AM 221: Advanced Optimization Spring 2016

Prof. Yaron Singer *Lecture 6 — February 10th*

1 Overview

In our previous lecture we saw the application of the strong duality theorem to game theory, and in particular how one can use the result to prove the minimax theorem, and then conclude the existence of mixed Nash equilibrium in zero-sum games. In this lecture we will see how duality and the minimax theorem can be applied to learning theory. In particular, we will introduce the concept of *boosting*, informally proving that classifiers that perform only slightly better than random can be turned into a classifier that is never wrong.

2 Learning Theory

We now study an application of duality to learning. More precisely we consider a classification problem over a space $\mathcal{X}$ for which we are given a set of hypothesis (possible classifiers) $\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}$. The data is sampled from an unknown distribution q over $\mathcal{X}$.

The *weak learning assumption* states that the set of hypothesis $\mathcal{H}$ is good in the following sense: for any distribution q there exists a hypothesis $h \in \mathcal{H}$ which is wrong less than half the time (i.e is better than a uniformly random classifier). Formally:

$$\exists \gamma,\ \forall \mathbf{q},\ \exists h \in \mathcal{H},\ \mathbb{P}_{x \sim \mathbf{q}}[h(x) \neq c(x)] \leq \frac{1}{2} - \frac{\gamma}{2}$$

where $c(x)$ is by definition the correct class (the true answer) of $x \in \mathcal{X}$.

3 Boosting

Surprisingly the weak learning assumption implies something much stronger: it is possible to combine the classifiers in $\mathcal{H}$ to construct a classifier which is always right. This is known as *strong learning*. This is a consequence of the minimax theorem.

Theorem 1. *Let $\mathcal{H}$ be a set of hypothesis satisfying the weak learning assumption, then there exists a distribution p on $\mathcal{H}$ such that the weighed majority classifier:*

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

is always correct, i.e, *$c_{\mathbf{p}}(x) = c(x)$ for all $x \in \mathcal{X}$.*

Proof. All we have to do is to find the probability $\mathbf{p}$ distribution, that is, weights to assign to each classifier $h \in \mathcal{H}$. Let us define the matrix $M \in \{-1, +1\}^{|\mathcal{X}| \times |\mathcal{H}|}$ such that M_{ij} is $+1$ when classifier h_j is wrong on data x_i and -1 otherwise.

Note that weak learning can be written as:

$$\sum_{i=1}^{|\mathcal{X}|} q_i \delta_{h_j(x_i) \neq c(x_i)} \leq \frac{1}{2} - \frac{\gamma}{2} \tag{1}$$

where $\delta_{h_j(x_i) \neq c(x_i)}$ is $+1$ when $h_j(x_i) \neq c(x_i)$ and 0 otherwise. We have $M_{ij} = 2\delta_{h_j(x_i) \neq c(x_i)} - 1$, hence (1) can be written more compactly as:

$$\mathbf{q}^\intercal M e_j \leq -\gamma,$$

where e_j is the j-th basis vector of $\mathbb{R}^{|\mathcal{H}|}$. In the same manner of the proof of the minimax theorem, we know that we can find such a j for all $\mathbf{q}$, i.e:

$$\min_{\mathbf{q}} \max_{j} \mathbf{q}^\intercal M e_j = \min_{\mathbf{q}} \max_{\mathbf{p}} \mathbf{q}^\intercal M \mathbf{p} \leq -\gamma$$

where $\mathbf{q}$ is a distribution on $\mathcal{X}$ and p is a distribution on $\mathcal{H}$ and where the equality follows from the fact that the basis vectors are the basic feasible solutions of the standard simplex. By the minimax theorem:

$$\max_{\mathbf{p}} \min_{\mathbf{q}} \mathbf{q}^\intercal M \mathbf{p} = \max_{\mathbf{p}} \min_{i} e_i^\intercal M \mathbf{p} \leq -\gamma$$

i.e, there exists $\mathbf{p}$ a distribution over $\mathcal{H}$ such that for all i, $(M\mathbf{p})_i \leq -\gamma$. This implies that the weighted classifier defined in the statement of the theorem is always correct. □