EECS598: Prediction and Learning: It's Only a Game Fall 2013

Lecture 10: Boosting and Perceptron Algorithms

Prof. Jacob Abernethy *Scribe: Abhinav Sinha*

Recap

We discussed two algorithms for solving zero-sum games, defined by a pay-off matrix $M \in [0,1]^{n\times m}$.

1^{st} algorithm

Start with the uniform distribution in Δ_n i.e. $\underline{p}^1 = \langle \frac{1}{n}, \cdots, \frac{1}{n} \rangle$.

Update using

$$\underline{p}^{t+1} = \underline{p}^t \cdot \frac{\exp(-\eta M \cdot \underline{q}^t)}{Z_{t+1}} \qquad \underline{q}^{t+1} = \underline{q}^t \cdot \frac{\exp(\eta \underline{p}^t \cdot M)}{\bar{Z}_{t+1}}$$

where $\exp(\cdot)$ of a vector is just the point-wise exponential of each element of the vector and $Z_{t+1}, \bar{Z}_{t+1}$ are the normalisation factors.

2^{nd} algorithm

The only modification from above is using sequential best-response

$$\underline{q}^{t+1} = arg \max_{\underline{q}\in\Delta_m} \; \underline{p}^{t+1} M \underline{q}$$

Fact 10.1. *For both algorithms, the average strategies*

$$\frac{1}{T}\sum_{t=1}^{T} \underline{p}^t \quad and \quad \frac{1}{T}\sum_{t=1}^{T} \underline{q}^t$$

are 2ε–Nash equilibrium.

Definition 10.2. *p, q are ε–NE if*

$$p^T M q' - \varepsilon \le \underline{p}^T M \underline{q} \le p'^T M \underline{q} + \varepsilon \quad \forall \; \underline{p}', \underline{q}'$$

11 Follow the Leader (FTL)

At every time, put all the weight on the least cumulative loss expert

$$p^{t+1} = arg \min_{\underline{p}} \; \underline{p} \cdot \sum_{s=1}^{t} \ell^s$$

Question Why is FTL bad?

Answer through example: Take the loss sequence as

$$\text{expert 1} \;\; - \;\{0.5, 0, 1, 0 \ldots\} \qquad \text{expert 2} \;\; - \;\{0, 1, 0, 1 \ldots\}$$

then FTL will suffer a loss $\geq T - 1$ up to round T when each expert has only suffered a loss of $\frac{T+1}{2}$. This gives linear regret.

OPEN PROBLEM What if we try to find ε–NE with FTL?

(This was originally considered as a natural way to get at NE and the corresponding approach was known as "Fictitious play")

It was shown by Robinson (1956) - Convergence to ε–NE is at rate $\mathcal{O}(\varepsilon^{-(n+m)})$.

Karlin Conjecture: $\mathcal{O}(\frac{Poly(n,m)}{\varepsilon^2})$ is achievable with FTL.

12 Boosting

Input space $\mathcal{X}$ and labels $\left\{c : \mathcal{X} \to \{0, 1\}\right\}$. We have a weak hypothesis $h : \mathcal{X} \to \{0, 1\}$, $h \in \mathcal{H}$. Given a parameter $\gamma > 0$, we take the Weal Learning assumption,

Assumption 12.1 (WLA$_\gamma$)**.** *$\forall\, p \in \Delta(\mathcal{X})$, $\exists\, h_j \in \mathcal{H}$ such that*

$$\mathbb{P}_p\left(h_j(x) = c(x)\right) \geq \frac{1}{2} + \frac{\gamma}{2} \quad \Leftrightarrow \quad \mathbb{P}_p\left(h_j(x) \neq c(x)\right) \leq \frac{1}{2} - \frac{\gamma}{2}$$

If we assume $|\mathcal{X}| = N$ and $|\mathcal{H}| = H$ are finite and define matrix $M \in \{-1, 1\}^{N \times H}$ as

$$M_{ij} = \begin{cases} +1 & \text{if} \quad h_j(x_i) \neq c_j(x_i) \\ -1 & \text{otherwise} \end{cases}$$

Then WLA$_\gamma$ is equivalent to

$$p^T M e_j \leq -\gamma$$

Strong Learning

For all $x_i \in \mathcal{X}$ there exists a $\underline{q} \in \Delta(\mathcal{H})$ such that

$$\mathbb{P}_{h \sim \underline{q}}[h(x_i) = c(x_i)] \geq \mathbb{P}_{h \sim \underline{q}}[h(x_i) \neq c(x_i)] \quad \Leftrightarrow \quad e_i^T M \underline{q} < 0$$

We already know that Weak Learning implies Strong Learning using Strong Duality $\underline{p}^T M e_j \leq -\gamma \Rightarrow e_i^T M \underline{q} \leq -\gamma < 0$.

Boosting by Majority

Start with $p^1 = \langle \frac{1}{N}, \ldots, \frac{1}{N} \rangle \in \Delta(\mathcal{X})$.

For $t = 1, 2, \ldots$, for each $\underline{p}^t$ find $h_t \in \mathcal{H}$ such that

$$\mathbb{P}_{x \sim p^t}\left[h_t(x) \neq c(x)\right] \leq \frac{1}{2} - \frac{\gamma}{2}$$

(using WLA$_\gamma$ there will be at least one).

Update according to,

$$p_i^{t+1} = p_i^t \frac{\exp\left(\eta(-1)^{\mathbb{1}[h_t(x_i)=c(x_i)]}\right)}{Z_t}$$

Finally return

$$\underline{\hat{q}}^T = \frac{1}{T}\sum_{t=1}^{T} \underline{q}^t$$

where $\underline{q}^t$ puts weight 1 on h_t.

We already know that $\underline{\hat{q}}^T$ will be ε_T–NE, so

$$\forall\ i \quad e_i^T M \underline{q} \leq \text{Value of Game} + \frac{\text{Regret}}{T} \leq -\gamma + \sqrt{\frac{\log n}{T}}$$

So if $T \geq \frac{\log n}{\gamma^2}$ then

$$\mathbb{P}[\text{incorrect}] < \mathbb{P}[\text{correct}]$$

which gives Strong Learning.

Diagram representing decision boundaries through various iterations of ADABOOST is uploaded on the course website.

13 Perceptron Algorithm (Linear Online Prediction)

We observe a sequence $(\underline{x}^1, y^1), \ldots, (\underline{x}^T, y^T) \in \mathbb{R}^d \times \{-1, 1\}$ and we would like to find a weight vector $\underline{w}$ such that

$$\text{sgn}(\underline{w} \cdot \underline{x}^t) = y^t \quad \forall\ t$$

This weight vector will give us a separating hyperplane between the set of negative and positive data points.

Perceptron Algorithm Start with $\underline{w}^1 = \bar{0} \in \mathbb{R}^d$.

For $t = 1, \ldots, T$, predict

$$\hat{y}^t = \mathrm{sgn}(\underline{w}^t \cdot \underline{x}^t)$$

If prediction is correct i.e. $\hat{y}^t = y^t$ then don't change weights $\underline{w}^{t+1} = \underline{w}^t$. Otherwise update weights as

$$\underline{w}^{t+1} = \underline{w}^t + y^t \underline{x}^t$$

Definition 13.1. *For any $\underline{w}$ that correctly classifies $\{(\underline{x}^1, y^1)\}_{t=1}^T$, the **margin** of $\underline{w}$ is the largest $\gamma > 0$ such that $y^t(\underline{w} \cdot \underline{x}^t) \geq \gamma \ \forall\ t$.*

To make above a proper definition, assume $\|x\|_2, \|w\|_2 \leq 1$.

Theorem 13.2. *Assuming there exists a $\underline{w}^\star$ with margin γ, the number of mistakes made by the Perceptron algorithm is less than γ^{-2}.*

Let M_t be the mistakes up to round t. We will prove the theorem using the following claims

Claim 13.3 (a). $\underline{w}^t \cdot \underline{w}^\star \geq \gamma M_t$

Claim 13.4 (b). $\|\underline{w}^t\|^2 \leq M_t$

Proof of Claim (a). We will use induction on the rounds. If there is no mistake then it is trivial, so assume we are on a mistake round.

$$\hat{y}^t \neq y^t \quad \Rightarrow \quad \underline{w}^{t+1} \cdot \underline{w}^\star = (\underline{w}^t + y^t \underline{x}^t) \cdot \underline{w}^\star \geq \gamma M_t + \gamma = \gamma(M_t + 1) = \gamma M_{t+1}$$

So using induction we are done. □

Proof of Claim (b). We use induction again, and for non-mistake round it is trivial so we consider a mistake round

$$\|w^{t+1}\|^2 = \|w^t + \underline{x}^t y^t\|^2 = \|w^t\|^2 + \|\underline{x}^t y^t\|^2 + 2w^t \cdot \underline{x}^t y^t \leq M_t + 1 + \underbrace{2w^t \cdot \underline{x}^t y^t}_{-ve} < M_t + 1 = M_{t+1}$$

□

Proof of Theorem 13.2. Now with the two claims, we have

$$\gamma M_t \leq \underline{w}^T \cdot \underline{w}^\star \leq \|\underline{w}^T\| \cdot \|\underline{w}^\star\| \leq \sqrt{M_T} \cdot 1 = \sqrt{M_T} \quad \Rightarrow \quad \gamma \leq \frac{1}{\sqrt{M_T}} \quad \Leftrightarrow \quad M_T \leq \frac{1}{\gamma^2}$$

□