

Towards Black-box Iterative Machine Teaching

Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, Jim Rehg, Le Song

出典: ICML 2018

概要

- 先生が、生徒の特徴空間を知ることができない (そもそも特徴量空間が一致していない) 状態で Machine teaching を行えるようにした。
- 新しく提案する先生のモデル active teacher は生徒にクエリ (例えば、確認テスト) を投げて、その回答により、生徒の現在のモデルを推定して、より早く収束するようにする。

選択した理由

- Iterative Machine teaching の続きの話だから。

1 Introduction

- Machine teaching の研究として、Iterative Machine teaching があるが、ここでは先生は生徒のモデルを完全に捉えることができ、理論的には興味深い。
- しかし、実際には先生が、生徒のモデルを完全に捉えているのは、難しい。
- Machine teaching の一番の問題は、ブラックボックスである生徒をどうやって教育するか、ということ。これを *black-box machine teaching* と呼ぶ。
- black-box machine teaching* を解くために、我々は *cross-space machine teaching* を提案する。特徴は以下の通り。
 - 先生と生徒は異なる特徴空間を持つ。
 - 先生は生徒のモデルを直接観測することはできない。
- 我々は、*cross-space teaching problem* にアクセスできる先生のモデル active teacher を作る。active teacher は、各イテレーション中に生徒にクエリを投げて、生徒からの回答を得る。生徒の回答は、 $\langle w_t, \tilde{x} \rangle$ を受け取る。 w_t は t 時点でのモデルパラメータ、 \tilde{x} は、生徒の特徴空間のクエリに対する representation

2 Related Work

- いままでの Machine teaching は、生徒のモデルは iterative algorithm(SGD とか) によって決まって最適化される。しかし、実際には生徒はもっと早く学習する。
- 私たちの *cross-space teaching* は *black-box iterative teaching* の踏み石になる。

3 Cross-Space Iterative Machine Teaching

cross-space iterative machine teaching paradigm で使う記法を書く。

Teacher. 先生はサンプル \mathcal{A} (写真, テキスト...), 特徴ベクトル $x_{\mathcal{A}} \in \mathbb{R}^d$ とラベル $y \in \mathbb{R}$ を観察できる。また生徒のモデル (損失関数とか), 最適化アルゴリズム (SGD とか) と真のパラメータ v^* を知っている。先生の予測は $\hat{y}_{v^*} = \langle v^*, x \rangle$ と書く。

Leaner. 生徒はサンプル \mathcal{A} を観察できる。生徒の特徴ベクトルは $x_{\mathcal{A}} \in \mathbb{R}^s$, ラベルは $\tilde{y} \in \mathbb{R}$ と書く。生徒は線形モデル $\langle w, \tilde{x} \rangle$ で、更新は SGD を用いる。生徒のモデルの予測は t 時点で $\hat{y}_w^t = \langle w^t, \tilde{x} \rangle$ と書く。

Representation. 先生と生徒の特徴ベクトルは次元が一致していないので $\tilde{x} = \mathcal{G}(x)$ と one-to-one なマッピングしてくれる関数 \mathcal{G} を考える。

Interaction. 各イテレーション中に、先生は生徒に教材を提示して、教材を用いて、生徒を学習させていく。先生は生徒から $\langle w^t, \tilde{x} \rangle$ の回答を受け取る。この回答は、回帰や分類、もしくは confidence score $S(\langle w^t, \tilde{x} \rangle)$ のどれか。

先生が生徒に教材を生成する方法として 3 つあげる。

Three ways of generating teaching examples

Regression: $\mathcal{Y} = \mathbb{R}$, Classification $\mathcal{Y} = \{-1, 1\}$

- Synthesis-based teaching:

$$\mathcal{X} = \{x \in \mathbb{R}^d, \|x\| \leq R\}$$

- Combination-based teaching:

$$\mathcal{X} = \{x \mid \|x\| \leq R, x = \sum_{i=1}^m \alpha_i x_i, x_i \in \mathcal{D}\},$$

with $\mathcal{D} = \{x_i\}_{i=1}^m$

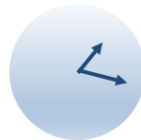
- Rescaled Pool-based teaching:

$$\mathcal{X} = \{x \mid \|x\| \leq R, x = \gamma x_i, x_i \in \mathcal{D}\},$$

with $\mathcal{D} = \{x_i\}_{i=1}^m$



Synthesis



Combination



Rescaled Pool

13/21

図 1:

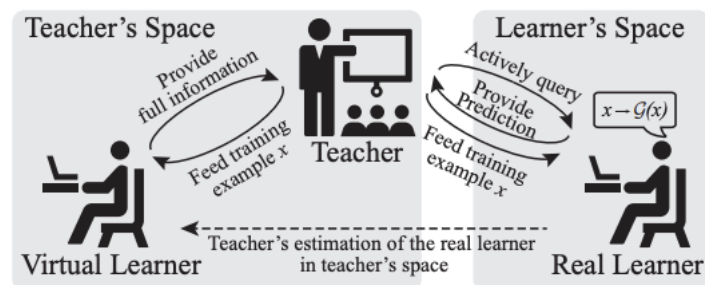


図 2: 概要図

4 The Active Teaching Algorithm

Active teaching algorithm の概要図を図 2 で示す。

- 先生は生徒にクエリを投げ、生徒からの回答を用いて、先生は生徒のモデルと先生が想定している仮想の生徒のモデルを一致させていく。
- 理想的な仮想生徒は $\langle v, x \rangle = \langle w, \tilde{x} \rangle$ になっている。 $\tilde{x} = \mathcal{G}(x) \Leftrightarrow v = \mathcal{G}^T(w)$ が言える。(??)
- active teaching の中で一番のやりたいことは、仮想の生徒をできる限り、本物の生徒に近づけること。
- 仮想の生徒を作ることができたら、IMT での Omniscient teacher を使うことができる。

生徒からのフィードバックから仮想生徒を作る方法を二つ提案する。

1. Exact Recovery of virtual learner

先生は生徒からの回答 $F(\langle w, \tilde{x} \rangle)$ を受け取ることができる。 $\langle w, \tilde{x} \rangle$ から $F(\langle w, \tilde{x} \rangle)$ の情報の欠損がないかぎり、正確に仮想生徒を作ることができる。 $(F(\cdot))$ が one-to-one map だったら \rightarrow ヒンジ関数とかだったら復元できない。

2. Approximate recovery of the virtual learner

生徒からの回答 $F(\cdot)$ が one-to-one じゃなかった時はこっちで再現する。少ないサンプルから $\mathcal{G}^T(w)$ を推定するちょっと賢いテクニックをつかう。(Settles, 2010) の方法をつかう。ただこの能動学習アルゴリズムは影響がない重みのノルムの予想ラベルに基づくモデルパラメータを得てしまう問題がある。こうすると先生が教材が何を選べばいいかわかんなくなってしまう。だからちょっと工夫として、仮想生徒を復元するためにもっとも必要な重みのノルムを見つける必要がある。

5 Theoretical Result

教材の提示方法について書かれている。

6 Discussions and Extensions

- 生徒の学習率を推定するためにはまず生徒の初期パラメータを推定する必要がある。その後は先生によって更新されたパラメータを使っていく。そのときの学習率 η は次のように計算される。

$$\eta = \frac{1}{d} \sum \left((w_1 - w_2) \cdot \nabla_{w\ell} \left(w_1^T x, y \right) \right)$$

\cdot は要素同士の割り算。合計は w_1 の次元に合わせて合計する。

- 忘却率も考える。生徒の学習は $w^{t+1} = w^t + \nabla_{w\ell} (\langle w^t, x \rangle, y)$ で学習していくが、これは先生が教えていることを完全に理解し、忘れないと想定している。実際にはこんなことはないのでパラメータを下のように変える。

$$\hat{w}^{t+1} = w^{t+1} + \epsilon_t$$

ϵ_t は t 時点でのランダムな偏差ベクトル。詳細は Appendix D に。

7 Experiments

The omniscient teacher をベースライン (upper bound) としてつかう。人工データでは、 $\|\mathcal{G}^T(w^t) - v^*\|_2$ を使い、実際の画像ではテストデータでの accuracy を使う。

元論文へ

8 Conclusions

- black-box machine teaching を解くために、cross-space teaching を提案した。従来の Machine teaching より現実に近いようにした。
- Mahchine teaching の恩恵はさらに 2 つ使える..
 - 新しいタスクで新しいモデルを教えるための既製のモデルをより良くすることができる。これは転移学習に使える。
 - 人間の学習でどうあがっていくか、どう教えるのが効率いいかわかる。

9 感想

Iterative machine teaching のときよりもより現実世界の話に近い machine teaching の話だった。いろいろやっててなんかつらくなった。