Learnability with Indirect Supervision Signals (NeurIPS 2020)

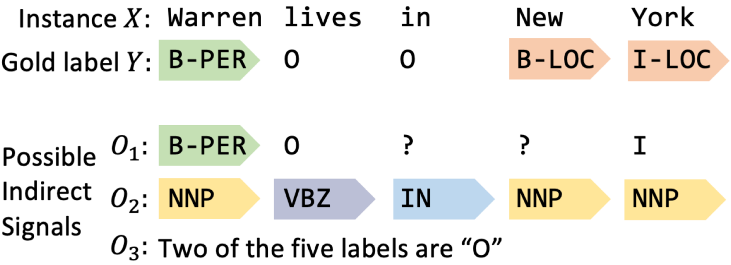
Kaifu Wang, Qiang Ning and Dan Roth

February 1, 2021

Motivation

- In our work, we consider a classification task where we predict the target label Y of an instance variable X.
- An *indirect supervision signal* is any random variable (denoted by O) that is correlated to the target label Y.
- We assume the learner only receives samples of (X, O) but does not observe Y directly.

Taking the named entity recognition (NER) tagging as an example:



Instance X:		Warren	lives	in	New	York
Gold label Y:		B-PER	O	O	B-LOC	I-LOC
Possible Indirect Signals	O_1:	B-PER	O	?	?	I
	O_2:	NNP	VBZ	IN	NNP	NNP
	O_3:	Two of the five labels are “O”				

Problems

Learnability problem concerns whether we can learn the optimal classifier in our model given sufficiently many indirect supervision samples (just like using gold labels).

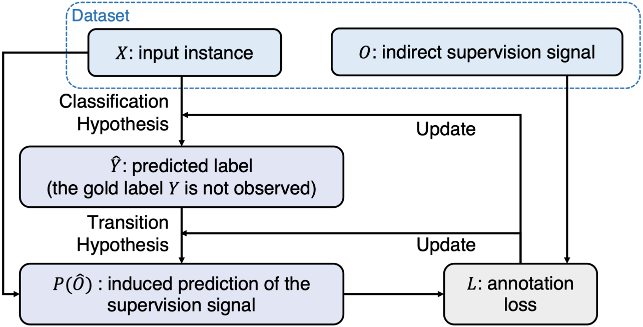
- Intuitively some indirect signals cannot guarantee learnability since they are *weak*. For example, O_3 only tells a statistics of the label but there can be a lot of wrong predictions that satisfy this constraint.
- In contrast, O_1 seems to be a promising choice if the missing rate is not too high.
- How do we find a way to distinguish them?

Learning

- To understand this problem, we need to have an algorithm that uses samples of (X, O) to supervise our learning process.
- A natural choice is to maximize the likelihood of the observed samples of O.
- In a standard classification problem, we have a *hypothesis class* $\mathcal{H}$ which contains candidate classifiers $h : X \mapsto \widehat{Y}$.
- To further compute the likelihood of O, we also need to model $\mathbb{P}(O|X, \widehat{Y})$. We call it *transition hypothesis*, denoted by $\mathcal{T}$.
- Sometimes $\mathbb{P}(O|X, \widehat{Y})$ is known (e.g., O_3). In this case, $\mathcal{T}$ only contains a single mapping.
- In other cases, we need a non-parametrized (e.g., O_1) or parametrized (e.g., O_2) model.

Learning (contd.)

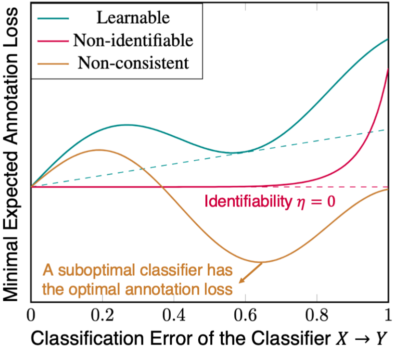
Our learning framework is concluded in the figure. The learner uses the prediction of Y to induce predictions about O. This prediction is then evaluated by the observed dataset. The annotation loss is used to update the classifier and the transition hypothesis.



Dataset
X: input instance
O: indirect supervision signal
Classification Hypothesis
Update
$\hat{Y}$: predicted label
(the gold label Y is not observed)
Transition Hypothesis
Update
$P(\hat{O})$: induced prediction of the supervision signal
L: annotation loss

Learnability Condition: Overview

To illustrate the learnability condition, we plot the the relationship between the classification error of a hypothesis h and the minimum annotation loss (risk) it can have (over choices of transition hypotheses).



Learnability Condition 1: Consistency

The optimal classifier should be able to induce an optimal prediction of the indirect signal. Formally, we require:

Condition 1

The optimal classifier $h_0 \in \underset{h \in \mathcal{H}, T \in \mathcal{T}}{\operatorname{argmin}} \operatorname{Risk}_{\mathcal{O}}(T \circ h)$.

Remark: When the consistency condition does not hold (this can happen when our signal is very noisy), maximizing the likelihood of the observable will contradict our goal of maximizing the likelihood of the true label.

Learnability Condition 2: Identifiability

A suboptimal classifier should induce higher annotation loss than the lowest annotation loss on average. Formally, we require

Condition 2

Define and let

$$\eta := \inf_{h \in \mathcal{H}, T \in \mathcal{T} : R(h) > 0} \frac{\mathrm{Risk}_{\mathcal{O}}(T \circ h) - \inf_{T \in \mathcal{T}} \mathrm{Risk}_{\mathcal{O}}(T \circ h_0)}{\mathrm{Risk}(h)} > 0$$

Remark: O_2 may fail to satisfy this condition (with some models) because it is possible that we can predict PoS tagging without the knowledge of NER.

Learnability Condition 3: Complexity

Our model should not be too complex. Complexity of a model can be described by VC-dimension.

Condition 3

We assume $\ell_{\mathcal{O}} \circ \mathcal{T} \circ \mathcal{H}$ is weak VC-major with dim $d < \infty$.

Learning Bound

Now we are able to state the first main result:

Theorem (Learnability)

If the above three conditions are satisfied, then for any $\delta < 1$, with probability of at least $1-\delta$, we have:

$$\mathrm{Risk}(\mathrm{ERM}(S^{(m)})) \leq \frac{2b}{\eta}\left(\sqrt{\frac{2\overline{\Gamma}_m(d)}{m}} + \frac{4\overline{\Gamma}_m(d)}{m} + \sqrt{\frac{2\log(4/\delta)}{m}}\right)$$

where $\overline{\Gamma}_m(d)$ is defined as

$$\overline{\Gamma}_m(d) := \log\left[2\sum_{j=0}^{\min\{d,m\}}\binom{m}{j}\right] = d\log m(1+o(1)) \text{ as } m \to \infty$$

This implies $R(\mathrm{ERM}(S^{(m)})) \to 0$ in probability as $m \to \infty$.

In words, we can find the optimal classifier as we have a large training set.

Separation: Illustration

To check the first two conditions with the learner's prior knowledge, we further propose the **separation** condition. We illustrate the definition using the example of partial label O_1 for a 3-class classification problem where O_1 is identified as a subset of $\{1, 2, 3\}$.

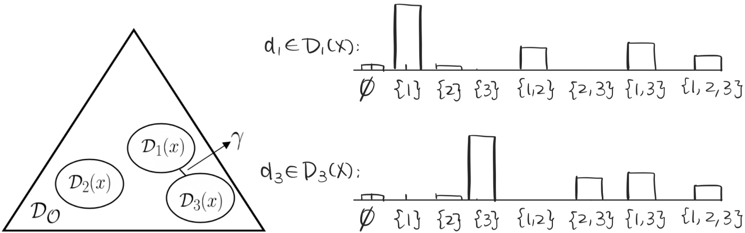


Figure: A (predicted) label y_i will induce a distribution family over $\mathcal{O}$, denoted as $\mathcal{D}_i(x)$. Different families are separated by a minimal "distance" $\gamma > 0$.

Separation: Definition

Formally:

Theorem (Separation)

For all $x \in \mathcal{X}$, we denote the induced distribution families by label y_i as $\mathcal{D}_i(x) = \{(T(x))_i : T \in \mathcal{T}\} \subseteq \mathcal{D}_{\mathcal{O}}$, and the set of all possible predictions of y as $\mathcal{H}(x) = \{h(x) : h \in \mathcal{H}\} \subseteq \mathcal{Y}$. If

$$\gamma = \inf_{(x,i,j):p(x,y_i)>0, j\neq i, y_j \in \mathcal{H}(x)} \mathrm{KL}(\mathcal{D}_i(x) \,\|\, \mathcal{D}_j(x)) > 0 \tag{1}$$

Then the first two conditions are satisfied with $\eta \geq \gamma > 0$ via the ERM of the cross-entropy loss for the observables.

Moreover, if (1) is not satisfied, then it can be shown the learning problem can be arbitrarily difficult since different labels can induce arbitrarily similar distributions over $\mathcal{O}$. In other words, the observation of O cannot help us to distinguish different labels.

Application: Partial Annotation with Noise

In this way, we can include many learning conditions for specific scenarios proposed in literature as special cases of separation:

- Partial with noise (the true label may not appear in O), where $\mathcal{O} = 2^{\mathcal{Y}}$, i.e., the annotation is a subset in $\mathcal{Y}$:

$$\gamma_C = \inf_{p(x,y_i)>0, i\neq j} \mathbb{P}(y_i \in O|x, y_i) - \mathbb{P}(y_j \in O|x, y_i) > 0$$

- Massart condition for binary classification with label noise, where $\mathcal{O} = \mathcal{Y}$ and the labels are flipped with certain probability:

$$\gamma_C = \sup_{p(x,y_i)>0, i\neq j} \mathbb{P}(O \neq y_i|x, y_i) < \frac{1}{2}$$

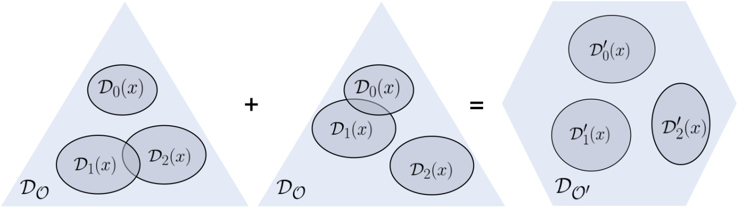
- Small ambiguity degree condition for noise-free superset problem, where $\mathbb{P}(y_i \in O|x, y_i) = 1$:

$$\gamma_C = \sup_{p(x,y_i)>0, i\neq j} \mathbb{P}(y_j \in O|x, y_i) < 1$$

Application: Joint Supervision

If a single source of supervision signal cannot ensure learnability, it should be used jointly with other signals. We show that a joint supervision can:

- Harm the separation if supervision signals are simply mixed. This is due to the convexity of the KL-divergence.
- Preserve the pairwise separation if modelled properly. This effect is visualized in the following figure, where each signal cannot separate one pair of labels, but can be combined to ensure global separation.



- Create new separation: If there are constraints between different signals, these constraints can be utilized to supervise the learning.

Thank you