Learning Functional Distributions with Private Labels

Changlong Wu (CSoI, Purdue University)

Joint with

Yifan Wang, Ananth Grama and Wojciech Szpankowski
(CSoI, Purdue University)

ICML2023, Honolulu, Hawaii

Center for Science of Information
NSF Science & Technology Center

Motivation

Consider how *clinical features* (e.g., age, gender) impact the probability of manifesting various *consequences* after catching a disease.

Motivation

Consider how *clinical features* (e.g., age, gender) impact the probability of manifesting various *consequences* after catching a disease.

Rate compared to 18-29 years old[1]	0-4 years old	5-17 years old	18-29 years old	30-39 years old	40-49 years old	50-64 years old	65-74 years old	75-84 years old	85+ years old
Hospitalization[3]	0.7x	0.2x	Reference group	1.5x	1.8x	3.1x	5.0x	9.3x	15x

Motivation

Consider how *clinical features* (e.g., age, gender) impact the probability of manifesting various *consequences* after catching a disease.

Rate compared to 18-29 years old[1]	0-4 years old	5-17 years old	18-29 years old	30-39 years old	40-49 years old	50-64 years old	65-74 years old	75-84 years old	85+ years old
Hospitalization[3]	0.7x	0.2x	Reference group	1.5x	1.8x	3.1x	5.0x	9.3x	15x

- Let $\mathcal{X}$ be the set of features, $\mathcal{Y}$ be the set of consequences and $\Delta(\mathcal{Y})$ be the set of distributions over $\mathcal{Y}$.

Motivation

Consider how *clinical features* (e.g., age, gender) impact the probability of manifesting various *consequences* after catching a disease.

Rate compared to 18-29 years old[1]	0-4 years old	5-17 years old	18-29 years old	30-39 years old	40-49 years old	50-64 years old	65-74 years old	75-84 years old	85+ years old
Hospitalization[3]	0.7x	0.2x	Reference group	1.5x	1.8x	3.1x	5.0x	9.3x	15x

- Let $\mathcal{X}$ be the set of features, $\mathcal{Y}$ be the set of consequences and $\Delta(\mathcal{Y})$ be the set of distributions over $\mathcal{Y}$.
- The goal is to find a map $p : \mathcal{X} \to \Delta(\mathcal{Y})$ by observing data sampled from real patients.

Motivation

Consider how *clinical features* (e.g., age, gender) impact the probability of manifesting various *consequences* after catching a disease.

Rate compared to 18-29 years old[1]	0-4 years old	5-17 years old	18-29 years old	30-39 years old	40-49 years old	50-64 years old	65-74 years old	75-84 years old	85+ years old
Hospitalization[3]	0.7x	0.2x	Reference group	1.5x	1.8x	3.1x	5.0x	9.3x	15x

- Let $\mathcal{X}$ be the set of features, $\mathcal{Y}$ be the set of consequences and $\Delta(\mathcal{Y})$ be the set of distributions over $\mathcal{Y}$.
- The goal is to find a map $p : \mathcal{X} \to \Delta(\mathcal{Y})$ by observing data sampled from real patients.

Privacy concerns:

- The feature $\mathcal{X}$ only weakly impacts label $\mathcal{Y}$. Therefore, the features are much *less sensitive* than the true labels.

Motivation

Consider how *clinical features* (e.g., age, gender) impact the probability of manifesting various *consequences* after catching a disease.

Rate compared to 18-29 years old[1]	0-4 years old	5-17 years old	18-29 years old	30-39 years old	40-49 years old	50-64 years old	65-74 years old	75-84 years old	85+ years old
Hospitalization[3]	0.7x	0.2x	Reference group	1.5x	1.8x	3.1x	5.0x	9.3x	15x

- Let $\mathcal{X}$ be the set of features, $\mathcal{Y}$ be the set of consequences and $\Delta(\mathcal{Y})$ be the set of distributions over $\mathcal{Y}$.
- The goal is to find a map $p : \mathcal{X} \to \Delta(\mathcal{Y})$ by observing data sampled from real patients.

Privacy concerns:

- The feature $\mathcal{X}$ only weakly impacts label $\mathcal{Y}$. Therefore, the features are much *less sensitive* than the true labels.
- We consider label differential privacy by adding noise only to labels while reveal feature explicitly.

Motivation

Consider how *clinical features* (e.g., age, gender) impact the probability of manifesting various *consequences* after catching a disease.

Rate compared to 18-29 years old[1]	0-4 years old	5-17 years old	18-29 years old	30-39 years old	40-49 years old	50-64 years old	65-74 years old	75-84 years old	85+ years old
Hospitalization[3]	0.7x	0.2x	Reference group	1.5x	1.8x	3.1x	5.0x	9.3x	15x

- Let $\mathcal{X}$ be the set of features, $\mathcal{Y}$ be the set of consequences and $\Delta(\mathcal{Y})$ be the set of distributions over $\mathcal{Y}$.
- The goal is to find a map $p : \mathcal{X} \to \Delta(\mathcal{Y})$ by observing data sampled from real patients.

Privacy concerns:

- The feature $\mathcal{X}$ only weakly impacts label $\mathcal{Y}$. Therefore, the features are much *less sensitive* than the true labels.
- We consider label differential privacy by adding noise only to labels while reveal feature explicitly.

Goal: Design a noisy process that prevent inferring the true labels while still learning the underlying true mapping p.

Problem Setup

- Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a set of *hypothesis* that models the underlying truth, i.e., we assume the true map $p \in \mathcal{F}$.

Problem Setup

- Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a set of *hypothesis* that models the underlying truth, i.e., we assume the true map $p \in \mathcal{F}$.
- Let P be a set of *random processes* over $\mathcal{X}^T$ that models the feature generating process.

Problem Setup

- Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a set of *hypothesis* that models the underlying truth, i.e., we assume the true map $p \in \mathcal{F}$.
- Let P be a set of *random processes* over $\mathcal{X}^T$ that models the feature generating process.

We consider the online learning scenario that happens as follows:

1. At beginning *Nature* selects $p \in \mathcal{F}$ and $\mu \in \mathrm{P}$.
2. At time step t, Nature generates $\mathbf{x}_t \sim \mu$ and reveal it to a *predictor*.
3. The predictor predicts $\hat{p}_t \in \Delta(\mathcal{Y})$ based on history observe thus far.
4. Nature generates $y_t \sim p(\mathbf{x}_t)$ and reveals $\tilde{y}_t = \mathcal{K}_\eta(y_t)$ to predictor, where $\mathcal{K}_\eta$ is a *noisy kernel* (channel).

Problem Setup

- Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a set of *hypothesis* that models the underlying truth, i.e., we assume the true map $p \in \mathcal{F}$.
- Let P be a set of *random processes* over $\mathcal{X}^T$ that models the feature generating process.

We consider the online learning scenario that happens as follows:

1. At beginning *Nature* selects $p \in \mathcal{F}$ and $\mu \in \mathsf{P}$.
2. At time step t, Nature generates $\mathbf{x}_t \sim \mu$ and reveal it to a *predictor*.
3. The predictor predicts $\hat{p}_t \in \Delta(\mathcal{Y})$ based on history observe thus far.
4. Nature generates $y_t \sim p(\mathbf{x}_t)$ and reveals $\tilde{y}_t = \mathcal{K}_\eta(y_t)$ to predictor, where $\mathcal{K}_\eta$ is a *noisy kernel* (channel).

Goal: Find a prediction rule $\hat{p}^T$ that minimizes the expected KL-risk:

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{P}) = \sup_{\mu \in \mathsf{P}, p \in \mathcal{F}} \mathbb{E}\left[\sum_{t=1}^{T} \mathsf{KL}(p(\mathbf{x}_t), \hat{p}_t(\mathbf{x}^t, \tilde{y}^{t-1}))\right].$$

Related Work

- Our setup can be understood as an extension for the *randomized response* scenario of (Warner, 1965) by allowing features to influence outcome distributions.
- Label differential privacy was studied in (Chaudhuri & Hsu, 2011; Esfandiari et al., 2022; Ghazi et al., 2021; Wu et al., 2023). But only for the classification problems.
- Learning *conditional* distributions was studied in the context of *sequential probability assignment* in (Yang & Barron, 1998; Cesa-Bianchi & Lugosi, 2006; Rakhlin & Sridharan, 2015; Bilodeau et al., 2020; Wu et al., 2022b; Bhatt & Kim, 2021; Bilodeau et al., 2021). But considers only the regret formulation.

Main Results

Let $\mathcal{Y}$ be a finite set of size M, and $\mathcal{K}_\eta$ be a *random mapping* such that for all $y \neq y' \in \mathcal{Y}$

$$\Pr[\mathcal{K}_\eta(y) = y] = 1 - \eta,$$

and

$$\Pr[\mathcal{K}_\eta(y) = y'] = \frac{\eta}{M-1}.$$

Theorem 1: Let $\mathcal{F}$ be **any finite class** and the features are generated **adversarially**. Then for the noisy kernel $\mathcal{K}_\eta$, we have

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{P}) \leq O\left(\frac{\log(MT)\sqrt{T\log|\mathcal{F}|}}{1 - \frac{M\eta}{M-1}}\right).$$

Moreover, for any $k \leq T$, there exists class $\mathcal{F}$ with $|\mathcal{F}| = 2^K$ such that

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{P}) \geq \Omega(\sqrt{T\log|\mathcal{F}|}).$$

Main Results

Let $\mathcal{G}$ be a set of functions map $\mathcal{X}^* \to \Delta(\mathcal{Y})$. We say $\mathcal{G}$ stochastic sequential covers $\mathcal{F}$ w.r.t. P at confidence δ and scale α, if

$$\forall \mu \in \mathsf{P},\ \Pr_{\mathbf{x}^T \sim \mu}\left[\exists p \in \mathcal{F} \forall g \in \mathcal{G} \exists t \in [T], \mathsf{TV}(p(\mathbf{x}_t), g(\mathbf{x}^t)) > \alpha\right] \leq \delta.$$

Theorem 2: Let $\mathcal{F}$ and P be **arbitrary classes** and $\mathcal{G}_\alpha$ be the **stochastic sequential** cover of $\mathcal{F}$ w.r.t. P at scale α and confidence $\delta = \frac{1}{TM}$. Then for the noisy kernel $\mathcal{K}_\eta$, we have

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{P}) \leq O\left(\frac{\log(MT)\sqrt{T \inf_{\alpha \geq 0}\{M\alpha^2 T/\eta + \log|\mathcal{G}_\alpha|\}}}{1 - \frac{M\eta}{M-1}}\right).$$

Example

Let $\mathcal{H} \subset [N]^{\mathcal{X}}$ be a class of functions that classifies $\mathcal{X}$ into N categories.

The Hidden Classification Model $\mathcal{F}$ w.r.t. $\mathcal{H}$ is defined as

$$\mathcal{F} = \left\{ p_{h,\mathbf{q}}(\mathbf{x}) = q_{h(\mathbf{x})} : h \in \mathcal{H}, \mathbf{q} = \{q_1, \cdots, q_N\} \in \Delta(\mathcal{Y})^M \right\}.$$

Theorem 3 Let $\mathcal{H} \subset [N]^{\mathcal{X}}$ be any class with Pseudo-dimension $\mathsf{Pdim}(\mathcal{H})$ and P be the class of all *i.i.d.* **processes**. If $\mathcal{F}$ is the *hidden classification model* w.r.t. $\mathcal{H}$. Then for the noisy kernel $\mathcal{K}_\eta$, we have

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{P}) \leq \tilde{O}(\sqrt{T(\mathsf{Pdim}(\mathcal{H}) + NM)}).$$

Moreover, there exists class $\mathcal{H}$ such that

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{P}) \geq \Omega(\sqrt{T \max\{\mathsf{Pdim}(\mathcal{H}), NM\}}).$$

Thanks!