Learning as Compression

Daniel Khashabi

Fall 2016
Last Update: October 19, 2016

1 Introduction

Here we will give an information definition of *compression scheme*. Consider the training set as the ordered sequence:

$$T = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

We want to analyze the compression bound for an algorithm $\mathcal{A}$ which takes the training data and returns a hypothesis function $h \in \mathcal{H}$. Define a subsequence of this training set to be T_I where $I \in [m]$. The smallest I such that

$\mathcal{A}(T) = \mathcal{A}(T_I)$ (i.e. the two algorithms learn 'the same hypothesis function on the training data)

is called *compression set* for T. For example, for perceptron/Winnow algorithm, the comppression set is their mistake bound (number of instances they have to visit until they learn the target class).

For any subset of instances I define the empirical risk:

$$\hat{R}_{T_I}(h) \leq \frac{1}{|I|} \sum_{i \in I} \ell(h(X_i), Y_i)$$

Often times we minimize the empirical risk on trining data in order to find the target hypothesis. Often times the compression scheme is defined in terms of minimizing the empirical risk on a subset I of the training instances. For the *realizable* case the compression scheme is defined as choosing such subset, such that $\hat{R}_{T_I}(h) = 0$. For the *unrealizable/agnostic* case, one can give slightly softer condition:

$$\hat{R}_{T_I}(h) \leq \hat{R}_T(h)$$

The following theorem shows that if we find a hypothesis perfect on the subset I (i.e., minimizer of the empirical risk on the subset), the actual risk of the resulting function is not too bad, for any random draw of training data.

Intuitively compression scheme for the realizable case implies the existence of compression scheme for the agnostic setting:

Lemma 1.1. *Let $\mathcal{H}$ be a hypothesis class for binary classification and assume that it has compression scheme of size k in the realizable case. Then it has a compression scheme of size for the unrealizable case as well.*

It is worth stressing that this result holds only for *binary classification.* The claim doesn't hold in general (See for more details in [**?**]). Here is the proof:

Proof. For a training dataset in the agnostic case, suppose there is a hypothesis $h \in \mathcal{H}$ consider all the instances on which h doesn't make mistakes and use them to construct another hypothesis h'. Since this hypotheis won't make on any of these instances, in overall the empirical risk of h' is as good as h. ■

Theorem 1.2 (Compression Bound for Realizable Case)**.** *If the loss function is bounded in* $[0, 1]$*, with probablity at least* $1 - \delta$*, for any* I *such that* 1

$$\hat{R}_{T_I}(h) \leq \frac{1}{n-1}\left((l+1)\log n + \log\frac{1}{\delta}\right)$$

with l *being the size of the compression, and the probability is with respect to a random draw of* T*.*

Proof. TODO ■

References