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

Lecture 9: Applications of Minimax: LP and Boosting

Prof. Jacob Abernethy *Scribe: Chansoo Lee*

Announcements

- Lecture on 10/16 is rescheduled to 9:30AM-11AM. Location TBA.

9.1 Minimax Theorem Review

Minimax Theorem states

$$\min_{\mathbf{p}} \max_{\mathbf{q}} \mathbf{p}^T M \mathbf{q} = \max_{\mathbf{q}} \min_{\mathbf{p}} \mathbf{p}^T M \mathbf{q}.$$

It means

$$\exists\ \mathbf{p}^* \ \forall\ \mathbf{q} \ \text{ s.t. } \ {\mathbf{p}^*}^T M \mathbf{q} \leq V \quad \text{and} \quad \exists\ \mathbf{q}^* \ \forall\ \mathbf{p} \ \text{ s.t. } \ \mathbf{p} M \mathbf{q}^* \geq V$$

Definition 9.1 (ϵ-Optimal)**.** *The mixed strategies* $\mathbf{p}$ *and* $\mathbf{q}$ *are* ϵ*-optimal if there exists* V *such that for all* p' *and* q'*,*

$$\mathbf{p}^T M \mathbf{q}' \leq V + \epsilon \ \ \textit{and} \ \ \mathbf{p}'^T M \mathbf{q} \geq V - \epsilon$$

9.2 Approximately Solving a Linear Program

Definition 9.2 (Linear Program)**.** *A Linear program is a problem that can be expressed in the canonical form:*

$$\begin{array}{ll} \textit{maximize} & \mathbf{c}^T \mathbf{x} \\ \textit{subject to} & {\mathbf{a}^j}^T \mathbf{x} \leq b^j \quad \forall\ j \in I \quad \textit{and} \\ & x_i \geq 0 \qquad \forall\ i \in [n] \end{array}$$

Lemma 9.3. *Without loss of generality, we can make the following modifications to the canonical form:*

(i) Add an additional constraint that $\mathbf{x} \in \triangle_n$ *(due to linearity)*

(ii) Assume $b_j = 0\ \forall j \in I$*, by spreading* b_j *into* $\mathbf{a}^j$ *using (i).*

Definition 9.4 (Linear Program)**.** *An alternate form for linear programs is:*

$$\begin{array}{ll} \textit{maximize} & d \\ \textit{subject to} & {\mathbf{a}^j}^T \mathbf{x} \leq 0 \ \ \forall\ j \in I, \ \ \mathbf{x} \in \triangle_n, \ \ \textit{and} \\ & \mathbf{c}^T x \geq d. \end{array}$$

We will use the above definition unless noted otherwise.

Definition 9.5 (Feasibility)**.** *An LP is feasible if there exists* $\mathbf{x} \in \triangle_n$ *that satisfies all the constraints.*

Given a feasibility checker, we can perform a binary search on d within the interval $[\min(\mathbf{c}), \max(\mathbf{c})]$ (where the min and max are taken over the coordinates of $\mathbf{c}$) to find an ϵ-optimal solution to LP with $\log(1/\epsilon)$ added time complexity.

Goal: Find an algorithm $\mathcal{A}$ such that for any feasibility checker, $\mathcal{A}$ returns $\mathbf{x} \in \triangle_n$ such that $\mathbf{a}_j^T \mathbf{x} \leq \epsilon$ for all $j \in I$, or INFEASIBLE if there is no such $\mathbf{x}$.

Algorithm 1: EWA-LP

$\mathbf{x}^{(1)} \leftarrow (\frac{1}{n}, \ldots, \frac{1}{n})$
for $t = 1$ *to* T **do**
  $\mathbf{a}_*^{(t)} \leftarrow \arg\max_{\mathbf{a} \in \{\mathbf{a}_j : j \in I\}} \mathbf{a}^T \mathbf{x}$
  if $\mathbf{a}_*^{(t)} \mathbf{x}^{(t)} \leq \epsilon$ **then**
    return $\mathbf{x}^{(t)}$
  else
    $x_i^{(t+1)} \leftarrow \dfrac{x_i^{(t)} \exp(-\eta \mathbf{a}_i)}{\Phi^{(t+1)}}$
return INFEASIBLE

Proof. The regret bound of EWA-LP is

$$\frac{1}{T} \sum \mathbf{x}^{(t)} \mathbf{a}_*^{(t)} \leq \frac{1}{T} \text{Regret}_T$$

Suppose that LP is feasible but EWA-LP returns INFEASIBLE and $T > c^2 \log n / \epsilon^2$. Then,

$$\epsilon = \frac{1}{T}(T\epsilon) \leq \frac{1}{T} \sum \mathbf{x}^{(t)} \mathbf{a}_*^{(t)} \leq \frac{1}{T} \text{Regret}_T = c\sqrt{\frac{\log n}{T}} < \epsilon,$$

a contradiction. (The second inequality follows from the assumption that the LP is feasible). □

Remark: The number of experts in EWA-LP is independent of the number of constraints.

9.3 Boosting via Minimax Duality

Setup: Let $\mathcal{X}$ be a data space (e.g. $\mathbb{R}^d$). We have a set of hypothesis $\left\{c : \mathcal{X} \to \{0,1\}\right\}$, which contains a correct hypothesis. Let $C(x)$ be the true label $\forall\, x \in \mathcal{X}$. We want to find $\hat{c} \in c$ such that the error rate

$$P_{x \sim q}[\hat{c}(x) \neq C(x)]$$

is small for any distribution $q \in \Delta(\mathcal{X})$.

Weak Hypotheses: It is easy to find weak hypotheses $\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}$.

Example: If $\mathcal{X} \subseteq \mathbb{R}^n$, then define $H = \{h_{i,c} : i = 1, \ldots, n \text{ and } c \in \mathbb{R}\}$ where

$$h_{i,c}(x) = \begin{cases} 1 & \text{if } x_i \geq c \\ 0 & \text{otherwise} \end{cases}$$

The function $h_{i,c}$ is called a *decision stump*.

Weak Learning Assumption: For a positive constant γ, the weak learning assumption states: For any distribution $q \in \Delta(\mathcal{X})$, there exists $h \in \mathcal{H}$ such that

$$P_{x \sim q}[h(x) \neq C(x)] \leq \frac{1}{2} - \frac{\gamma}{2}$$

Question: Is there a distribution p on $\mathcal{H}$ such that the weighted majority

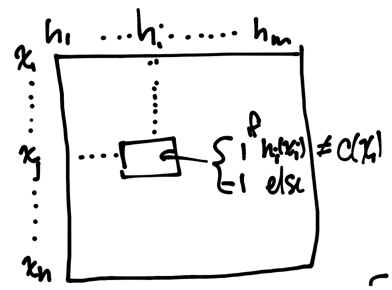
$$c_p(x) = \begin{cases} 1 & \text{if } \sum_{h \in \mathcal{H}} p(h)h(x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

achieves zero error (a.k.a. *strong learning*)?

Theorem 9.6. *Weak learning implies strong learning.*

Proof. Suppose $\mathcal{H}$ satisfies the weak learning assumption, and let $x_1, \ldots, x_n$ be the data. Let M be a $n \times |\mathcal{H}|$ matrix such that

$$M_{ij} = \begin{cases} +1 & \text{if } h_i(x_j) \neq c(x_j) \\ -1 & \text{otherwise} \end{cases}$$



[1]

The weak learning assumption states that for any $\mathbf{q} \in \triangle_n$ there exists $j \in [m]$ such that

$$P_{x \sim q}[h_j(x) \neq C(x)] \leq \frac{1}{2} - \frac{\gamma}{2}$$

This is equivalent to

$$\mathbf{q}^T M \mathbf{e}_j \leq -\gamma$$

which in turn is equivalent to

$$\min_{\mathbf{q}} \max_j \ \mathbf{q}^T M \mathbf{e}_j \leq -\gamma$$

By the minimax theorem, the above is true iff it's dual is. The dual

$$\exists \ \mathbf{p} \in \Delta_n \ \text{ s.t. } \ \mathbf{e}_i^T M \mathbf{p} \leq \gamma \ \ \forall \ i \in [n]$$

is exactly strong learning. □

[1]Diagram credit: Cat Saint Croix