Lecturer: Siu On Chan *Based on Rocco Servedio's and Varun Kanade's notes*

Notes 15: AdaBoost

AdaBoost (**Ada**ptive **Boost**ing)

Fix training samples $S = \{(x^1, c(x^1)), \dots, (x^m, c(x^m))\}$ (independent samples from $\mathrm{EX}(c, \mathcal{D})$)

Fix current distribution $\mathcal{D}_t$ over S

Suppose current hypothesis h_t has error $\varepsilon \leqslant \frac{1}{2} - \gamma$ under $\mathcal{D}_t$

Question: What should updated distribution $\mathcal{D}_{t+1}$ be?

$\mathcal{D}_{t+1}$ should force weak learner A to output hypothesis h_{t+1} to reveal information not available in h_t

Key idea: Make old hypothesis h_t have error exactly 1/2 under $\mathcal{D}_{t+1}$

Since A outputs hypothesis with advantage $\gamma > 0$ under any distribution, including $\mathcal{D}_{t+1}$

h_{h+1} is guaranteed to carry new information

Since h_t errs on ε prob. mass and is correct on $1 - \varepsilon$ prob. mass under $\mathcal{D}_t$

Multiply weight of every sample h_t errs by $\sqrt{\frac{1-\varepsilon}{\varepsilon}}/Z$ (raised)

Multiply weight of every sample h_t is correct by $\sqrt{\frac{\varepsilon}{1-\varepsilon}}/Z$ (reduced)

Z = normalization constant to keep total mass of new $\mathcal{D}_{t+1}$ at 1

Total mass that h_t errs on under $\mathcal{D}_{t+1} = \varepsilon\sqrt{\frac{1-\varepsilon}{\varepsilon}}/Z = \sqrt{\varepsilon(1-\varepsilon)}/Z$

Total mass that h_t is correct on under $\mathcal{D}_{t+1} = (1-\varepsilon)\sqrt{\frac{\varepsilon}{1-\varepsilon}}/Z = \sqrt{\varepsilon(1-\varepsilon)}/Z$ (same!)

Hence $\sqrt{\varepsilon(1-\varepsilon)}/Z = 1/2 \iff Z = 2\sqrt{\varepsilon(1-\varepsilon)}$

Multiplicative weight update algorithm, like Weighted Majority

Raise weight of samples x^i that current hypothesis errs on

Reduce weight of samples x^i that current hypothesis already good at

Weighted Majority	AdaBoost
i-th expert, $1 \leqslant i \leqslant m$	i-th sample, $1 \leqslant i \leqslant m$
t-th round	t-th run of weak PAC algorithm A
prediction of i-th expert in round t	$h_t(x^i)$
weight of i-th expert in round t	$\mathcal{D}_t(x^i)$

Question: How to combine $h_1, \dots, h_R$ into final hypothesis h?

(Weighted) majority vote!

To simplify calculations, suppose $h_t : X \to \{-1, +1\}$ (as opposed to $\{0, 1\}$)

Also assume labels $y^i \in \{-1, +1\}$ (as opposed to $\{0, 1\}$)

Define $\mathrm{sign} : \mathbb{R} \to \{-1, 1\}$ as $\mathrm{sign}(z) = 1$ if $z \geqslant 0$ and $\mathrm{sign}(z) = -1$ if $z < 0$

Output hypothesis $h(x) \stackrel{\text{def}}{=} \mathrm{sign}\left(\sum_{1 \leqslant t \leqslant R} \alpha_t h_t(x)\right)$ for some positive weights $\alpha_t > 0$

Let $f(x) = \sum_{1 \leqslant t \leqslant R} \alpha_t h_t(x)$ so that $h(x) = \mathrm{sign}(f(x))$

AdaBoost

Draw independent training samples $S = \{(x^1, y^1), \dots, (x^m, y^m)\}$ from $\mathrm{EX}(c, \mathcal{D})$

Initially set $\mathcal{D}_1$ = uniform distribution over S

Repeat $t = 1, \dots, R$ times:

  Run A on samples from $\mathrm{EX}(c, \mathcal{D}_t)$ to get hypothesis h_t

  Compute $\varepsilon_t = \mathrm{err}_{\mathcal{D}_t}(h, c)$ (empirical error under $\mathcal{D}_t$)

  Set $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$ and $Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$

  Update $\mathcal{D}_{t+1}(x^i) = \mathcal{D}_t(x^i) \cdot \exp(-\alpha_t h_t(x^i) y^i)/Z_t$

Set $f(x) = \sum_{1 \leqslant t \leqslant R} \alpha_t h_t(x)$ and output hypothesis $h(x) = \mathrm{sign}(f(x))$

If $h_t(x^i) = y^i$ (correct), then $h_t(x^i)y^i = 1$, $\exp(-\alpha_t h_t(x^i) y^i) = \exp(-\alpha) = \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}$ (reduced)

If $h_t(x^i) \neq y^i$ (mistake), then $h_t(x^i)y^i = -1$, $\exp(-\alpha_t h_t(x^i) y^i) = \exp(\alpha) = \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$ (raised)

Claim: $\frac{1}{m}|\{1 \leqslant i \leqslant m \mid h(x^i) \neq y^i\}| = \frac{1}{m}\sum_{1 \leqslant i \leqslant m} \mathbb{1}(y^i f(x^i) \leqslant 0) \leqslant \frac{1}{m}\sum_{1 \leqslant i \leqslant m} \exp(-y^i f(x^i))$

Reason: $\mathbb{1}(z \leqslant 0) \leqslant \exp(-z)$ for any $z \in \mathbb{R}$

Claim: $\frac{1}{m}\sum_{1\leqslant i\leqslant m}\exp(-y^i f(x^i)) = Z_1 Z_2 \cdots Z_R$

Reason:

$$\mathcal{D}_{R+1}(x^i) = \frac{\exp(-\alpha_R h_R(x^i) y^i)}{Z_R}\mathcal{D}_R(x^i) = \quad \text{(keep expanding } \mathcal{D}_R, \ldots, \mathcal{D}_2\text{)}$$

$$= \frac{\exp(-\alpha_R h_R(x^i) y^i)}{Z_R} \cdots \frac{\exp(-\alpha_1 h_1(x^i) y^i)}{Z_1}\mathcal{D}_1(x^i)$$

Sum over all x^i, using $\mathcal{D}_1(x^i) = \frac{1}{m}$ and $\mathcal{D}_{R+1}$ has total mass 1,

$$1 = \frac{1}{m}\sum_{1\leqslant i\leqslant m} \frac{\exp(-\alpha_R h_R(x^i) y^i)}{Z_R} \cdots \frac{\exp(-\alpha_1 h_1(x^i) y^i)}{Z_1}$$

$$Z_1 \cdots Z_R = \frac{1}{m}\sum_{1\leqslant i\leqslant m} \exp(-y^i \underbrace{(\alpha_1 h_1(x^i) + \cdots + \alpha_R h_R(x^i))}_{f(x^i)})$$

Claim: $Z_1 \cdots Z_R = \sqrt{1-4\gamma_1^2}\cdots\sqrt{1-4\gamma_R^2}$ where $\gamma_t \stackrel{\text{def}}{=} \frac{1}{2} - \varepsilon_t \geqslant \gamma$

Reason: $Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = \sqrt{2\varepsilon_t 2(1-\varepsilon_t)} = \sqrt{(1-2\gamma_t)(1+2\gamma_t)} = \sqrt{1-4\gamma_t^2}$

Previous three Claims imply that training error of h on S is

$$\frac{1}{m}|\{1 \leqslant i \leqslant m \mid h(x^i) \neq y^i\}| \leqslant \left(\sqrt{1-4\gamma^2}\right)^R < (e^{-4\gamma^2})^{R/2} \leqslant \varepsilon \qquad \text{if } R \geqslant \frac{1}{2\gamma^2}\ln\frac{1}{\varepsilon}$$

e.g. If $\varepsilon = \frac{1}{m}$, then h is correct on all of S

But our goal is to get hypothesis with small (true) error, not training error!

By Theorem in Notes13, suffices to show the following hypothesis class $\mathcal{H}_R$ has small VC dimension

$$\mathcal{H}_R = \left\{ \text{sign}\left(\sum_{1\leqslant t\leqslant R} \alpha_t h_t \;\middle|\; \alpha_t \in \mathbb{R}, h_t \in \mathcal{H} \text{ for } 1 \leqslant t \leqslant R \right) \right\}$$

Here $\mathcal{H}$ denotes the hypothesis class of weak learner A

Functions in $\mathcal{H}_R$ are (± 1 version of) centered linear threshold functions of at most R hypotheses of A

Proposition 1. *If* $\text{VCDim}(\mathcal{H}) \leqslant d$, *then* $\text{VCDim}(\mathcal{H}_R) \leqslant O(Rd\log R)$

This proposition can be proved by considering growth function (next lecture)