A Challenge in Machine Teaching

Jerry Zhu

University of Wisconsin-Madison

Dagstuhl Seminar on Machine Learning and Formal Methods
2017
Teaching problems, organized as a tensor

The problem at the origin:

\[
\min_{D \in \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n} \|D\|_0 \\
\text{s.t.} \quad A(D) = \{f^*\}
\]

- \(A(D) = \{f \in \mathcal{F} : f \text{ fits } D\}\), the version space learner
- the objective value is the teaching dimension of \(f^*\)
- applications: education, data poisoning attacks, . . .
Dimension 1: The learner $A$

- $A(D) = \{ f \in \mathcal{F} : f \text{ fits } D \}$, the version space learner
- Bayesian learner $A(D) = P(f \mid D)$
- Regularized empirical risk minimizer

$$A(D) = \arg\min_{\theta} \sum_{i=1}^{n} \ell(\theta, x_i, y_i) + \lambda \|\theta\|$$

- Online learner (e.g. SGD): $D$ is a sequence
- Reinforcement learner
- ...
Dimension 2: Search space

- continuous (can lie): \( D \in \mathcal{D} = \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \)
- continuous (honest): \( D \in \mathcal{D} = (\bigcup_{n=1}^{\infty} (\mathcal{X})^n, Y = f^*(X)) \)
- pool-based teaching: \( D \in \mathcal{D} = 2\{(x_i, y_i)\}_{1:N} \) (multiset)
- alter existing training set \( D_0 + \delta \)
- ...
Dimension 3: Constraints and objectives

- \( \min_D \| D \|_0 \text{ s.t. } A(D) = \{ f^* \} \)
- \( \min_D \text{cost}(D) \text{ s.t. } \| A(D) - f^* \| \leq \epsilon \)
- \( \min_D \| A(D) - f^* \| \text{ s.t. } \text{cost}(D) \leq k \)
- \( \min_D \text{cost}(D) + \lambda \| A(D) - f^* \| \)
- ...
Dimension 4: Protocol between teacher and learner

- a single, known, standard learner
- a learner that expects to be taught
- same lecture on a classroom of different learners (optimal partition)
- a learner with unknown learning algorithm (teaching + probing)
- teach both features and items
- ...
A challenge: finding a teaching set $D$

Example: teach SVM $\theta^* = 0$

- $A=$SVM
- pool $= \{x_1, \ldots, x_6\}$
- semantics: $|A(D) - \theta^*| \leq \epsilon$
- syntax:
  1. $n = x? \in \{x_1, x_2, x_3\}$
  2. $p = x? \in \{x_4, x_5, x_6\}$
  3. return $\text{abs}((n + p)/2) \leq \epsilon$