Weak Learnability and Boosting

PAC (Strong) Learnability

Consider a hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ for binary classification.
$\mathcal{D}$ is $\mathcal{H}$-realisable if there is $h \in \mathcal{H}$ such that $\mathbb{P}_{X,Y\sim\mathcal{D}}[h(X) = Y] = 1$.

Definition (Strong Learnability)

Algorithm $\mathcal{A}$ **PAC learns** $\mathcal{H}$ with sample complexity $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ if for any $\mathcal{H}$-realisable $\mathcal{D}$, any $(\epsilon, \delta) \in (0,1)^2$ and any $m \geq m_{\mathcal{H}}(\epsilon, \delta)$

$$\mathbb{P}_{S^m \overset{\text{iid}}{\sim} \mathcal{D}} \left\{ L_{\mathcal{D}}(h_{\mathcal{A},S}) \leq \epsilon \right\} \geq 1 - \delta.$$

Computational Complexity of Learning

- So far we have discussed only the “sample complexity” of learning
- ERM was often the learning algorithms
- ERM is an optimization algorithm
- Computational complexity of ERM was not analysed

- A hypothesis class may have small sample complexity – but computational complexity of ERM may be high

Computational Complexity of Learning

- Can we trade-off accuracy of learning to reduce computational time of learning algorithm?

- Can we use a simpler hypothesis class *B*, with a lower computational cost, to learn with a lower accuracy?

Weak Learnability

Definition (γ-Weak Learnability)

Algorithm $\mathcal{A}$ **γ-weakly learns** $\mathcal{H}$ with sample complexity $m_{\mathcal{H}} : (0,1) \to \mathbb{N}$ if for any $\mathcal{H}$-realisable $\mathcal{D}$, any $\delta \in (0,1)$ and any $m \geq m_{\mathcal{H}}(\delta)$

$$\mathop{\mathbb{P}}_{S^m \overset{\text{iid}}{\sim} \mathcal{D}} \left\{ L_{\mathcal{D}}(h_{\mathcal{A},S}) \leq \tfrac{1}{2} - \gamma \right\} \geq 1 - \delta.$$

Are weakly learnable classes PAC learnable?

Proposition

$\mathcal{H}$ is PAC learnable iff it is weak learnable.

Proof.

- If $\text{VCdim}(\mathcal{H}) < \infty$ then $\mathcal{H}$ is PAC learnable and hence weak learnable.
- If $\text{VCdim}(\mathcal{H}) = \infty$, then by the Fundamental Theorem the sample complexity at (ϵ, δ) is at least of order

$$\geq \frac{\text{VCdim}(\mathcal{H}) + \ln \frac{1}{\delta}}{\epsilon}$$

which is infinite even for $\epsilon = \frac{1}{2} - \gamma$.

□

Example: Weak Learnability

Say

$$\mathcal{H} \;=\; \{\text{Three-piece classifiers}\}$$

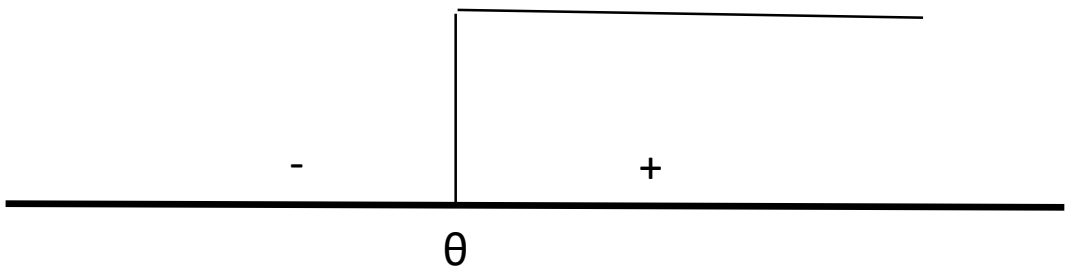
Let $\mathcal{X} = \mathbb{R}$ and let $\mathcal{H}$ be the class of 3-piece classifiers, namely, $\mathcal{H} = \{h_{\theta_1,\theta_2,b} : \theta_1, \theta_2 \in \mathbb{R}, \theta_1 < \theta_2, b \in \{\pm 1\}\}$, where for every x,

$$h_{\theta_1,\theta_2,b}(x) = \begin{cases} +b & \text{if } x < \theta_1 \text{ or } x > \theta_2 \\ -b & \text{if } \theta_1 \le x \le \theta_2 \end{cases}$$

An example hypothesis (for $b = 1$) is illustrated as follows:

Example: Weak Learnability

$\mathcal{B}$ = {Two-piece classifiers}



Let B be the class of Decision Stumps, that is, $B = \{x \mapsto \text{sign}(x - \theta) \cdot b : \quad \theta \in \mathbb{R}, b \in \{\pm 1\}\}$. In the following we show that ERM_B is a γ-weak learner for $\mathcal{H}$, for $\gamma = 1/12$.

Decision Stumps are Weak Learners of Three-Piece Hypothesis Class

For every $\mathcal{H}$-realisable $\mathcal{D}$ there is a hypothesis $f_{\mathcal{B}}^* \in \mathcal{B}$ with $L_D(f_{\mathcal{B}}^*) \leq \frac{1}{3}$.

As $\mathrm{VCdim}(\mathcal{B}) = 2$, we can agnostic(!) PAC learn $\mathcal{B}$ to accuracy $\epsilon = \frac{1}{12}$ with sample size of order $\epsilon^{-2} \ln \frac{1}{\delta}$.

With probability $1 - \delta$, get

$$L_{\mathcal{D}}(h_S) \leq L_{\mathcal{D}}(f_{\mathcal{B}}^*) + \frac{1}{12} \leq \frac{1}{3} + \frac{1}{12} = \frac{1}{2} - \frac{1}{12}$$

So we can γ-weak learn $\mathcal{H}$ for $\gamma = \frac{1}{12}$.

Efficient Learning Algorithm for Decision Stumps

- Decision Stumps can be (efficiently) learned in O($m$log$m$) time
  - m – number of samples
  - Assume realizability
- Idea:
- Sort the training examples
- Find the boundary between the classes - θ

Boosting a Weak Learner

Idea: perhaps ERM for $\mathcal{B} \subseteq \mathcal{H}$ is a weak learner for $\mathcal{H}$.
Can we **boost** an **efficient** weak learner for $\mathcal{H}$ to an **efficient** strong learner for $\mathcal{H}$?

AdaBoost

In particular, do not want to assume knowledge of γ up front.

AdaBoost instead **computes** the empirical error:

$$\epsilon_t = \sum_{i=1}^{m} w_t^i \mathbf{1}\{h_t(x_i) \neq y_i\}$$

AdaBoost

Fix

- Aggregating Algorithm
- A γ-weak learner $\mathcal{W}$ for $\mathcal{H}$.
- A sample $S = (x_i, y_i)_{i=1}^m$.

AdaBoost

input:

training set $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$

weak learner WL

number of rounds T

initialize $\mathbf{D}^{(1)} = (\frac{1}{m}, \ldots, \frac{1}{m})$.

for $t = 1, \ldots, T$:

invoke weak learner $h_t = \mathrm{WL}(\mathbf{D}^{(t)}, S)$

compute $\epsilon_t = \sum_{i=1}^{m} D_i^{(t)} \mathbb{1}_{[y_i \neq h_t(\mathbf{x}_i)]}$

let $w_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$

update $D_i^{(t+1)} = \frac{D_i^{(t)} \exp(-w_t y_i h_t(\mathbf{x}_i))}{\sum_{j=1}^{m} D_j^{(t)} \exp(-w_t y_j h_t(\mathbf{x}_j))}$ for all $i = 1, \ldots, m$

output the hypothesis $h_s(\mathbf{x}) = \mathrm{sign}\left(\sum_{t=1}^{T} w_t h_t(\mathbf{x})\right)$.

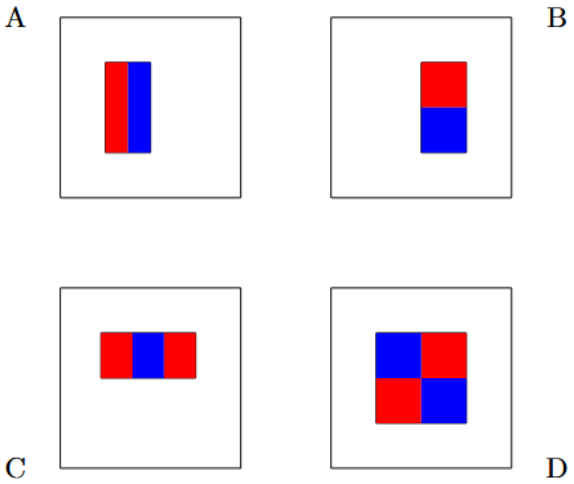
AdaBoost Results

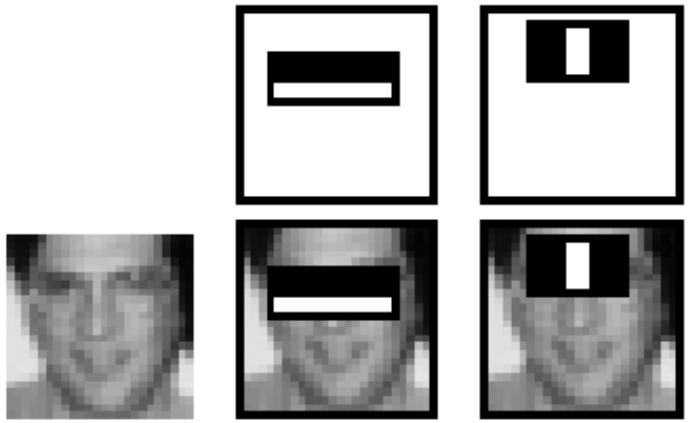
Theorem

Suppose $\mathcal{W}$ γ-weak learns $\mathcal{H}$, i.e. $\epsilon_t \leq \frac{1}{2} - \gamma$. Then the training error after T rounds of AdaBoost is at most

$$L_S(h_s) \leq e^{-2\gamma^2 T}.$$

Viola-Jones Face Detector





Each hypothesis in the base class is of the form $h(x) = f(g(x))$, where f is a decision stump hypothesis and $g : \mathbb{R}^{24,24} \to \mathbb{R}$ is a function that maps an image to a scalar. Each function g is parameterized by

- An axis aligned rectangle R. Since each image is of size 24×24, there are at most 24^4 axis aligned rectangles.
- A type, $t \in \{A, B, C, D\}$. Each type corresponds to a mask, as depicted in Figure 10.1.