Semi-supervised learning with adversarially missing label information

by U. Syad and B. Taskar @ Univ. Penn

Akisato Kimura, Ph.D
[E-mail] akisato@ieee.org    [Twitter] @_akisato
Abstract

- Semi-supervised learning in an adversarial setting
  - Consider a general framework, including
    - Covariate shift: Missing labels are not necessarily selected at random.
    - Multiple instance learning: Correspondences between labels and instances are not provided explicitly.
    - Graph-based regularization: Only information about which instances are likely to have the same label.
  - but not including
    - Malicious label noise setting: Labelers are allowed to mislabel a part of the training set.
    - Labelers can remove labels, but cannot change them.
Framework

- $(\hat{x}, \hat{y}) \sim \mathcal{D}^m : m$ labeled training examples
  - $\hat{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)^\top \in \chi^m$ : examples
  - $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)^\top \in \gamma^m$ : labels
  - $\mathcal{D} \in \mathcal{P}(\chi \times \gamma)$ : unknown PDF

- $R(q) \in R(\hat{x}, \hat{y}) : label$ regularization function
  - $q \in \mathcal{P}(I_m \times \gamma) : soft$ labeling of the training examples $\hat{x}$
    - $q(i, y) : probability$ so that $\hat{x}_i$ has label $y \in \gamma$
  - The algorithm can access to only the examples $\hat{x}$ and the label regularization function $R$.
  - Correct labeling $\hat{y}$ might be near the minimum of $R$, but not necessarily (thus it is called “adversarial”).
Label regularization functions (1)

- **Fully supervised learning**
  - $\mathcal{R}(\hat{x}, \hat{y}) = \{R_{\hat{y}}\}, \quad R_{\hat{y}} = 0 \iff q = \hat{y}, \quad R_{\hat{y}} = \infty \iff \text{otherwise}
  - [Note] Labels $\hat{y}$ also denotes the correct soft labeling.

- **Semi-supervised learning**
  - $\mathcal{R}(\hat{x}, \hat{y}) = \{R_{I}\}, \quad I \subseteq I_m$
  - $R_I(q) = 0 \iff q(i, y) = 1\{y = \hat{y}_i\} \forall i \in I, \forall y \in Y$
  - $R_I(q) = \infty \iff \text{otherwise}

- **Ambiguous labeling** (≥ multiple instance learning)
  - $\mathcal{R}(\hat{x}, \hat{y}) = \{R_{\hat{Y}}\}, \quad \hat{Y}_i = \text{a set of possible labels of the example } \hat{x}_i$
  - $R_{\hat{Y}}(q) = 0 \iff \text{supp}(q(i, \cdot)) \subseteq \hat{Y}_i \forall i \in I_m, \quad R_{\hat{Y}}(q) = \infty \iff \text{otherwise}

Don’t mind how $q$ labels examples not in $I$

A set of indexes of unlabeled examples

A set of indexes of non-zero components
Label regularization functions (2)

- **Laplacian regularization**
  - \( R_L(q) = \sum_{y \in \mathcal{Y}} q(\cdot, y)^\top L(\hat{x}) q(\cdot, y) \), \( L(\hat{x}) \) : positive semi-definite
  - \( L(\hat{x})(i, j) \) is small \( \iff \hat{x}_i \) and \( \hat{x}_j \) are believed to have the same label

- Any combinations are also in this framework.
  - Ex. \( R_I(q) + R_L(q) \) : SSL with Laplacian regularization

- We don’t specify how the function is selected.
  - Ex. Semi-supervised learning :
    Usually, we can’t know how the unlabeled samples were selected (“covariate shift” setting).
Loss functions for model learning

- Any types of loss functions can be applicable, but particularly we focus on:
  - Negative log likelihood of a log-linear PDF
    \[ L_{\text{like}}(\theta, x, y) = -\log p_\theta(y|x) = -\log \frac{\exp(\theta^T \phi(x, y))}{\sum_{y'} \exp(\theta^T \phi(x, y'))} \]
  - Binary loss function of a linear classifier
    \[ L_{\text{bin}}(\theta, x, y) = 1 \left\{ \arg \max_{y' \in \mathcal{Y}} \theta^T \phi(x, y') \neq y \right\} \]
- \( \theta \in \mathcal{R}^d \) : a model/classifier parameter
- \( \phi(x, y) \in \mathcal{R}^d \) : a feature function ("basis function" in PRML)
Algorithm

Objective: Find a parameter $\theta^*$ achieving

$$\min_{\theta} \max_{q \in \mathcal{P}(I_m \times Y)} \left\{ E_{\tilde{x},q}[L(\theta, x, y)] - R(q) + \alpha \|\theta\|^2 \right\}$$

- Its validity will be disclosed later.
- For brevity, we abbreviate the objective as $F(\theta, q)$.

Basic algorithm: Game for Adversarially Missing Evidence (GAME)
1. Constants $\epsilon_1, \epsilon_2 > 0$ are given in advance.
2. Find $\tilde{q}$ s.t. $\min_{\theta} F(\theta, \tilde{q}) \geq \max_{q \in \mathcal{P}(I_m \times Y)} \min_{\theta} F(\theta, q) - \epsilon_1$
3. Find $\tilde{\theta}$ s.t. $F(\tilde{\theta}, \tilde{q}) \leq \min_{\theta} F(\theta, \tilde{q}) + \epsilon_2$
4. Output the parameter estimate $\tilde{\theta}$

The order is opposite!!
Algorithm in detail

- Consider the case of the negative log likelihood

- **Step 3**
  - The label regularization function can be ignored.
  - This step is equivalent to maximization of the likelihood of a log-linear model.

- **Step 2**
  - Take the dual of the inner minimization
    \[
    \max_{q \in \mathcal{P}(\mathcal{I}_m \times \mathcal{Y})} \min_{\theta} F(\theta, q) = \max_{(p, q) \in \mathcal{P}(\mathcal{I}_m \times \mathcal{Y})^2} G(p, q)
    \]
    \[
    G(p, q) = H(p) - \frac{1}{\alpha} \|\Delta_\phi(p, q)\|^2 - R(q)
    \]
    \[\Delta_\phi(p, q) = E_{\hat{x}, p}[\phi(x, y)] - E_{\hat{x}, q}[\phi(x, y)], \quad H(p) : \text{Shannon entropy}\]
Convergence properties (1)

- **Theorem 1** (Converse theorem)
  - If $(\hat{x}, \hat{y}) \sim \mathcal{D}^m$, then for all parameters $\theta (\|\theta\| \leq 1)$ and label regularization functions $R \in \mathcal{R}(\hat{x}, \hat{y})$
    
    $$
    E_{\mathcal{D}}[L(\theta, x, y)] \leq \max_{q \in \mathcal{P}(I_m \times Y)} \left\{ E_{\hat{x}, q}[L(\theta, x, y)] - R(q) \right\} + R(\hat{y}) + \epsilon(\delta, m)
    $$
  
  satisfies with probability $\geq 1 - \delta$.

  Similar to the objective function of the algorithm
Theorem 2 (Direct theorem: lower bound)

Consider the binary loss function, and a finite set $X$. For all learning algorithms and example distributions $D_X$, there exists a labeling function $h^*$ that satisfies

$$E_{D_X \cdot h^*}[L(\hat{\theta}, x, y)] \geq \frac{1}{4} \max_{q \in \mathcal{P}(I_m \times Y)} \left\{ E_{\hat{\theta}, q}[L(\hat{\theta}, x, y)] - R(q) \right\} + \min_{q \in \mathcal{P}(I_m \times Y)} R(q) - \epsilon(\delta, m)$$

with probability $\geq 1/4 - 2\delta$ under some assumptions.

- The assumptions can be seen in the paper (1, 2 and 3).
- Some detailed conditions are omitted for the simplicity.
- The parameter $\hat{\theta}$ is obtained by a learning algorithm.
Theorem 3 (Direct theorem: upper bound)

Consider the binary loss function. If \((\hat{x}, \hat{y}) \sim \mathcal{D}^m\), there exists a family \(\mathcal{R}\) of label regularization function and a learning algorithm that satisfy

\[
E_\mathcal{D}[L(\hat{\theta}, x, y)] 
\leq \max_{q \in \mathcal{P}(I_m \times \mathcal{Y})} \left\{ E_{\hat{x},q}[L(\hat{\theta}, x, y)] - R(q) \right\} + \min_{q \in \mathcal{P}(I_m \times \mathcal{Y})} R(q) + \epsilon(\delta, m) - 1
\]

with probability \(\geq 1 - \delta\) under some assumptions.

Assumption 3 can be removed.

Some detailed conditions are omitted for the simplicity.

The parameter \(\hat{\theta}\) is obtained by the learning algorithm.
Objective: Find a parameter $\theta^*$ achieving
$$\min_{\theta} \max_{q \in \mathcal{P}(I_m \times Y)} \left\{ E_{\hat{x},q} [L(\theta, x, y)] - R(q) + \alpha \|\theta\|^2 \right\}$$

Minimum upper bound in the converse theorem
$$\min_{\theta : \|\theta\| \leq 1} \max_{q \in \mathcal{P}(I_m \times Y)} \left\{ E_{\hat{x},q} [L(\theta, x, y)] - R(q) \right\} + R(\hat{y}) + \epsilon(\delta, m)$$

- Do not minimize over $\theta$ ($\|\theta\| \leq 1$).
- Instead, add a regularization part to leave $\theta$ unconstrained.
Experiments

- **Exp. 1: Binary classification**
  - Datasets: COIL (object recognition), BCI (EEG scans)

- **Exp. 2: Multi-class categorization**
  - Dataset: Labeled Faces in the Wild
Results

BCI

COIL

Faces

Accuracy vs. Fraction of training set labeled for BCI, COIL, and Faces datasets.
この論文から何を考えるか？

「一般化」論文は得てして面白くない。
- その特殊ケースの結果を知っているので、「おお、そうだったのか」という驚きを得られにくい。

- が、一般化が新しい見方を提供することは見逃せない。
  - 知られていない特殊ケースの発見
  - 特殊ケースの組み合わせによる新しい知見/手法

- さらに、一般化で技術を俯瞰できるようになることも重要。

- なので、ここからもう一歩先を考えてみましょう！
Fin.

- Thank you for your kind attention.