The impact of Record Linkage on Learning from Feature Partitioned Data

[1]Google Research

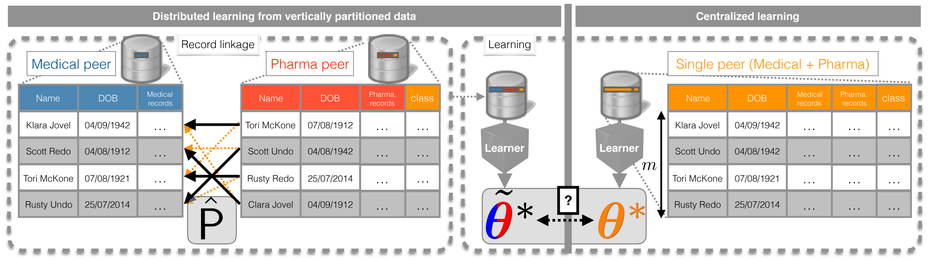
Richard Nock[1], Stephen Hardy[2], Wilko Henecka[2], Hamish Ivey-Law[3], Jakub Nabaglo[3], Giorgio Patrini[4], Guillaume Smith[2] & Brian Thorne[2]

[3]Australian National University

[2]AMBIATA

[4]sensity

Setting



- **Our framework**: **vertical partition** (**VP**): *features* split among 2 peers, **M** and **P** (1+, e.g. **P**, holds label **C**)
- **Record linkage** (**RL**) needed before batch supervised learning
- **Baseline framework**: batch learning: labeled sample to learn classifier
- Example: (Med+Pharma) single ***peer*** holds all data

Problems, summarised: compare $\tilde{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^*$? Improve RL to get better $\tilde{\boldsymbol{\theta}}^*$?

Summary

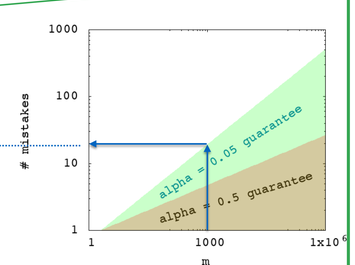
Theory

- Linear models: sufficient conditions to get on training:

$$\frac{\|\tilde{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}^*\|_2} \leq O\left(\frac{1}{m^\alpha}\right)$$

- Regularisation is key (*for broad set of losses*)
- Optimisation of RL prior to ML, e.g. min. *between-classes* RL errors
- Large margins on $\boldsymbol{\theta}^* \Rightarrow$ right class on $\tilde{\boldsymbol{\theta}}^*$ ("RL immunity for large margins")
- Results hold in the small data regime
- A small # of RL mistakes does not impact ML

RL does less than ~20 mistakes for 1000 ex. implies $\alpha = 0.05$.



Experiments

- Exps include case where 1 peer does not hold class or a noisy estimate
- Simple approaches to complete / correct label prior to RL can $\Rightarrow$ leverage
- Potential (estimated) ***dials*** to evaluate the value of RL prior to ML
  - inaccuracy between matched observations
  - error between classes for matched observations
- ***Margin immunity observed experimentally***:
  - strong advocacy for distributed / federated ML
  - may offer further cheap dials to evaluate the ML "potential" of datasets

Theory

- **Loss optimized** = Taylor loss + Ridge regularisation. For training sample $\hat{S} \doteq \{(\hat{\boldsymbol{x}}_i, y_i), i \in [m]\}$, the loss is:

$$\ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma) \doteq \mathbb{E}_i[F(y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)] + \boldsymbol{\theta}^\top \Gamma \boldsymbol{\theta}\,, \text{with } F(z) \doteq a + bz + cz^2\,, \text{and } a \in \mathbb{R}, b, c \in \mathbb{R}_*\,, \text{and } \Gamma \text{ symmetric PSD}$$

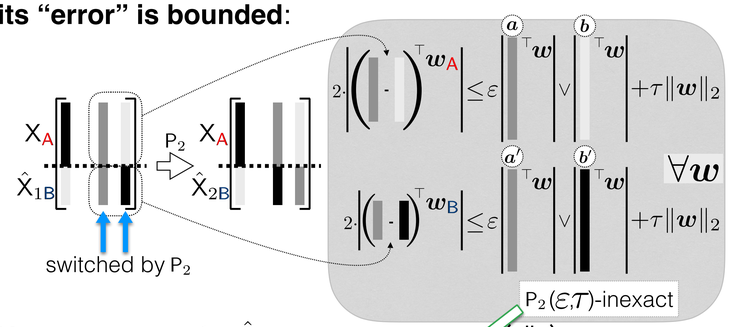
- **Record Linkage** setting: observation matrix $\mathsf{X} \in \mathbb{R}^{d \times m}$, split between two peers A and B. RL reconstructs $\hat{\mathsf{X}}$ but makes mistakes as (unknown) permutation matrix $\hat{\mathsf{P}} \in \{0,1\}^{m \times m}$

$$\hat{\mathsf{X}} \doteq \begin{bmatrix} \mathsf{X}_\mathsf{A} \\ \hline \hat{\mathsf{X}}_\mathsf{B} \doteq \mathsf{X}_\mathsf{B}\hat{\mathsf{P}} \end{bmatrix} \text{(RL)} \Leftarrow \text{A} \Leftrightarrow \text{B} \Leftrightarrow \mathsf{X} \doteq \begin{bmatrix} \mathsf{X}_\mathsf{A} \\ \hline \mathsf{X}_\mathsf{B} \end{bmatrix} \text{ (unknown)}$$

Decomposition as elementary permutations:

Each elementary permutation P_t switches parts between 2 observations. It is called (ε, τ)-**inexact** iff **its "error" is bounded**:

$$\hat{\mathsf{P}} = \prod_{t=1}^{T} \mathsf{P}_t$$



Key parameters for $\hat{\mathsf{P}}$: T = size *and* $\xi(\hat{\mathsf{P}}) \doteq \min\left\{\varepsilon + \frac{\tau}{X_*} : \hat{\mathsf{P}} \text{ is } (\varepsilon,\tau)\text{-inexact}\right\}$ (all t)

never 'large': $\xi(.) \leq 2$; max column-norm in X

- **Theorem:** if ℓ_F 'sufficiently' regularised & m not too 'small', $m \geq 16$ if $b = c$

$$\frac{\|\tilde{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}^*\|_2} \leq \frac{\xi^{\frac{1}{4}}}{m} \cdot T^2 \cdot \left(\xi^{\frac{3}{4}} + \frac{\delta_{\hat{\mathsf{P}},\ell}}{\delta_{\boldsymbol{\theta}}}\right)$$

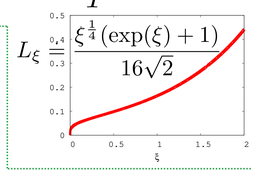
#errs on observations w/ ***different*** classes

where $\delta_{\boldsymbol{\theta}} \doteq \|\boldsymbol{\theta}^*\|_2 X_*$ and $\delta_{\hat{\mathsf{P}},\ell} \doteq \frac{\rho|b|L_\xi}{|c|}$ *and* $\rho \doteq \frac{T_+}{T} \in [0,1]$

- Furthermore, $\forall \kappa > 0$, if m 'large enough':

$$\frac{m}{\xi^{\frac{1}{4}} T^2} > \frac{\xi^{\frac{3}{4}} \delta_{\boldsymbol{\theta}} + \delta_{\hat{\mathsf{P}},\ell}}{\kappa}$$

$L_\xi \doteq \frac{\xi^{\frac{1}{4}}(\exp(\xi)+1)}{16\sqrt{2}}$



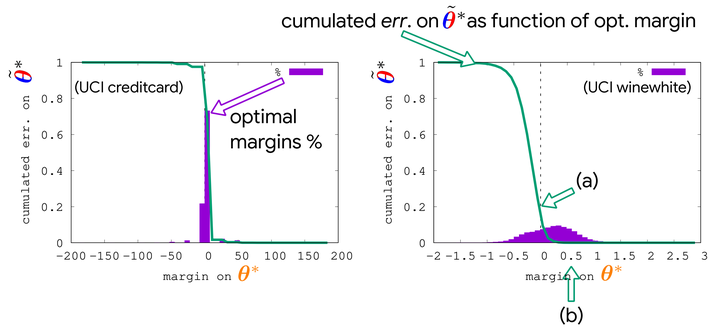
then $\forall (\boldsymbol{x}, y), \left(y \boldsymbol{\theta}^{*\top} \boldsymbol{x} > \kappa\right) \Rightarrow \left(y \tilde{\boldsymbol{\theta}}^{*\top} \boldsymbol{x} > 0\right)$

('large' optimal margin) ⇒ (**right class** despite RL mistakes)

($\tilde{\boldsymbol{\theta}}^*$ is *immune to record linkage at margin* κ)

Experiments

- Experiments on RL immunity for large margin classification

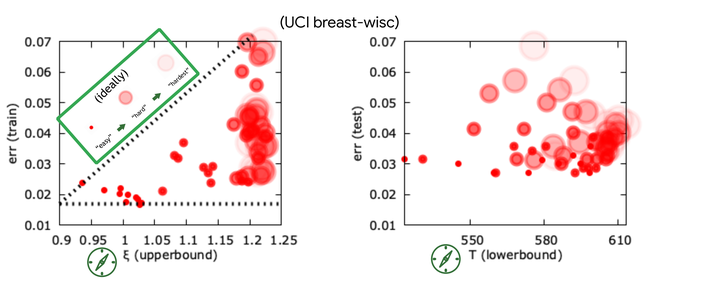


(a): <=***20%*** errors happen on examples with >0 opt. margin

(b): almost ***no error*** on examples with optimal margin > 0.5

Justification for ML on VP setting: VP = more features, so opt. margins ***increase***, beneficial for a ***good*** RL+ML pipeline

- Experiments on computable / estimable dials to 'link' the difficulty of RL to that of ML



Main parameters of $\hat{\mathsf{P}}$ = good dials for ML

Could be of use to select / improve RL techs

Key parameters for $\hat{\mathsf{P}}$:
- size T
- "inaccuracy" ξ
- "error" ρ