Boosting

Bias-Variance Tradeoff

$$\underbrace{\mathbb{E}[(h_D(x)-y)^2]}_{\text{test error}} = \underbrace{\mathbb{E}[(h_D(x)-\bar{h}(x))^2]}_{\text{variance}} + \underbrace{\mathbb{E}[(\bar{h}(x)-\bar{y}(x))^2]}_{\text{bias}} + \underbrace{\mathbb{E}[(\bar{y}(x)-y)^2]}_{\text{noise}}$$

$h \in \mathcal{H}$ — hypoth. class; high bias "weak learners"

Kearns, 88: $\mathcal{H} \longrightarrow \mathcal{H}^*$ ("strong learners" / low bias classifier)

Schapire 90: Yes!

$$H_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x) = \underbrace{\alpha_1 h_1(x)}_{\text{if. 1}} + \underbrace{\alpha_2 h_2(x)}_{\text{if 2}} + \cdots$$

(α_t: step-size)

$$\ell(H) = \frac{1}{n}\sum_{i=1}^{n} \ell(H(x_i), y_i)$$

imagine: $\operatorname{argmin}_{h_{t+1}\in\mathcal{H}} \ell(H_t + \alpha h_{t+1}(x))$ $\qquad H_{t+1} = H_t + \alpha h_{t+1}$

Gradient Descent in Function Space

Goal: given: H: partial ensemble; find h: weak learner

Solution: Taylor approx of $\ell(H+\alpha h) \approx \ell(H) + \alpha \langle \nabla \ell(H), h \rangle$ ($\alpha \to 0.1$)

$$\operatorname{argmin}_{h\in\mathcal{H}} \ell(H+\alpha h) \approx \operatorname{argmin}_{h\in\mathcal{H}} \ell(H) + \langle \nabla \ell(H), h \rangle$$

$$\langle g, k \rangle = \int g(x)k(x)dx \qquad \langle g, k \rangle = \sum_{i=1}^{n} g(x_i)k(x_i)$$

$$= \sum_{i=1}^{n} \frac{\partial \ell}{\partial H}(x_i)h(x_i) = \sum_{i=1}^{n} \frac{\partial \ell}{\partial [H(x_i)]} h(x_i)$$

$$h_{t+1} = \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} r_i h(x_i) \qquad \sum_{i=1}^{n} r_i h(x_i) < 0$$

$$h_{t+1} = \mathcal{A}(\{x_i, r_i\}_{i=1}^{n}) \longrightarrow h_{t+1}$$

Generic Boosting

Input: ℓ, α, $\{(x_i, y_i)\}_{i=1}^{n}$, $\mathcal{A}$

$H_0 = 0$

for $t = 0 : T-1$ do

$\quad \forall i,\ r_i = \frac{\partial \ell}{\partial [H(x_i)]}$

$\quad h_{t+1} = \mathcal{A}(\{x_i, r_i\}_{i=1}^{n})$

$\quad$ if $\sum_{i=1}^{n} r_i h_{t+1}(x_i) < 0$ then

$\quad\quad H_{t+1} = H_t + \alpha h_{t+1}$

$\quad$ else

$\quad\quad$ return H_t

$\quad$ end

end

return H_T

#1 Gradient Boosted Regression Trees (GBRT)

Setting

* $y \in \{-1, 1\}$ $\quad$ ($y \in \mathbb{R}$)
* fixed-depth (say 4) CART, $h(x) \in \mathbb{R}$
* α — fixed to small constant (hyperparameter)
* ℓ: diff., convex, decomposes: $\ell(H) = \sum_{i=1}^{n} \ell(H(x_i))$

Goal: find $h = \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} r_i h(x_i)$

Assumptions

1. $\sum_{i=1}^{n} h^2(x_i) = \text{constant}$ — why?
2. h is negation closed: $\forall h \in \mathcal{H} \Rightarrow -h \in \mathcal{H}$

$t_i = -r_i$

$$= \operatorname{argmin}_{h\in\mathcal{H}} -2\sum_{i=1}^{n} t_i h(x_i)$$

$$= \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} \underbrace{t_i^2}_{\text{constant}} - 2t_i h(x_i) + \underbrace{h(x_i)^2}_{\text{constant}}$$

$$= \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} (h(x_i) - t_i)^2$$

(the bracketed part: CART)

GBRT

Input: ℓ, α, $\{(x_i, y_i)\}$

$H_0 = 0$

for $t = 1 : T$ do

$\quad \forall i:\ t_i = y_i - H(x_i)$ $\quad (-\frac{\partial \ell}{\partial [H(x_i)]}$ in general)

$\quad h = \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} (h(x_i) - t_i)^2$

$\quad H \leftarrow H + \alpha h$

end

return H

#2 AdaBoost

Setting

* $y \in \{-1, 1\}$
* $h \in \mathcal{H}$, $h(x_i) \in \{-1, 1\}$
* α: optimal α every iteration
* ℓ: exponential loss: $\ell(H) = \sum_{i=1}^{n} e^{-y_i H(x_i)}$

Finding the best h

$$r_i = \frac{\partial \ell}{\partial H(x_i)} = -y_i e^{-y_i H(x_i)}$$

$$w_i = \frac{1}{Z} e^{-y_i H(x_i)}$$

normalizing constant: $Z = \sum_{i=1}^{n} e^{-y_i H(x_i)}$ $\qquad \sum_{i=1}^{n} w_i = 1$

$$h_{t+1} = \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} r_i h(x_i)$$

$$= \operatorname{argmin}_{h\in\mathcal{H}} -\sum_{i=1}^{n} y_i e^{-H(x_i)y_i} h(x_i) \qquad (w_i Z)$$

$$= \operatorname{argmin}_{h\in\mathcal{H}} -\sum_{i=1}^{n} y_i w_i h(x_i)$$

$y_i h(x_i) \in \{-1, 1\}$; $h(x_i) = y_i$

$$= \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i: h(x_i)\neq y_i} w_i - \sum_{i: h(x_i)=y_i} w_i$$

$$\sum_{i: h(x_i)=y_i} w_i = 1 - \sum_{i: h(x_i)\neq y_i} w_i$$

$$= \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i: h(x_i)\neq y_i} w_i$$

(weighted classification error ε)

Find the best α

$$\alpha = \operatorname{argmin}_{\alpha} \ell(H + \alpha h)$$

$$= \operatorname{argmin}_{\alpha} \underbrace{\sum_{i=1}^{n} e^{-y_i [H(x_i) + \alpha h(x_i)]}}_{\mathcal{L}}$$

$$\nabla_\alpha \mathcal{L} = -\sum_{i=1}^{n} y_i h(x_i) e^{-[y_i H(x_i) + \alpha y_i h(x_i)]} \Rightarrow 0$$

$y_i h(x_i) \in \{-1, 1\}$

$$= \sum_{i: h(x_i)y_i = 1} e^{-(y_i H(x_i) + \alpha \underbrace{y_i h(x_i)}_{1})} + \sum_{i: h(x_i)y_i = -1} e^{-(y_i H(x_i) + \alpha \underbrace{(y_i h(x_i))}_{-1})} = \frac{0}{Z}$$

$$= \sum_{i: h(x_i)y_i = 1} \underbrace{w_i e^{-\alpha}}_{\frac{1}{Z} e^{-y_i H(x_i)}} + \sum_{i: h(x_i)y_i = -1} w_i e^{\alpha} = 0$$

$$-(1-\varepsilon)e^{-\alpha} + \varepsilon e^{\alpha} = 0$$

$$e^{2\alpha} = \frac{1-\varepsilon}{\varepsilon}$$

$$\alpha = \frac{1}{2} \ln \frac{1-\varepsilon}{\varepsilon}$$

$\varepsilon = \sum_{i: h(x_i)y_i = -1} w_i$

Re-normalization

$H_{t+1} = H_t + \alpha h$

unnormalized weights: $\hat{w}_i \leftarrow \hat{w}_i * e^{-\alpha h(x_i) y_i}$

$Z \leftarrow Z * 2\sqrt{\varepsilon(1-\varepsilon)}$

$w_i \leftarrow w_i \frac{e^{-\alpha h(x_i) y_i}}{2\sqrt{\varepsilon(1-\varepsilon)}}$

Further analysis

weight update

$h(x_i) y_i \in \{-1, 1\}$

$e^{\alpha} > 1$ if $h(x_i)y_i = -1$ $\quad \uparrow w_i$

$e^{-\alpha} < 1$ if $h(x_i)y_i = 1$ $\quad \downarrow w_i$

normalization update

$$\ell(H) = Z = n \prod_{t=1}^{T} 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

$\varepsilon = \max_t \varepsilon_t$

$$\ell(H) \le n\left[2\sqrt{\varepsilon(1-\varepsilon)}\right]^T$$

$\frac{1}{2} = \operatorname{argmax}_{\varepsilon} \varepsilon(1-\varepsilon)$ $\qquad \varepsilon_t < \frac{1}{2}\ \forall t$

$\varepsilon(1-\varepsilon) < \frac{1}{4}$

$\varepsilon(1-\varepsilon) = \frac{1}{4} - \gamma^2$

$$\ell(H) \le n\underbrace{\left[1 - 4\gamma^2\right]^{\frac{T}{2}}}_{<1}$$

training error: $\delta_{H(x_i)\neq y_i} < e^{y_i H(x_i)}$

training loss: $\sum_{i=1}^{n} e^{y_i H(x_i)} = \ell(H) \le n\left[1-4\gamma^2\right]^{\frac{T}{2}} < 1$

AdaBoost (exp.)

Input: ℓ, $\{(x_i, y_i)\}$

$H_0 = 0$

$\forall i\ w_i = \frac{1}{n}$

for $t = 0 : T-1$ do

$\quad h = \operatorname{argmin}_h \sum_{i: h(x_i)\neq y_i} w_i$

$\quad \varepsilon = \sum_{i: h(x_i)\neq y_i} w_i$

$\quad$ if $\varepsilon < \frac{1}{2}$ then

$\quad\quad \alpha = \frac{1}{2} \ln\left(\frac{1-\varepsilon}{\varepsilon}\right)$

$\quad\quad H_{t+1} = H_t + \alpha h$

$\quad\quad \forall i\ w_i \leftarrow \frac{w_i e^{-\alpha h(x_i) y_i}}{2\sqrt{\varepsilon(1-\varepsilon)}}$

$\quad$ else

$\quad\quad$ return H_t

$\quad$ end

end

return H_T

Summary

$h \in \mathcal{H}$ — low bias; high bias

Final Points

* Stochastic gradient boosting: $h = \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n/2} \ldots$
* Test time: $H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$

$O(\log(n))$ → train error must be 0

$$T > \frac{2\log(n)}{\log\left(\frac{1}{1-4\gamma^2}\right)}$$