\gamma}\right)}_{>1}^{\rho} \uparrow \quad \rho\uparrow$$

if $\rho<\gamma$:

$$(1-2\gamma)^{1-\rho}(1+2\gamma)^{1+\rho} < (1-2\gamma)^{1-\gamma}\cdot(1+2\gamma)^{1+\gamma}. \qquad \gamma\in(0,\tfrac{1}{2})$$

$$\leq 1 \quad (\gamma=0)$$

RHS decreases exponentially with T.

T越大, $\hat{R}_{S,\rho}$ 越小 $\to 0$, R越小

与γ有关! $\varepsilon_t \to \frac{1}{2}$ $\gamma\to 0$. $\Rightarrow$ uninformative

$\varepsilon_t \uparrow$

$$\boxed{\gamma^* = \frac{\rho_{max}}{2}}$$

$\frac{1}{3}\cdot\frac{3}{8}$

(converge)

AdaBoost achieves an L_1-geometric margin of γ over samples.

$$\forall i: \quad \frac{y_i f(x_i)}{\|\alpha\|_1} \geq \rho_f \longrightarrow \gamma$$

* AdaBoost does NOT achieve the maximum margin ρ_{max}.

为什么训练误差为零后, 测试误差继续降低?

↳ $\hat{R}_S(f)=0 \overset{\times}{\Rightarrow} \hat{R}_{S,\rho}(f)=0 \Rightarrow \rho\uparrow \Rightarrow \gamma\uparrow$

0/1-loss

上界

ρ-margin-loss

margin edge

0-1损失误差为零后

ρ-margin损失误差仍在下降 (i.e. margin ↑) !

$\boxed{\text{Margin-Maximization}}$

$$\rho = \max_{\alpha} \min_{i\in[m]} \frac{y_i\cdot(\alpha\cdot h(x_i))}{\|\alpha\|_1}$$

L_1-margin.

$$\Rightarrow \max_{\alpha} \ \rho \quad \text{s.t.} \quad \frac{y_i(\alpha \cdot h(x_i))}{\|\alpha\|_1} \geq \rho, \ \forall i \in [m]$$

$\|\alpha\|_1 \to 1$

$$\Rightarrow \max_{\alpha} \ \rho \quad \text{s.t.} \quad y_i(\alpha \cdot h(x_i)) \geq \rho, \ \forall i \in [m]$$

经验风险 (+正则化项)

$\|\alpha\|_1 = 1 \Rightarrow \alpha_t \geq 0, \ \sum_{t=1}^{T} \alpha_t = 1$

Linear Programming (LP)

AdaBoost outperforms LP.

Similar to SVM !

$$\min \ \frac{1}{2}\|w\|_2^2 + C \cdot \sum_{i=1}^{m} \xi_i^p$$

$\Downarrow$

L1-regularization AdaBoost.

Motivation: $f = \sum_{t=1}^{T} \alpha_t h_t(x)$

$h_t : \alpha \to 1$ (dominate)

$h_t : \alpha \to 0$

"even distribution"

Objective:

$$F(\bar{\alpha}) = \frac{1}{m} \sum_{i=1}^{m} e^{-y_i f(x_i)}$$

Coordinate Descent

$$\boxed{G(\bar{\alpha}) = \underbrace{\frac{1}{m} \sum_{i=1}^{m} e^{-y_i f(x_i)}}_{\text{经验风险}} + \underbrace{\lambda \|\alpha\|_1}_{\text{正则化项}}}$$

Generalization Bound:

$$R(f) \leq \hat{R}_{S,\rho}(f) + \frac{2}{\rho} \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \log_2 \frac{2}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \quad \text{for all } \rho$$

$$\leq \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{y_i f(x_i) \leq \rho}$$

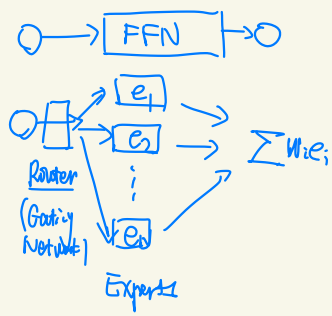
$$\leq \frac{1}{m} \sum_{i=1}^{m} e^{1 - \frac{y_i f(x_i)}{\rho}}$$

$\mathbb{1}_{u \leq 0} \leq e^{-u}$

$\|\alpha\|_1 \leq 1$

↳ 只与 ρ 有关 (ρ 越大)

* MoE (Mixture-of-Experts)



load-balancing problem

auxiliary loss

$KL(d \| p)$

f 和 $\frac{f}{\rho}$ 的符号一样 $\Rightarrow$ $\frac{1}{m}\sum_{i=1}^{m} e^{-f(x_i)}$ $\quad \|\vec{\alpha}\|_1 \le \frac{1}{\rho}$

$$\Rightarrow \text{minimize} \quad \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f(x_i)} = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \alpha_j h_j(x_i)}$$

subject to $\|\alpha\|_1 \le \frac{1}{\rho}$

Lagrangian :

$$\boxed{\text{minimize} \quad \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \alpha_j h_j(x_i)} + \lambda \|\alpha\|_1}$$

Intuition : KKT condition identical. $\quad$ L1-regularized AdaBoost.

AdaBoost seeks the large margin, but cannot achieve maximal margin ρ_{max}.

$$\boxed{\gamma^* = \frac{\rho_{max}}{2}}$$

def (edge) The edge for base-learner h_t

for distribution $\mathcal{D}$ over the training sample $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$

$$\star \quad \gamma_t(\mathcal{D}) = \frac{1}{2} - \varepsilon_t = \frac{1}{2}\left((1-\varepsilon_t) - \varepsilon_t\right) = \frac{1}{2}\sum_{i=1}^{m} y_i h_t(x_i) \mathcal{D}(i)$$

正确 $\quad$ 错误

$y_i h_t(x_i) = 1 \quad y_i h_t(x_i) = -1$

$\sum_{i:\text{正确}} \mathcal{D}(i) + \sum_{i:\text{错误}} \mathcal{D}(i)$

$\forall t \in [T]$, $\gamma_t(\mathcal{D}) \ge \gamma$ $\quad \exists \gamma$

AdaBoost's Weak Learning Condition. $\quad (\forall h_t : \varepsilon_t \le \frac{1}{2} - \gamma)$

Game-Theory

零和博弈

def (Zero-sum Game) loss matrix $M \in \mathbb{R}^{m \times n}$.

行 row player : m actions

列 column player : n actions

M_{ij} : loss for row player when row i column j

(payoff for column player)

e.g. rock-paper-scissors.

row		○	✓	□
$\frac{1}{2}$	○	0	-1	$+1$
$\frac{1}{3}$	✓	$+1$	0	-1
$\frac{1}{6}$	□	-1	$+1$	0

混合策略

def (Mixed Strategy) $\begin{cases} \text{row player}: \ p \in \mathbb{R}^m \\ \text{column player}: \ q \in \mathbb{R}^n \end{cases}$ distribution.

expected loss

$$\mathbb{E}_{\substack{i\sim p \\ j\sim q}}[M_{ij}] = \sum_{i=1}^{m}\sum_{j=1}^{n} p_i M_{ij} q_j = p^T M q$$

def (Von Neumann's Minimax Theorem) finite zero-sum game (M)

☆☆ $$\text{Value} = \boxed{\min_p \max_q p^T M q = \max_q \min_p p^T M q}$$

行: ↓ min

列: ↑ max

不管谁先, 期望结果都一样.

proof in Chapter 8.

f: convex-concave $\begin{cases} f(\cdot, y): \text{convex} \\ f(x, \cdot): \text{concave} \end{cases}$

$$\min_x \max_y f(x,y) = \max_y \min_x f(x,y)$$

pure strategy (deterministic action) : $\min_p \max_{j\in[n]} p^T M e_j = \max_q \min_{i\in[m]} e_i^T M q$

AdaBoost is a zero-sum game.

		Action	Mixed Strategy
row player:	training example x_i, $i \in [m]$,	$\rightarrow$	$\mathcal{D}$
column player:	base learner h_t, $t \in [T]$,	$\rightarrow$	α

loss matrix $M \in \mathbb{R}^{m \times T}$, $M_{it} \in \{-1, +1\}$

$$\boxed{M_{it} = y_i h_t(x_i)}$$

$f = \sum_{t=1}^{T} \alpha_t h_t$: Column player's strategy

分类正确为1

错误为-1

Column: 列 loss ↓ ⇒ 行 loss ↑ 尽可能分类正确

Row: 走成率越难 ⇒ 行 loss ↓ 尽可能分类错误

max] Game min

by Von-Neumann's Minimax Theorem

$$\min_{\mathcal{D}} \max_{t} \sum_{i=1}^{m} \mathcal{D}(i) \cdot y_i h_t(x_i) = \max_{\alpha} \min_{i} \left(\sum_{t=1}^{T} \frac{\alpha_t}{\|\alpha\|_1} y_i h_t(x_i) \right)$$

$f_t(x_i)$

column player: pure

row player: pure

$$\min_{\mathcal{D}} \max_{t} 2\gamma_t(\mathcal{D})$$

$y_i f_t(x_i)$: confidence margin

$\frac{y_i f_t(x_i)}{\|\alpha\|_1}$: L_1-geometric margin

$$\boxed{2\gamma^* = \rho_{max}} = \max_{\alpha} \underbrace{\min_{i} \rho_f(x_i)}_{L_1\text{-margin}}$$

$\gamma^* = \frac{1}{2} \rho_{max}$: AdaBoost cannot achieve maximal margin

objective of AdaBoost.

9 Multi-Class Classification

二分类 (Binary Classification) → Chapter 9. 多分类

$h: \mathcal{X} \to \{-1, +1\}$ → Chapter 11. 回归 (regression) $y \in \mathbb{R}$

SVM. AdaBoost

Chapter 8. 在线学习 (Online Learning)

多分类:

☆ mono-label: $y = \{1, \cdots, k\}$ {山猫, 野猪, 鲤鱼}

multi-label: $y = \{-1, +1\}^k$ < CV: 图像识别 / NLP: 文本分类

error:

(mono) → $R(h) = \mathbb{E}_{x \sim \mathcal{D}}\left[\mathbb{1}_{h(x) \neq f(x)} \right]$ h: hypothesis f: target

(multi) → $R(h) = \mathbb{E}_{x \sim \mathcal{D}}\left[\sum_{i=1}^{k} \mathbb{1}_{h(x)_i \neq f(x)_i} \right]$ "不合" (disagreement)

Hamming Distance $x, y \in \mathbb{R}^n$ $d_H(x, y) = \sum_{i=1}^{n} \mathbb{1}_{x_i \neq y_i}$

$\Rightarrow R(h) = \mathbb{E}_{x \sim \mathcal{D}}\left[d_H(h(x), f(x)) \right]$ generalization error

$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} d_H(h(x_i), y_i)$ empirical error

Potential Issues:

① computationally costly

② unbalanced class 类别不平衡 → { cost (代价) / weight (权重) / sampling (采样) } !

"focal loss"

③ relationship between classes (hierarchical) numerical × graphical ✓

Mono-label $\mathcal{Y}=\{1,\cdots,k\}$

$$x \mapsto \operatorname*{argmax}_{y\in\mathcal{Y}} h(x,y) \Leftarrow h(x)$$

分数 score

$\Downarrow$

margin $\rho_h(x,y) = h(x,y) - \max_{y'\neq y} h(x,y')$

h misclassify $h(x,y)$ iff $\rho_h(x,y)\leq 0$ ☆

↳ Relationship Learning 关联学习

PRML

西瓜书 Chapter 16

confidence margin

$\rho_h(x) = y\cdot h(x)$

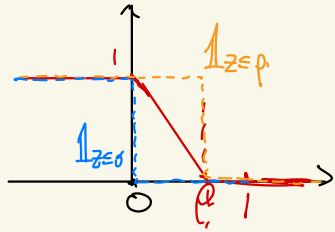
$\begin{cases} \text{予测判对的话} \Rightarrow \rho_h(x)>0 \\ \text{予测判错了} \Rightarrow \rho_h(x)\leq 0 \end{cases}$

Empirical Margin Loss $\hat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m}\Phi_\rho(\rho_h(x_i,y_i))$

Generalised ~: $R_\rho(h) = \mathbb{E}_S[\hat{R}_{S,\rho}(h)]$

$\Phi_\rho(z) = \min\left(1, \max\left(0, 1-\frac{z}{\rho}\right)\right)$

$$\mathbb{E}[\mathbb{1}_{\rho_h(x,y)\leq 0}] \leq R_\rho(h) \leq \mathbb{E}[\mathbb{1}_{\rho_h(x,y)\leq\rho}]$$



Lemma $\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_l$: l hypothesis sets in $\mathbb{R}^{\mathcal{X}}$

$$G = \{\max\{h_1,\cdots,h_l\} : h_i\in\mathcal{F}_i,\ i\in[l]\} \Rightarrow \boxed{\hat{R}_S(G) \leq \sum_{i=1}^{l}\hat{R}_S(\mathcal{F}_i)}$$

max → Σ

Proof: 只证 $l=2$ 情况

$$\max\{h_1,h_2\} = \frac{1}{2}(h_1+h_2+|h_1-h_2|)$$

$$\hat{R}_S(G) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{h_1\in\mathcal{F}_1\\ h_2\in\mathcal{F}_2}}\sum_{i=1}^{m}\sigma_i \max\{h_1,h_2\}\right]$$

$$= \frac{1}{2m} \mathbb{E}_\sigma\left[\sup_{\substack{h_1\in\mathcal{F}_1\\ h_2\in\mathcal{F}_2}} \sum_{i=1}^m \sigma_i \left(h_1(x_i)+h_2(x_i)+|h_1(x_i)-h_2(x_i)|\right)\right]$$

$$\le \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{F}_1)+\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{F}_2)+\frac{1}{2m}\mathbb{E}_\sigma\left[\sup_{h_1,h_2}\sum_{i=1}^m \sigma_i |h_1(x_i)-h_2(x_i)|\right]$$

$\boxed{\sup(a+b)\le \sup a+\sup b}$

sub-additivity

$y=|x|$: 1-lipschitz

函数复合

By Talagrand's Lemma, $\hat{\mathcal{R}}_S(\Phi\circ\mathcal{H}) \le \ell\cdot\hat{\mathcal{R}}_S(\mathcal{H})$ (Φ: ℓ-Lipschitz)

$$\frac{1}{2m}\mathbb{E}_\sigma\left[\sup_{h_1,h_2}\sum_{i=1}^m \sigma_i |h_1(x_i)-h_2(x_i)|\right] \le \frac{1}{2m}\mathbb{E}_\sigma\left[\sup_{h_1,h_2}\sum_{i=1}^m \sigma_i (h_1(x_i)-h_2(x_i))\right]$$

$$\le \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{F}_1)+\frac{1}{2m}\mathbb{E}_\sigma\left[\sup_{h_2}\sum_{i=1}^m -\sigma_i h_2(x_i)\right] = \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{F}_1)+\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{F}_2)$$

$$\Rightarrow \hat{\mathcal{R}}_S(\mathcal{G}) \le \hat{\mathcal{R}}_S(\mathcal{F}_1)+\hat{\mathcal{R}}_S(\mathcal{F}_2)$$

* $\max\{h_1,\cdots,h_\ell\} = \max\{h_1, \max\{h_2,\cdots,h_\ell\}\}$ recurrently □

$g=\max\{h_1,\cdots,h_\ell\}$ $\quad \hat{\mathcal{R}}_S(\mathcal{G}) \le \sum_{i=1}^{\ell} \hat{\mathcal{R}}_S(\mathcal{F}_i)$

scoring 函数

Hypothesis Space: $\boxed{\Pi_1(\mathcal{H}) = \{x\mapsto h(x,y),\ y\in\mathcal{Y},\ h\in\mathcal{H}\}}$

"label-augmented hypothesis space"

$\boxed{\mathcal{H}\to\mathcal{H}\times\mathcal{Y}}$

将标签空间(label space)的复杂性也纳入考虑之中

Thm $\mathcal{H}\subseteq\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$, $\mathcal{Y}=\{1,\cdots,k\}$

fix $\rho>0$, $\forall\delta>0$, w.p. $\ge 1-\delta$:

$$\forall h\in\mathcal{H}:\ \boxed{R(h) \le \hat{R}_{S,\rho}(h) + \frac{4k}{\rho}\cdot\mathcal{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$$

ρ-margin loss

☆☆ 多分类问题泛化界

Proof: $\rho_{\theta,h}(x,y) = \min_{y'}\left(h(x,y) - h(x,y') + \theta \cdot \mathbb{1}_{y'=y}\right)$

约束转化成无约束

surrogate margin

$\rho_h(x,y) = h(x,y) - \max_{y' \neq y} h(x,y')$

① Upper Bound:

$$\mathbb{E}\left[\mathbb{1}_{\rho_h(x,y)\le 0}\right] \le \mathbb{E}\left[\mathbb{1}_{\rho_{\theta,h}(x,y)\le 0}\right]$$

θ

$y'=y \Rightarrow \rho_{\theta,h}(x,y) \le \rho_h(x,y)$ constraint.

$y'\neq y \Rightarrow \rho_{\theta,h}(x,y) = \rho_h(x,y)$

$$\rho_{\theta,h}(x,y) \le \rho_h(x,y) \Leftarrow \rho_{\theta,h}(x,y) = \min_{y'}\left(h(x,y) - h(x,y') + \theta \cdot \mathbb{1}_{y'=y}\right)$$

$$\le \min_{y'\neq y}\left(h(x,y) - h(x,y') + \theta \cdot \mathbb{1}_{y'=y}\right)$$

taking the minimum on a smaller set

$$= \min_{y'\neq y}\left(h(x,y) - h(x,y')\right) = \rho_h(x,y)$$

$$\le \hat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m}\Phi_\rho\left(\rho_h(x_i,y_i)\right)$$

θ margin

$$\mathbb{E}[g(Z)] \le \frac{1}{m}\sum_{i=1}^{m} g(Z_i) + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\tilde{\mathcal{H}} = \left\{(x,y) \mapsto \rho_{\theta,h}(x,y) : h \in \mathcal{H}\right\}$$

$$\tilde{\mathcal{H}} = \left\{\Phi_\rho \circ h : h \in \tilde{\mathcal{H}}\right\}$$

$$\Rightarrow \mathbb{E}_{(x,y)}\left[\Phi_\rho\left(\rho_{\theta,h}(x,y)\right)\right] \le \frac{1}{m}\sum_{i=1}^{m}\Phi_\rho\left(\rho_{\theta,h}(x_i,y_i)\right) + 2\mathcal{R}_m(\tilde{\mathcal{H}}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$\Phi_\rho(x) \ge \mathbb{1}_{x\le 0}$

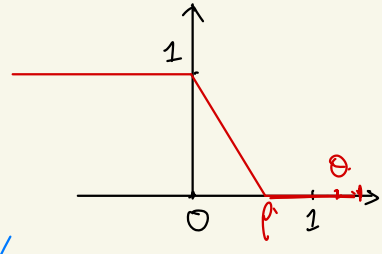
$$\ge \mathbb{E}_{(x,y)}\left[\mathbb{1}_{\rho_{\theta,h}(x,y)\le 0}\right] = \mathbb{E}_{(x,y)}\left[\mathbb{1}_{\rho_h(x,y)\le 0}\right] = R(h)$$

misclassification

Φ_ρ : $\frac{1}{\rho}$-Lipschitz. $\mathcal{R}_m(\tilde{\mathcal{H}}) \le \frac{1}{\rho}\mathcal{R}_m(\mathcal{H})$. (by Talagrand's Lemma)

$\Phi_\rho(\rho_{\theta,h}(x,y)) = \Phi_\rho(\rho_h(x,y))$ since

$$\begin{cases} y' \neq y: & \rho_{\theta,h}(x,y) = \rho_h(x,y) \\ y' = y: & \theta = \rho_{\theta,h}(x,y) \leq \rho_h(x,y) \end{cases}$$



fix $\theta = 2\rho > \rho$: $\frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(\rho_h(x_i,y_i))$ $\Downarrow$

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathfrak{R}_m(\tilde{\mathcal{H}}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \qquad \text{w.p.} \geq 1-\delta$$

只需证 $\mathfrak{R}_m(\tilde{\mathcal{H}}) \leq 2k \cdot \mathfrak{R}_m(\Pi_1(\mathcal{H}))$

$$\mathfrak{R}_m(\tilde{\mathcal{H}}) = \frac{1}{m}\mathbb{E}_S\left[\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\,\rho_{\theta,h}(x_i,y_i)\right]\right]$$

$$= \frac{1}{m}\mathbb{E}_S\left[\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(x_i,y_i) - \max_y\left(h(x_i,y) - \overset{\theta}{2\rho}\,\mathbb{1}_{y=y_i}\right)\right]\right]$$

$$\leq \underbrace{\frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(x_i,y_i)\right]}_{①} + \underbrace{\frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i \max_{y\in\mathcal{Y}}\left(h(x_i,y) - 2\rho\,\mathbb{1}_{y=y_i}\right)\right]}_{②}$$

$$① = \frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}}\sigma_i h(x_i,y_i)\mathbb{1}_{y=y_i}\right]$$

$$\leq \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(x_i,y_i)\mathbb{1}_{y=y_i}\right]$$

$$= \sum_{y\in\mathcal{Y}}\frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(x_i,y_i)\cdot\left(\frac{\varepsilon_i}{2}+\frac{1}{2}\right)\right]$$

$\varepsilon_i = 2\cdot\mathbb{1}_{y_i=y} - 1 \in \{-1,+1\}$

$$\leq \sum_{y\in\mathcal{Y}}\left(\underbrace{\frac{1}{2m}\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\varepsilon_i h(x_i,y_i)\right]}_{\sigma_i} + \frac{1}{2m}\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(x_i,y_i)\right]\right.$$

$\le k\, \mathcal{R}_m(\Pi_1(\mathcal{H}))$

ε_i 和 σ_i 同分布 $\{-1,+1\}$

② by Lemma, $\le \sum_{y\in\mathcal{Y}} \frac{1}{m} \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i \left(h(x_i,y) - 2\rho \mathbb{1}_{y=y_i}\right)\right]$

$2\rho\sigma_i$ 期望为0

$\le \sum_{y\in\mathcal{Y}} \frac{1}{m} \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i,y)\right]$

$= k\cdot \mathcal{R}_m(\Pi_1(\mathcal{H}))$

Summary $\Rightarrow$ $\mathcal{R}_m(\tilde{\mathcal{H}}) \le 2k\cdot \mathcal{R}_m(\Pi_1(\mathcal{H}))$

$\rho_{\theta,h}(x,y) \to h(x,y)$

$$\Rightarrow \boxed{R(h) \le \hat{R}_{S,\rho}(h) + \frac{4k}{\rho}\mathcal{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}} \quad \square$$

fixed ρ

for any ρ: $+\sqrt{\frac{\log\log_2\frac{2}{\rho}}{m}}$

想要 ρ 小

想要 ρ 大

Margin Bound

类别数 $|\mathcal{Y}|$: 越多, 界越弱!

Kernalized

$K: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ PDS kernel

Φ: feature mapping

linear $w_1,\cdots,w_k \in \mathbb{H}$

$x \mapsto w_0\cdot\Phi(x) \to \operatorname{argmax}_{y\in\mathcal{Y}} w_y\cdot\Phi(x)$

W: weight matrix 矩阵

$W = (w_1,\cdots,w_k)^T \in \mathbb{R}^{k\times d_{\mathbb{H}}}$

$\forall j\in[k]$ $\|w_k\|_{\mathbb{H}} \le \Lambda$

def (Group Norm) $L_{p,q}$ Group Norm: $\|W\|_{p,q} = \left(\sum_{i=1}^{k} \|w_i\|_p^q\right)^{\frac{1}{q}}$

row-vector

* 矩阵范数 $\begin{cases} \text{inductive norm 诱导范数} & \text{e.g. Spectral Norm } \|M\|_2 = \sup_x \frac{\|Mx\|_2}{\|x\|_2} \\ \text{entry-wise norm 逐元素范数} & \text{e.g. Frobenius Norm } \|M\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}^2} \end{cases}$

$\mathcal{H}_{K,p} = \{ (x,y) \in \mathcal{X} \times \{1,\cdots,k\} \mapsto w_y \cdot \Phi(x) : W = (w_1,\cdots,w_k)^T, \boxed{\|W\|_{\mathbb{H},p} \le \Lambda} \}$

Proposition: $R_m(\Pi_1(\mathcal{H}_{K,p})) \le \sqrt{\frac{r^2\Lambda^2}{m}}$

$K(x,x) \le r^2$

$\Updownarrow$

$\|\Phi(x)\|_{\mathbb{H}} \le r$

Cor (Kernelized Margin Bound) w.p. $\ge 1-\delta$

$$R(h) \le \hat{R}_{S,\rho}(h) + 4k\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

Multi-Class Classification Algorithms

- uncombined algorithms: single optimization problem
- aggregated algorithms: multiple binary classifier.

Uncombined Algorithms

① Multi-Class SVM.

$$R(h) \le \frac{1}{m}\sum_{i=1}^{m} \xi_i + 4k\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \quad \text{w.p.} \ge 1-\delta$$

有k个 $w_l \Phi(x_i)$ $\forall l \in [k]$

$w_{y_i} \cdot \Phi(x_i) - \max_{y' \ne y_i} w_{y'} \cdot \Phi(x_i) \le 1$

$\begin{cases} \le 0: \text{错误} \\ (0,1): \text{正确with less confidence} \end{cases}$

二维: $y_i \cdot w \cdot \Phi(x_i) \le 1$

$\xi_i = \max(1 - y_i \cdot w \cdot \Phi(x_i), 0)$

$$\xi_i = \max\left(1-\left(w_{y_i}\Phi(x_i) - \max_{y\neq y_i} w_y \Phi(x_i)\right), 0\right) \quad \forall i\in[m].$$

$$\boxed{\begin{aligned} &\min_{w,\xi} \frac{1}{2}\sum_{l=1}^{k}\|w_l\|_{\mathbb{H}}^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t.} \quad \forall i\in[m],\ \forall l\in \mathcal{Y}-\{y_i\} \\ &\qquad w_{y_i}\Phi(x_i) \geq w_l\Phi(x_i) + 1 - \xi_i \\ &\qquad \xi_i \geq 0 \end{aligned}}$$

SVM:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.}\ y_i(w\cdot\Phi(x_i)) \geq 1-\xi_i.$$

$$\xi_i \geq 0. \quad \forall i\in[m].$$

$\Theta(mk)$ 个约束

$\Downarrow$ QP

$\Omega(mk)$

若类别数 k 很大

则解需要大代价！

$\Downarrow$

Lagrange multiplier $\alpha\in\mathbb{R}^{m\times k}$ $\quad \alpha_i$ 是 α 的第 i 行

Dual:

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i\vec{e}_{y_i} - \frac{1}{2}\sum_{i=1}^{m}(\alpha_i\cdot\alpha_j)\cdot K(x_i,x_j)$$

$$\text{s.t.}\ \forall i\in[m].\quad 0\leq \alpha_{iy_i}\leq C$$

$$\alpha_{ij}\leq 0. \quad \forall j\neq y_i$$

$$\vec{\alpha}_i\cdot\vec{1} = 0$$

② Multi-Class Boosting

AdaBoost.MH : multi-label $\mathcal{Y}=\{-1,+1\}^k$

AdaBoost: $\mathcal{H}=\{h_1,\cdots,h_N\}$ $\quad \bar{\alpha}=(\bar{\alpha}_1,\bar{\alpha}_2,\cdots,\bar{\alpha}_N)$. $\quad f_N=\sum_{j=1}^{N}\bar{\alpha}_j h_j$

二分类 $\quad F(\bar{\alpha}) = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f_N(x_i)} = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i\sum_{j=1}^{N}\bar{\alpha}_j h_j(x_i)}$

$\Downarrow$

多分类 $\quad F(\bar{\alpha}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{l=1}^{k} e^{-y_i[l] f_N(x_i,l)} = \frac{1}{m}\sum_{i=1}^{m}\sum_{l=1}^{k} e^{-y_i[l]\sum_{j=1}^{N}\bar{\alpha}_j h_j(x_i,l)}$

l^{th} coordinate of y_i

m: 样本数
k: 类别数 $\Rightarrow$ mk: 总样本数
N: 轮数(基学习器个数)

$(x_i, y_i) \longrightarrow \boxed{\text{Splitting}} \longrightarrow ((x_i, l), y_i[l]), l \in [k]$

分割成k个标记样本

AdaBoost.MH($S = ((x_1, y_1), \ldots, (x_m, y_m))$)

for $i \leftarrow 1$ **to** m **do**
  for $l \leftarrow 1$ **to** k **do**
    $\mathcal{D}_1(i, l) \leftarrow \frac{1}{mk}$
for $j \leftarrow 1$ **to** N **do**
  $h_j \leftarrow$ base classifier in $\mathcal{H}$ with small error $\epsilon_j = \mathbb{P}_{(i,l)\sim\mathcal{D}_j}[h_j(x_i, l) \neq y_i[l]]$
  $\bar{\alpha}_j \leftarrow \frac{1}{2} \log \frac{1-\epsilon_j}{\epsilon_j}$
  $Z_t \leftarrow 2[\epsilon_j(1-\epsilon_j)]^{\frac{1}{2}}$ ▷ normalization factor
  for $i \leftarrow 1$ **to** m **do**
    for $l \leftarrow 1$ **to** k **do**
      $\mathcal{D}_{j+1}(i, l) \leftarrow \frac{\mathcal{D}_j(i,l)\exp(-\bar{\alpha}_j y_i[l] h_j(x_i, l))}{Z_j}$
$f_N \leftarrow \sum_{j=1}^{N} \bar{\alpha}_j h_j$
return $h = \mathrm{sgn}(f_N)$

ERM

$1 \to k$

$\forall l \in [k]$

调整分布

$$S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$$

$m \downarrow mk$

$$S' = \{((x_i, 1), y_i[1]), ((x_i, 2), y_i[2]), \cdots, ((x_i, k), y_i[k]) : i \in [m]\}$$

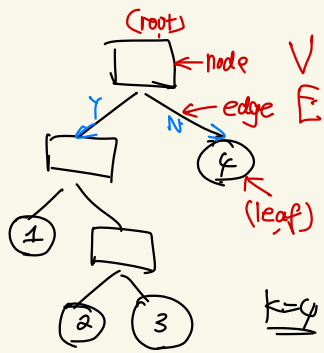
AdaBoost + Stump: $O(md \log m + mdN)$

AdaBoost.MH + Stump: $O(mkd \log(mk) + mkdN)$

- $k \uparrow$ impractical!
- base learner: $\varepsilon < \frac{1}{2}$ 弱学习器

③ Decision Tree

Ross-Quinlan.



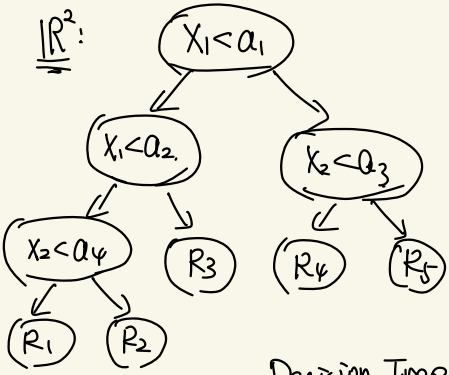
def (Binary Decision Tree) A tree representation of a partition of the feature space

划分

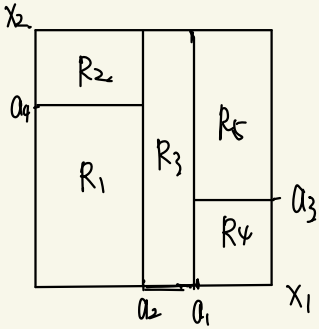
每一个叶结点代表一个标签(类别)

每一个中间结点代表一个问题 question ⇒ Bipartion (二分)

$\mathbb{R}^2$:



Decision Tree $\Longrightarrow$



Decision Tree: partition

- numerical question: $X_i \le a$
- categorical question: $X_i \in \{$山鸡, 野猪, 鼍$\}$

continuous or discrete (如何处理特征)
- 有序 → numerical
- 无序 → categorical

(A) Inference (推断)

go down the tree 推理 Reasoning
until get to a leaf!

(B) Training (训练)

1. Greedy (贪心策略)

每一轮选择一个结点 n, 构造一个划分 q.

```
tree ← {n_0}                    (root node)
for t = 1 to T:
    (n_t, q_t) ← argmax_(n,q) F(n, q)   (metric)
    SPLIT(tree, n_t, q_t)
return tree
```

$X_1 \le a$

不能划分? → leaf
- 样本全部同一类?
- 没有对应样本? (先验 prior)
- 样本相同

$(n_t, q_t) \to$ (node, question)

n_t: 维度 (特征); q_t: 问题 (划分)

Intuition: 划分后样本尽可能属于同一类 ("纯度" purity)

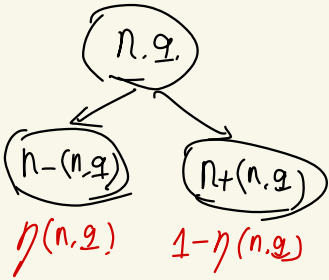
$F(n)$: node impurity (不纯洁度)

划分标准 $\Rightarrow$ $F(n, q)$: decrease of node impurity after splitting

$n_+(n,q)$: 右结点

$n_-(n,q)$: 左结点

impurity ↓



$$\tilde{F}(n,q) = F(n) - \Big(\eta(n,q)\cdot F(n_-(n,q)) + (1-\eta(n,q))F(n_+(n,q))\Big)$$

fraction: 比例

基于结点n和问题q能带来的 impurity decrease.

F ? $p_l(n)$: 结点n中属于第l类的样本比例

Impurity Measure:

$$F(n) = \begin{cases} 1-\max p_l(n) \\ -\sum_{l=1}^{k} p_l(n)\cdot \lg p_l(n) \\ \sum_{l=1}^{k} p_l(n)\cdot(1-p_l(n)) \end{cases}$$

misclassification 错误分类的比例

entropy 熵 ⇒ C4.5 → information gain (ID3)

$-\sum p \lg p$

$p=0$, $p=1$ 熵 $=0$

$p=\frac{1}{2}$ 最大 $\lg k$.

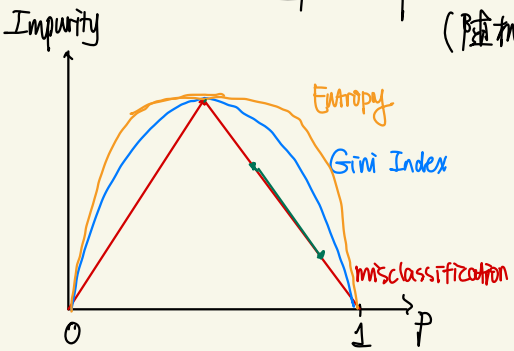
Gini Index 基尼指数

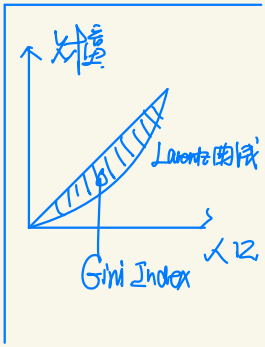
(随机抽两个样本，类别不一致的概率)

CART

⇓

Concave 凹性





$$\tilde{F} = F(n) - \Big(\eta(n,q)\cdot F(n_-(n,q)) + (1-\eta(n,q))\cdot F(n_+(n,q))\Big) \geq 0$$

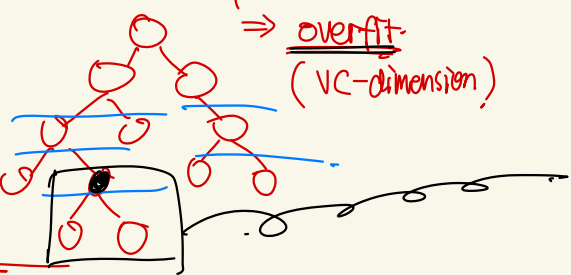
misclassification: $\tilde{F}=0$

2. Grow-then-Prune 先生长剪枝

regularization

$$G_\lambda(\text{tree}) = \sum_{n\in\text{tree}} |n|\cdot F(n) + \lambda\cdot|\text{tree}|$$

☆ deep tree ⇒ overfit (VC-dimension)



经验风险　　模型大小

已划分项

pruning

每一次剪枝：砍掉一个点（以其为根的子树 ⇒ 叶结点）

$F_{tree} = \sum_{n \in Tree} |n| F(n)$　　$|n|$: height

↓：划分

↑ ($\le \varepsilon$)：不划分

Decision Tree

pro：简单，高效

con：不适应，屡试屡败 (stable)

label：半监督学习

西瓜书 Chapter 4

样本权值，权重划分

---

uncombined

多 SVM

多 AdaBoost

Decision-Tree

impurity metric F

Learn:

(n, q)　Greedy $\tilde{F}_{argmax}$ → Pruning 子树替换叶子

aggregated algorithms：多个二类问题 组合得到多类问题

Occam's Razor

① One-versus-ALL (OVA) (rest)

K类 → K个学习器 { 正：第k类 / 负：y \ 第k类

$\forall x \in X,\ h(x) = \underset{l \in y}{argmax}\ f_l(x)$　"分数最高的类"

复杂度：K倍二分类复杂度

independent!

comparable?

normalization!

no principle justification

② One-versus-One (OVO)

K类 → $\binom{K}{2} = \frac{K(K-1)}{2}$ 个学习器 { 正：i / 负：j

$\{-1, +1\}$　正

$\forall x \in X$　$h(x) = \underset{l \in y}{argmax} \left| \{ l' : h_{ll'}(x) = 1,\ l' \in y \} \right|$

"锦标赛" (tournament)

↳ 将其作为正类而学习器中多少个就训为正类

OVO < $O(k^2)$个学习器 / $2\cdot\frac{m}{k}$ 个样本

= 分类学习器 < $O(m^\alpha)$ 学习 / $O(C_t)$ 推断

	Training	Testing
OVA	$O(k\cdot m^\alpha)$	$O(k\cdot C_t)$
OVO	$O(k^2\cdot(\frac{m}{k})^\alpha)$	$O(k^2\cdot C_t)$

↳ $O(k^{2-\alpha}\cdot m^\alpha)$

若优化问题复杂($\alpha \geq 1$): QP ($\alpha \geq 2$)

OVO学习有优势!

SVM通常采用OVO

③ MvM (多对多) → ECOC (Error-Correcting-Output-Code) 有纠错输出码

→ 对输出错误有一定容错能力

每个学习器把一部分类别作为正, 一部分类别作为负

$\forall x \in \mathcal{X}$, $\boxed{h(x) = \underset{l \in \mathcal{Y}}{\arg\min}\ d_H(M_l, h(x))}$

M矩阵 $M \in \mathbb{R}^{k\times c}$ < k: 类别数 / c: 学习器个数

codes

classes	1	2	3	4	5	6
1	-1	-1	-1	+1	-1	-1
2	+1	-1	-1	-1	-1	-1
3	-1	+1	+1	-1	+1	-1
4	+1	+1	-1	-1	-1	-1
5	+1	+1	-1	-1	+1	-1
6	-1	-1	+1	+1	-1	+1
7	-1	-1	+1	-1	-1	-1
8	-1	+1	-1	+1	-1	-1

5
6
1
4
3
3
3
4

Coding

$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$
-1	+1	+1	-1	+1	+1

new example x

→ $\boxed{h(x)} = 3$

$d(M_l, h(x))$ < Hamming distance / Euclidean distance / Manhattan distance

M_{ij}: 第i类在第j个学习器中是正类还是负类

trade-off < d大: 更复杂, 对错误的容忍度(纠错能力)更高 / d小: 更相似, 更精确

Complexity v.s. Accuracy

一. $h(x) = \arg\min_{l \in \mathcal{Y}} \sum_{j=1}^{c} L(\underbrace{M_{lj} f_j(x)}_{\text{confidence margin}})$

$L(x) = \frac{1 - \mathrm{sgn}(x)}{2} = \begin{cases} 1, & x \le 0 \\ 0, & x > 0 \end{cases}$

(0-1 loss)

$h(x) = \sum_{j=1}^{c} \frac{1 - \mathrm{sgn}(M_{lj} f_j(x))}{2} = d_h(M_l, h(x))$

Hamming Distance.

二. ternary ECOC. (三元ECOC)

$M_{lj} \in \{-1, 0, +1\}$

classifier j disregard of class l

三. $M_{lj} \in \mathbb{R}$ (continuous)

↓

learn matrix M.

区别性 经验风险

$$\min_{M, \xi} \|M\|_F^2 + C \cdot \sum_{i=1}^{m} \xi_i$$

$$\text{s.t. } K(f(x_i), M_{y_i}) - K(f(x_i), M_l) \ge 1 - \xi_i$$

$$\forall i \in [m],\ l \in \mathcal{Y}$$

☆

$f(x_i)$

M_{y_i}

余弦相似函

↓

similarity

↓

kernel method

PDS kernel $K(\cdot,\cdot)$

$h(x) = \arg\max_{l \in \mathcal{Y}} K(f(x), M_l)$