CS6781 - Theoretical Foundations of Machine Learning

Lecture 10: Representation Independent Hardness

February 20, 2020

Lecturer: Nika Haghtalab *Readings: Daniely [2016]*
Scribe: Raunak Kumar

1 Overview

In the last lecture we introduced two types of hardness results in learning: hardness of representation-dependent or “proper" learning, and hardness of representation-independent or “improper" learning. We also proved a representation-dependent hardness result, namely, that proper learning of intersection of two halfspaces in the realizable setting is hard. Today we will prove another representation-dependent hardness result (the hardness of properly agnostically learning a halfspace) and briefly go over a representation-independent hardness result (the hardness of improperly agnostically learning any “reasonable” halfspace).

2 Hardness of Proper Agnostic Learning of Halfspaces

Properly learning a halfspace in the realizable setting can be done in polynomial time. In lecture 2 we developed an efficient linear programming solution for this problem. We will show that in absence of the realizability assumption (in the agnostic setting) it is hard to properly learn halfspaces.

Recall that in the last lecture we took a two-step approach to showing the hardness of properly learning intersection of two halfspaces in the realizable setting. We first showed hardness of learning in the consistency model and then showed that this implies hardness in the realizable PAC setting. We adopt a similar strategy in this lecture, and only focus on the hardness of performing empirical risk minimization (ERM). Similar to the previous lecture, this can be used to show that proper agnostic learning of halfspaces is hard.

Theorem 2.1. *Given a class of hypothesis $\mathcal{C}$ that is the set of halfspaces in n dimension, it is NP-hard to find $h \in \mathcal{C}$ that has the smallest error on input set S.*

Definition 2.2 (Maximum Independent Set Problem)**.** Given a graph $G = (V, E)$, an independent set $V' \subset V$ is a set such that no two vertices in V' have an edge between them, i.e., $\forall u, v \in V', (u, v) \notin E$. The maximum independent set (MIS) problem is finding a independent set of maximum cardinality. This is an NP-Hard problem.

In order to prove theorem 2.1 we are going to show that any instance of the MIS problem can be mapped onto an instance of the problem of properly agnostically learning a halfspace. We are given a graph $G = (V, E)$. Without loss of generality, let $V = [n]$. We will map this into empirical risk minimization over halfspaces in n dimension. Construct the input (training set) for the learning algorithm as

$$\mathcal{S} = \big\{(\vec{e_i}, +) : i \in [n]\big\} \cup \big\{(\vec{e_i} + \vec{e_j}, -) : (i, j) \in E\big\} \cup \big\{(\vec{0}, -)\big\}.$$

where $\vec{e_i}$ is the vector with a one in the i'th index and zeros elsewhere. Notice that we are labelling the origin as -1. We need to show that finding an MIS in G is equivalent to performing ERM on $\mathcal{S}$ over the class of halfspaces. Let $\overline{\text{ERM}}$ be an algorithm that returns a halfspace minimizes error on positive instances and makes no mistakes on negative instances. Instead of directly showing that ERM over the class of halfspaces in the agnostic setting is hard, we will first show that $\overline{\text{ERM}}$ in this setting is hard and then show that this implies that ERM is hard.

Claim 2.3. *Let $G = ([n], E)$ and let the training set $\mathcal{S}$ be as defined above. Then, there is an independent set $P \subset [n]$ in G of size $|P|$ if and only if there is a halfspace that makes no mistakes on negative instances and $n - |P|$ mistakes on positive ones in $\mathcal{S}$.*

Note that the above claim and the definition of $\overline{\text{ERM}}$ imply that $\overline{\text{ERM}}$ gives us an MIS. Therefore, proving the claim will show hardenss of $\overline{\text{ERM}}$. The main idea in the proof of the claim is that halfspaces divide the space to two convex regions: That is when $\vec{0}$ is negative and $\vec{e_i}$ and $\vec{e_j}$ are positive, then $\vec{e_i} + \vec{e_j}$ is also positive.

Proof. $\Rightarrow$ direction: Given an independent set P, consider the halfspace $\vec{w} \cdot \vec{x} \geq \frac{1}{2}$, where

$$w_i = \begin{cases} +1 \text{ if } i \in P \\ -1 \text{ if } i \notin P \end{cases} .$$

We need to verify that this halfspace correctly classifies all negative instances and makes exactly $n - |P|$ mistakes on positive instances. Since $\vec{0} \cdot \vec{w} = 0 < \frac{1}{2}$, $(\vec{0}, -)$ is classified correctly. For each negative instance $(\vec{e_i} + \vec{e_j}, -)$, note that $\vec{w} \cdot (\vec{e_i} + \vec{e_j}) = w_i + w_j$. Since P is an independent set at most one of i and j is in P. This implies that at most one of w_i and w_j is +1. Therefore, $w_i + w_j \leq 0$, and $(\vec{e_i} + \vec{e_j}, -)$ is classified correctly. Thus, this halfspace classifies correctly classifies all negative instances. For each positive instance $(\vec{e_i}, +)$, note that $\vec{w} \cdot \vec{e_i} = w_i$. This is equal to $+1$ if $i \in P$, -1 otherwise. Therefore, the number of mistakes on positive instances is equal to $n - |P|$.
$\Leftarrow$ direction: Given a halfspace $\vec{w} \cdot \vec{x} \geq w_0$, where $\vec{w} = (w_1, w_2, \dots, w_n)$, let

$$P = \{i \mid w_i \geq w_0\}.$$

We have to show that if the halfspace makes no mistakes on the negative points then P is an independent set. Since the halfspace correctly classifies all negative instances, we have $\vec{0} \cdot \vec{w} = 0 < w_0$ and $\vec{w} \cdot (\vec{e_i} + \vec{e_j}) = w_i + w_j < w_0$. The second inequality implies that at most one of w_i and w_j can be $\geq w_0$. Given our construction of P at most one of i and j is in P. This shows that P is an independent set. For every positive instance $(\vec{e_i}, +)$ we have $\vec{w} \cdot \vec{e_i} = w_i$. If $w_i \geq w_0$ then $i \in P$, otherwise $i \notin P$. Since the halfspace makes $n - |P|$ mistakes on positive points, the size of the independent set P is $|P|$. □

To see why hardness of $\overline{\text{ERM}}$ also implies hardness of ERM, consider the following. Given the above training set $\mathcal{S}$, construct a new training set $\mathcal{S}'$ which contains $|\mathcal{S}|$ copies of each negative instance in $\mathcal{S}$ and one copy of each positive instance in $\mathcal{S}$. Observe that making a mistake on even a single negative instance is worse than making a mistake on all positive instances. Thus, ERM on $\mathcal{S}'$ will make no mistakes on negative instances and minimize error on positive instances, which is precisely our definition of $\overline{\text{ERM}}$. This shows that in the agnostic setting ERM over the class of halfspaces is hard and concludes the proof of theorem 2.1.

In light of theorem 2.1 in the proper and agnostic setting we cannot hope to efficiently find a halfspace with minimum empirical error. However, is it possible to find a halfspace that approximately minimizes the empirical error? For instance, if ERM outputs a halfspace with error η, can we efficiently find a halfspace with error at most 2η? If η is small, say, 0.001, then a halfspace with error at most 0.002 would work reasonably well in practice. Unfortunately, the following theorem states that this is impossible.

Theorem 2.4 (Hardness of Proper Agnostic Learning of Halfspaces, Guruswami and Raghavendra [2009])**.** *For arbitrary constants η and α, even if there is a halfspace with some constant error η, finding a halfspace with error $\frac{1}{2} - \alpha$ in the proper (representation-dependent) setting is NP-hard.*

Remark 2.5. We can get an error of $\frac{1}{2}$ with random coin flips. This theorem says that beating the predictions of a random coin flips by a bias of more than a constant is hard.

Remark 2.6. The above theorem states that we cannot find a good halfspace for every distributions. However, not all distributions are bad and we might be able to learn halfspaces that perform very well in practice on real-life distributions.

3 Hardness of Improper Learning of Halfspaces

In this section we will briefly discuss a representation-independent hardness result.

Theorem 3.1 (Daniely [2016])**.** *(Under some assumptions of hardness) For any constant η, even if there is a halfspace with a constant error η, it is hard to learner a predictor of any representation with error $\frac{1}{2} - \frac{1}{n^c}$ for any constant $c > 1$.*

Remark 3.2. In contrast to theorem 2.4 the above theorem states that it is hard to "predict" with low error without restricting to predictors that are halfspaces. The unrestricted "predict" part of the theorem corresponds to "improper" or "representation-independent" learning.

The proof of this theorem is detailed and the interested reader is encouraged to refer to Daniely [2016]. In this lecture, we only explain at high level how one can think about hardness of improper learning.

The main high level idea is that there are problems where distinguishing between something entirely random and something that has structure is hard. The notion of (lack of) structure is made precise by the following definition of scattered.

Definition 3.3 (Scattered)**.** A training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \{-1, 1\}^n \times \{-1, 1\}$ is called scattered if each y_i is a coin flip, i.e., it is equal to ± 1 with probability $\frac{1}{2}$. In other words, the labels have no correlation with the instances.

Let's call a problem $(\text{HYP}^{\eta}_{m}, \mathcal{H})$ “easy" if there exists an efficient randomized algorithm A such that given an input $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \{-1, 1\}^n \times \{-1, 1\}$ the algorithm satisfies the following:

- If there is a $h \in \mathcal{H}$ with small error $\text{err}_S(h) \leq \eta$ then $\Pr[A(S) =$ “almost realizable”$] \geq \frac{3}{4}$, where the probability is over the randomness in A.
- If S is scattered then with probability $1 - O\left(\frac{1}{n}\right)$ over the choice of labels, we have $\Pr[A(S) =$ “scattered”$] \geq \frac{3}{4}$ where the probability is over the randomness in A.

Lemma 3.4. *Assume that* $\text{HYP}^{\eta}_{n^a}$ *is hard for every* $a > 0$. *Even for distributions where there is an* $h \in \mathcal{H}$ *such that* $\text{err}_{\mathcal{D}}(h) \leq \eta$, *there is no efficient improper learning algorithm with error* $\text{err}_{\mathcal{D}}(h) \leq \frac{1}{2} - \frac{1}{n^c}$.

Proof. Roughly speaking, an algorithm is efficient if it does not access too many bits or the output of the algorithm is not a function of more than polynomial number of bits. In order to prove this theorem consider the following reduction. Let $\mathcal{L}$ be the improper learning algorithm that uses at most $n^{c'}$ bits for some constant c'. Define $q(n) = n^{2c+2c'}$. Let $m = q(n)$ and on input $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ do the following:

- Define $\mathcal{D}$ to be the uniform distribution on S. Call $\mathcal{L}$ on $\mathcal{D}$.
- Let h be the hypothesis returned by $\mathcal{L}$. If $\text{err}_S(h) \leq \frac{1}{2} - \frac{1}{n^c}$ then output “almost realizable”. Otherwise, output “scattered”.

Why does this work?

1. Case 1: $\exists h \in \mathcal{H}$ such that $\text{err}_S(h) \leq \eta$. By the definition of $\mathcal{L}$ the hypothesis returned, h, has error $\text{err}_{\mathcal{D}}(h) \leq \frac{1}{2} - \frac{1}{n^c}$. Since $\mathcal{D}$ is the uniform distribution on S, this means that $\text{err}_S(h) \leq \frac{1}{2} - \frac{1}{n^c}$ and the output is “almost realizable”.
2. Case 2: S is scattered. For any fixed hypothesis $h \in \mathcal{H}$, using the Hoeffding bound we get that

$$\Pr_{\text{draws of scattered } S}\left[\text{err}_S(h) \leq \frac{1}{2} - \frac{1}{n^c}\right] \leq \exp\left(\frac{-2q(n)}{n^{2c}}\right).$$

   Since $\mathcal{L}$ uses at most $n^{c'}$ bits there at most $2^{n^{c'}}$ hypotheses that could be returned by $\mathcal{L}$ regardless of how these hypotheses are represented . Using union bound, we get that

$$\Pr\left[\text{return } h \text{ s.t. } \text{err}_S(h) \leq \frac{1}{2} - \frac{1}{n^c}\right] \leq 2^{n^{c'}} \exp\left(\frac{-2q(n)}{n^{2c}}\right).$$

   Plugging in the definition of $q(n) = n^{2c+2c'}$ and simplifying yields that the right hand side is at most $2^{-n^{c'}} = 1 - O\left(\frac{1}{n}\right)$.

□

To use the above lemma, one first needs to show that the problem HYP$^{\eta}$ is hard for the class of halfspaces. At a high level Daniely [2016] proves this by assuming that HYP$^{\eta}$ is hard for the class of k-XOR functions. We encourage the interested reader to refer to Daniely [2016] for more details on the connection between the two classes.

k-XOR functions are ANDs of XORs, where each XOR has up to k variables in it. This is an example of a 2-XOR function.

$$(x_1 \oplus x_2) \wedge \neg(x_2 \oplus x_3) \wedge (x_1 \oplus x_4) \wedge \neg(x_2 \oplus x_4)$$

One can think about each of the XORs as a constraints expressed by one of m samples in the data set. As we say in Homework 1, it is easy to find a satisfying assignment if one exits. However, when no satisfying assignment exists, it is hard to find an assignment that satisfies the most number of XORs. The assumption Daniely [2016] uses is that when $m \leq n^{\sqrt{k}\log(k)}$ solving HYP$^{\eta}_{m}$ is hard for the class of k-XORs.

References

Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 105–117, New York, NY, USA, 2016. Association for Computing Machinery.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.