COMS 6998-4 Fall 2017 Presenter: Xuefeng Hu
Octorber 30, 2017 Scribe: Qinyao He

Hierarchical Concept Learning

1 Introduction

It has been shown that learning arbitrary polynomial-size circuits is computationally intractable [3]. To resolve this difficulty, we can introduce a more powerful teacher that breaks the target concept into subconcepts and teaches the learner in stages.

2 PAC-learning Recap

Definition 1. *Given an instance space $\mathcal{X}$ and probability distribution $\mathbb{P}$ on that space, we say a concept c' is ϵ-approximation of concept c if*

$$\Pr_{x\sim\mathbb{P}}\left(c'(x)\neq c(x)\right)\leq\epsilon.$$

Theorem 1. *Let $\mathcal{H}$ be a finite hypothesis set, and $\epsilon,\delta\in[0,1]$. Let*

$$m=\frac{1}{\epsilon}\left(\ln|\mathcal{H}|+\ln\frac{1}{\delta}\right).$$

With probability at least $1-\delta$, every concept $c\in\mathcal{H}$ that is consistent with a size-m iid sample from a distribution $\mathbb{P}$ labeled by a target concept c^* is an ϵ-approximation of c^*.

Proof. See [3, 1]. □

2.1 Reliable PAC-learning

In the original PAC-learning setting, the learner is required to output a prediction for every input instance. In the Reliable PAC-learning setting, we allow the learner to output a prediction (0 or 1) only when it is "sure" of the prediction; it may output "unsure" in other cases.

Definition 2. *A reliable learner outputs* 1 *only on positive instances, and outputs* 0 *only on negative instances.*

The fraction of instances a reliable learner output "unsure" is the error rate. So for reliable PAC-learning, the algorithm is required to output an classifier Q which say "unsure" only on at most ϵ fraction of all samples.

3 Problem Formulation

The idea of learning subconcept hierarchically can be seen as an imitation of human learning process. For human learning, we typically do not learn a complicated concept directly. Instead, we do by gradually learning the hierarchical decomposition of the original concept and then compose them up.

Take a simple example. Let concept c denote “a good PhD applicant”. In a very simplified setting, define this concept in natural language:

> A good PhD applicant must have high GPA, GRE scores, good recommendation letters. If the student is master student, he/she must have a conference paper.

We can break this complicated concept into small pieces and combine then up. Let

$$\begin{aligned} x_1 &= \text{undergraduate} \\ x_2 &= \text{master} \\ x_3 &= \text{good recommendation} \\ x_4 &= \text{conference paper} \\ x_5 &= \text{high GPA} \\ x_6 &= \text{high GRE} \end{aligned}$$

as boolean variables, so the target concept can be represented as

$$c = x_5 \wedge x_6 \wedge x_3 \wedge \big((x_2 \wedge x_4) \vee x_1\big).$$

We can introduce a set of boolean variables y_i and construct the target concept in a hierarchical manner

$$\begin{aligned} y_1 &= x_5 \wedge x_6 \\ y_2 &= y_1 \wedge x_3 \\ y_3 &= x_2 \wedge x_4 \\ y_4 &= y_3 \vee x_1 \\ y_5 &= y_2 \wedge y_4 = c. \end{aligned}$$

We then formally define our notion of hierarchical concept learning. We consider only the case where each level of subconcept is conjunction or disjunction of two variables (both can be negative) chosen from a set L, which contains original instance attributes and previously learned subconcepts. We assume the target concept c^* to be represented as a straight-line program like the previous example and contains s lines. Let the input to be $x_1, x_2, \ldots, x_n$, and output to be y_s. The ith line of the program is of the form

$$y_i = z_{i,1} \circ z_{i,2},$$

where $\circ$ may be one of $\wedge$ and $\vee$, and every $z_{i,k}$ is either an input literal x_j or $\bar{x_j}$, or previous computed y_j or $\bar{y_j}$.

Define $V_i = \{x_1, \ldots, x_n, y_1, \ldots, y_{i-1}\}$ to be the set of instance attributes and previously learned subconcepts, and $\mathrm{EB}(V_i) = \{\text{true}, \text{false}, z_{i,1} \circ z_{i,2} | z_{i,1}, z_{i,2} \in V_i\}$ is the set of all boolean formulas that are conjuction or disjunction of two variables in V_i, where "EB" stand for "Easy Boolean".

We should examine and bound the size of $\mathrm{EB}(V_i)$. $|V_i| = n + i - 1$, the size of $\mathrm{EB}(V_i)$ is equal to choose 2 from $n + i - 1$ elements, multiplied by a factor which represent the choice of $\wedge$ and $\vee$ and either of the two variables may be negated, then added by 2 which is two trivial cases of true and false:

$$|\mathrm{EB}(V_i)| = 8\binom{n+i-1}{2} + 2 \leq 8\binom{n+s-1}{2} + 2 = K.$$

We should note that K is polynomial in n and s which are the size of the concept class.

With those notations, we could write the hierarchical concept learning procedure as follow algorithm 1

Algorithm 1 Hierarchical Concept Learning

```
Teacher break the target concept $c^*$ into subconcepts $y_1, \ldots, y_s$.
for $i = 1$ to $s$ do
    Teacher draw random i.i.d samples $\big(x = (x_1, \ldots, x_n), y_i(x)\big)$.
    Learner learns $y_i$ and tell the teacher when this stage complete.
end for
return $y_s$.
```

For the ith stage, we are about to learn y_i in space $\mathrm{EB}(V_{i-1})$. However we are only given the sample $x_1, \ldots, x_n$, and the correct value of $y_1, \ldots, y_{i-1}$ are never shown to the learner. The reason not to do so is that providing this help to the learner requires large amount of extra information which is not very satisfying. We assume at each stage i the learner has successfully learned y_i, and the teacher continue to teach y_{i+1} and will never return back to y_i.

But we may have error when learning y_i at each stage. If we use any learned y_i at stage i for future learning, we may be influenced by the error of the learned concept.

To resolve this issue we should maintain a full list of all possible candidates, including the correct concept, for any y_i, i.e. the version space. As shown previously, the size of the version space is only at most polynomial in n and s, which makes it tractable to keep all of them.

We filter $\mathrm{EB}(V_i)$ to a version space F_i, which is consistent to all given samples in ith stage,

$$\mathrm{EB}(V_i) \to F_i = \{y | y \in \mathrm{EB}(V_i), y \text{ consistent with all } m \text{ samples}\}.$$

Since the previously learned $F_1, \ldots, F_{i-1}$ have many possible candidates y_i, their values on any sample x may be different, and we only want those with correct value. We say a sample

$x_1, \ldots, x_n$ is “good” if for every previously learn $F_j, 1 \leq j < i$, all y_j in F_j take the same truth value on $x_1, \ldots, x_n$. We only use “good” examples in every stage of learning.

Definition 3. *$F = \{f_1, \ldots, f_s\}$ is a set of boolean formulas defined in terms of $x_1, \ldots, x_n$. F is coherent on $x_1, \ldots, x_n$ if $f_1(x_1, \ldots, x_n) = f_2(x_1, \ldots, x_n) = \ldots = f_s(x_1, \ldots, x_n)$.*

Note that for every previously learned F_j, for any $(x_1, \ldots, x_n)$ if F_j is coherent on $x_1, \ldots, x_n$, then the common value must be the correct value $y_j(x_1, \ldots, x_n)$, since we know that the ground truth formula is actually contained in F_j. This can be concluded by noticing that F_j is constructed by selecting consistent hypothesis in $\mathrm{EB}(V_i)$.

4 Learning Algorithm

We first give the learning algorithm for each stage i, given all previous F_j as an input to learn F_i for the ith stage. See algorithm 2

Algorithm 2 CSL Coherent Set Learner

```
Input: $F_1, \ldots, F_{i-1}, i, K, \epsilon_i, \delta_i$
Output: $F_i$
  pick $m = \frac{1}{\epsilon_i}(\ln K + \ln \frac{1}{\delta_i})$.
  ask $2m$ samples from teacher, $a_1, \ldots, a_{2m}$.
  for $a_j \in \{a_1, \ldots, a_{2m}\}$ do
    if $F_1, \ldots, F_{i-1}$ coherent on $a_j$ then
      $a_j = \big(x_1, \ldots, x_n, F_1(a_j), \ldots, F_{i-1}(a_j)\big)$
    else
      discard $a_j$
    end if
  end for
  if $m$ samples is gathered then
    $F_i = \{f \in \mathrm{EB}(V_i) | f$ consistent with all $m$ samples$\}$
    return $F_i$
  else
    return failed
  end if
```

Given CSL algorithm, we can apply this on each stage i and learn F_i. See algorithm 3

With $F_1, \ldots, F_s$ learned, for any given instance x to query, we use algorithm 4 to return the classification result.

The algorithm return “unsure” when some of $F_1, \ldots, F_s$ is not coherent on input x. From previous discussion we know that, when all $F_1, \ldots, F_s$ are coherent on x, the output y_s must be the ground truth value. So the learner is reliable.

Algorithm 3 Reliable Learner

```
  $K \leftarrow 8\binom{n+i-1}{2} + 2$
  $\epsilon' \leftarrow \frac{\epsilon}{sK}$
  $\delta' \leftarrow \frac{\delta}{2s}$
  for $i \leftarrow 1$ to $s$ do
    $F_i \leftarrow \mathrm{CSL}(i, F_1, \ldots, F_{i-1}, K, \epsilon', \delta')$
  end for
  return $(F_1, \ldots, F_s)$
```

Algorithm 4 Q(x)

```
Input: $x \in \mathcal{X}$, $F_1, \ldots, F_s$.
  for $i \leftarrow 1$ to $s$ do
    if $F_i$ coherent on $(x_1, \ldots, x_n,\ y_1, \ldots, y_{i-1})$ then
      $y_i = F_i(x)$
    else
      return unsure
    end if
  end for
  return $y_s(x)$
```

5 Theoretical Analysis

Definition 4. *Let $\mathcal{X}_i(F_1, \ldots, F_{i-1}) = \{x | \text{all } F_j \text{ coherent on } x, x \in \{0,1\}^n, 1 \le j < i\}$, say F_i is accurate, if $\forall x \in \mathcal{X}_i, f \in F_i$, $\Pr(f(x) = y_i(x)) \ge 1 - \frac{\epsilon}{sK}$.*

Lemma 1. *If all $F_1, \ldots, F_s$ are accurate, then the unsure rate of $Q(x)$ is less than ϵ.*

Proof.

$$\begin{aligned}
&\Pr(Q(x) = c(x)) \\
=&\Pr(F_1, \ldots, F_s \text{ all coherent on } x) \\
=&\Pr(F_s \text{ coherent on } x | F_1, \ldots, F_{s-1} \text{ coherent on } x) \Pr(F_1, \ldots, F_{s-1} \text{ coherent on } x) \\
=&\left(\prod_{f \in F_s} \Pr(f(x) = y_s(x)) \right) \Pr(F_1, \ldots, F_{s-1} \text{ coherent on } x) \\
\ge&\left(1 - \frac{\epsilon}{sK}\right)^K \Pr(F_1, \ldots, F_{s-1} \text{ coherent on } x) \\
\ge&\left(1 - \frac{\epsilon}{s}\right) \Pr(F_1, \ldots, F_{s-1} \text{ coherent on } x) \ge \left(1 - \frac{\epsilon}{s}\right)^s \ge 1 - \epsilon.
\end{aligned}$$

□

Lemma 2. *If $F_1, \dots, F_{i-1}$ is accurate, then at stage i, with probability at least $1 - \frac{\delta}{2s}$, m samples are drawn in CSL, and CSL doesn't return "failed".*

Proof. For $\forall x \in \mathcal{X}$,

$$\Pr(F_1, \dots, F_{i-1} \text{ coherent on } x) \geq \left(1 - \frac{\epsilon}{s}\right)^i \geq \left(1 - \frac{\epsilon}{s}\right)^s \geq 1 - \epsilon,$$

thus we conclude

$$\Pr(x \text{ not selected}, x \in \mathcal{X}) \leq \epsilon.$$

According to Hoeffding's inequality, if we treat "x not selected" as a mistake,

$$\begin{aligned}
\Pr(\text{more than } m \text{ mistakes in } 2m \text{ trials}) &\leq \mathrm{e}^{-2\cdot 2m(\frac{1}{2}-\epsilon)^2} \\
&= \mathrm{e}^{-m(1-2\epsilon)^2} = \mathrm{e}^{-\frac{1}{\epsilon'}\ln(\frac{K}{\delta'})(1-2\epsilon)^2} \\
&= \left(\frac{K}{\delta'}\right)^{-\frac{1}{\epsilon'}(1-2\epsilon)^2} = \left(\frac{K}{\delta'}\right)^{-\frac{(1-2\epsilon)^2 sK}{\epsilon}} \\
&\leq \left(\frac{2sK}{\delta}\right)^{-sK} < \frac{\delta}{2s}.
\end{aligned}$$

□

Theorem 2. *For Reliable Learner algorithm, with probability at least $1 - \delta$, $F_1, \dots, F_s$ are accurate, and the final classifier has unsure rate less than ϵ.*

Proof. First we show with probability at least $1 - \delta$, $F_1, \dots, F_s$ are accurate. Note that

$$\begin{aligned}
&\Pr(F_i \text{ is accurate} | F_1, \dots, F_{i-1} \text{ are accurate}) \\
&\geq \Pr(\text{more than } m \text{ samples are drawn} | F_1, \dots, F_{i-1} \text{ are accurate}) \\
&\quad \cdot \Pr(F_i \text{ is accurate} | \text{more than } m \text{ samples}, F_1, \dots, F_{i-1} \text{ are accurate}).
\end{aligned}$$

We know that

$$\Pr(F_i \text{ is accurate} | \text{more than } m \text{ samples}, F_1, \dots, F_{i-1} \text{ are accurate}) \geq 1 - \frac{\delta}{2s},$$

which is direct conclusion of PAC-learning setting of each stage. And according to lemma 2

$$\Pr(\text{more than } m \text{ samples are drawn} | F_1, \dots, F_{i-1} \text{ are accurate}) \geq 1 - \frac{\delta}{2s},$$

thus

$$\Pr(F_i \text{ is accurate} | F_1, \dots, F_{i-1} \text{ are accurate}) \geq \left(1 - \frac{\delta}{2s}\right) \cdot \left(1 - \frac{\delta}{2s}\right) \geq 1 - \frac{\delta}{s}.$$

Then

$$\begin{aligned}\Pr(F_1, \ldots, F_s \text{ are accurate}) &= \prod_{i=1}^{s} \Pr(F_i \text{ is accurate}|F_1, \ldots, F_{i-1} \text{ are accurate})\\ &\geq \left(1 - \frac{\delta}{s}\right)^s \geq 1 - \delta.\end{aligned}$$

When all $F_1, \ldots, F_s$ are accurate, according to lemma 1, the final classifier has unsure rate less than ϵ. □

6 Remove Circuits Size as Input

In previous analysis, we actually spread our tolerance ϵ to each line of ϵ/s. However, in some case we may not know s for the target concept, then we must use different strategy to distribute the error into unknown number of stages. We use equality

$$\sum_{i=1}^{s} \frac{6\epsilon}{\pi^2 i^2} < \sum_{i=1}^{\infty} \frac{6\epsilon}{\pi^2 i^2} = \epsilon$$

and modify the ϵ' and δ' in Reliable Learner Algorithm 2 to

$$\begin{aligned}\epsilon' &= \frac{6\epsilon}{K\pi^2 i^2}\\ \delta' &= \frac{3\delta}{\pi^2 i^2}.\end{aligned}$$

We can apply similar approach as before to prove the Reliable PAC-learnability of the new algorithm without s as an input.

Bibliographic notes

Analysis of PAC-learning is due to [1, 3]. The main result of this note is from [2].

References

[1] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[2] Ronald L Rivest and Robert Sloan. A formal model of hierarchical concept-learning. *Information and Computation*, 114(1):88–114, 1994.

[3] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.