Inefficient Learning in Abstract Learner

Ming Feng Wan
University of Toronto
ming@cs.toronto.edu

Abstract

In this paper, we examine how efficient a popular learner is at learning grammatical rules.

1 Introduction

Daniely and Shalev-Shwartz (2014) proposed a popular multiclass learner with the best possible guarantee on error to be $\Theta(\frac{\mu_{\mathcal{H}}(m)}{m})$, with m being the number of samples seen. Here $\mu_{\mathcal{H}}(m) = \max\{\text{md}(G(\mathcal{H}|_S))|S \in \mathcal{X}^m\}$, where $G(\mathcal{H}|_S)$ is an *one-inclusion hypergraph* and md stands for *maximal average degree*. The one-inclusion hypergraph is constructed by the following rule: Given a set of m unlabelled samples $S = \{x_1, ..., x_m\}$, for every $i \in [m]$ and $h \in \mathcal{H}|_S$, let $e_{i,h} \subset \mathcal{H}|_S$ be all the hypotheses in $\mathcal{H}|_S$ whose restriction to $S \setminus \{x_i\}$ equals to $h|_{S\setminus\{x_i\}}$: $h_0 \in e_{i,h}$ iff for all $j \neq i, h_0(x_j) = h(x_j)$. It is therefore true that $h_0 \in e_{i,h}$ then $e_{i,h_0} = e_{i,h}$. Now we have $E = \{e_{i,h}\}_{i\in[m],h\in\mathcal{H}|S}$ as a collection of hyperedges of a hypergraph $G = (V, E)$, and the corresponding vertex set is $V = \mathcal{H}|_S$.

Now to their definition of *maximal average degree*. Consider a hypergraph $G = (V, E)$. They only consider hypergraphs with E as an antichain (there does not exist $e_1, e_2 \in E$ such that $e_1 \subsetneq e_2$). Define the induced hypergraph, $G[U], U \subseteq V$, as the hypergraph whose vertex set is U and whose edge set is all sets $e \subseteq U$ such that $e = U \cap e_0$ for some $e_0 \in E, |e| \geq 2$, and e is maximal with respect to these conditions. The degree of a vertex $v \in V$ in a hypergraph $G = (V, E)$ is the number of hyperedges $e \in E$ that contains v and $|e| \geq 2$. The average degree of G is therefore $d(G) = \frac{1}{|V|}\sum_{v\in V} d(v)$. The maximal average degree of G is $\text{md}(G) = \max_{U\subseteq V:|U|<\infty} d(G[U])$.

2 Applying the Hypergraph

Consider a simple classification problem of active voice and passive voice with a set S of 6 sentences (Table 1). Let $\mathcal{H}$ be a classifier space that determines if the sentence is in active voice or not. Of course, there exists a $h^* \in \mathcal{H}|_S$ such that h^* always gives the correct value. Now consider two

#	**Sentence**	h^*
1	I was given	No
2	I give	Yes
3	I have given	Yes
4	I have been being given	No
5	I have been giving	Yes
6	I have been given	No

Table 1: An example dataset of active voice classification

hypothesis that belongs to $\mathcal{H}|_S$:

- h_0: A sentence is in passive voice (not in active voice) if any variation of *be* (such as being, been, was) is in front of the main verb.
- h_1: A sentence is in active voice if the sentence does not contain the word *given*.

h_0 is true expect for sentence 5, and h_1 is true expect for sentence 3. We can construct a hypergraph of a multiclass learner defined in the previous section. As one can clearly see, the maximal average degree of Figure 1 is not zero. $\mathcal{H}$ is therefore expected to make errors in prediction given S.

Maybe the active voice passive voice problem is too complex with too few samples. Now let us consider classifier space $\mathcal{H}'$ with a simpler classification problem: determine if any of the previous sentences is in perfect tense. Here $h'^* \in \mathcal{H}'|_S$ can simply be 'a sentence is in perfect tense if the sentence contains the word have'.

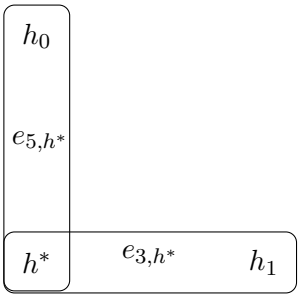

Figure 1: Hypergraph based on $\mathcal{H}|_S$

#	**Sentence**	h'^*
1	I was given	No
2	I give	No
3	I have given	Yes
4	I have been being given	Yes
5	I have been giving	Yes
6	I have been given	Yes

Table 2: An example dataset for perfect tense classification

Now consider two new hypothesis that belongs to $\mathcal{H}'|_S$:

- h_2: A sentence is in perfect tense if it does not contain the words *given* and *giving*.
- h_3: A sentence is in perfect tense if it does not contain the word *was*.

h_2 is true expect for sentence 1, and h_3 is true expect for sentence 2. Again, we can construct a hypergraph defined in the previous section with maximal average degree greater than zero (Figure 2). $\mathcal{H}'$ is therefore also expected to make errors in prediction given S. Now consider a new classifier class H'' that attempts to classify both perfect tense and active voice ($|\mathcal{Y}''| = 4$). We can again construct a hypergraph with even greater maximal average degree (Figure 3).

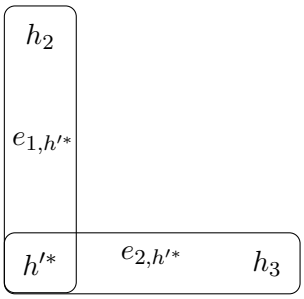

Figure 2: Hypergraph based on $\mathcal{H}'|_S$

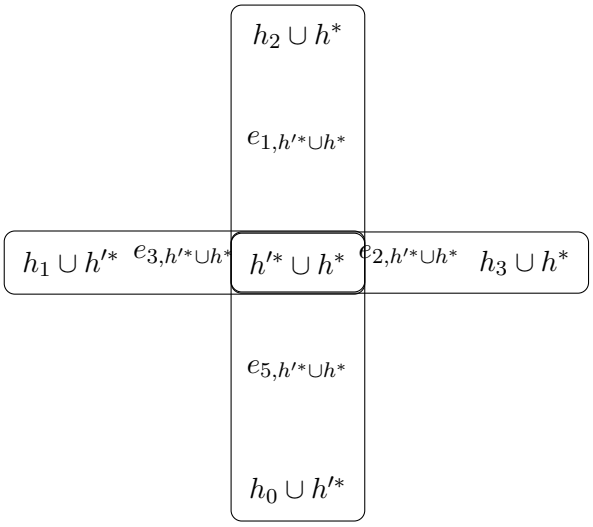

Figure 3: Hypergraph based on $\mathcal{H}''|_S$

All of the three previous problems above can be easily solved by a parser with FSM-like grammar rule; this creates an obvious problem that hypothesis derived from a FSM can be potentially infinite. Consider the problem determining whether the sentence is grammatical or not. Assume the sentence have a lexicon of at most three words: *you, tell, to*. An example sentence would be *you tell you to tell you to tell you*. A simple FSM can be used to solve this problem without error. However, this is not the case for multiclass learner. In fact, it is probably not hard to come up with datasets that have maximal average degree greater than zero, such as the one in Table 3. In table 3, Θ derived from the hypothesis 'a sentence is grammatical if it does not start with *tell*' is true without sentence 5, and the hypergraph constructed would have maximal average degree greater than zero.

#	**Sentence**	h^*
1	you tell you	Yes
2	you tell to	No
3	you to tell you	No
4	you tell you to tell you	Yes
5	tell	Yes
6	tell to	No

Table 3: An example dataset for grammaticality

A believer in *poverty of the stimulus* would argue that the ideal classifier class need to be way more restrictive than the classifier class given above. In particular, we need have a classifier class that only look at the presence of *have* as an auxiliary verb for classifying perfect tense. However, making

a classifier restrictive with syntactic information essentially makes it a parser.

3 Conclusion

In this paper, we showed that a popular abstract learner is inefficient at learning grammatical rules.

References

Amit Daniely and Shai Shalev-Shwartz. 2014. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR.