In all Likelihoods: Robust Selection of Pseudo-Labeled Data

Julian Rodemann, Christoph Jansen, Georg Schollmeyer, Thomas Augustin
Foundations of Statistics and Their Applications, Department of Statistics, LMU Munich, Germany

July 12, 2023

ISIPTA 2023, Oviedo, Spain

Outline

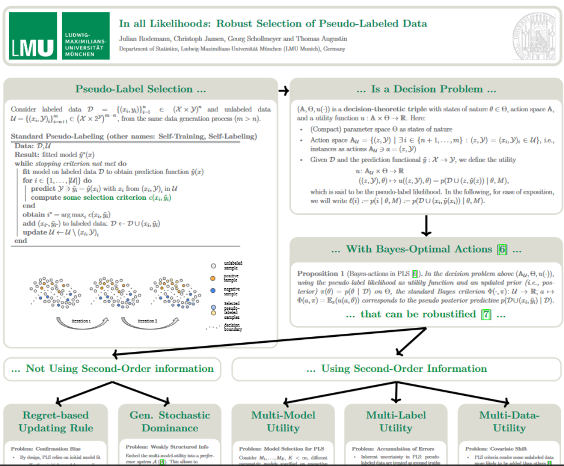


Figure: One poster, one sentence.

Contents

Contents

Intro: What's Pseudo-Labeling?

- Semi-Supervised Learning (Classification)
- Consider labeled data

  $$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

  and unlabeled data

  $$\mathcal{U} = \{(x_i, \mathcal{Y})\}_{i=n+1}^{m}$$

  from the same data generating process, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the categorical target space
- Aim: Use unlabeled data for training

Pseudo-Labeling

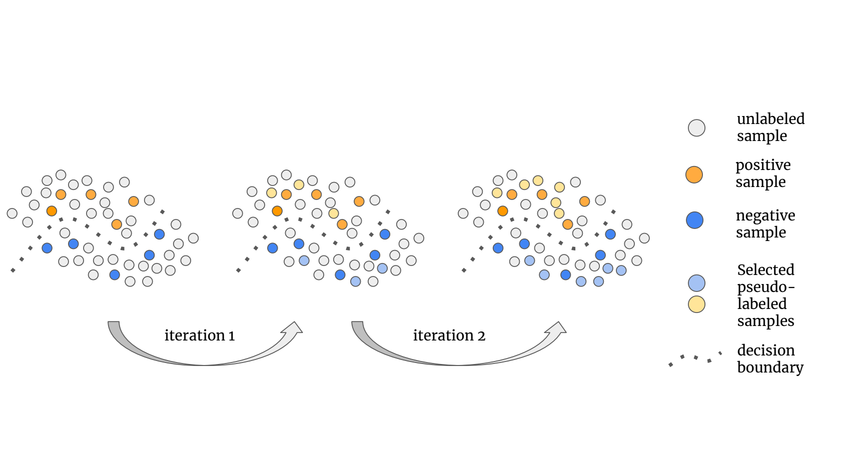


Figure: Sketch of Pseudo-Labeling for Binary Classification.

Pseudo-Labeling

Standard Pseudo-Labeling[1]

while *stopping criterion not met* **do**
  fit model on labeled data $\mathcal{D}$ to obtain prediction function $\hat{y}(x)$
  for $i \in \{1, \ldots, |\mathcal{U}|\}$ **do**
    predict $\mathcal{Y} \ni \hat{y}_i = \hat{y}(x_i)$ with x_i from $(x_i, \mathcal{Y})$ in $\mathcal{U}$
    compute some selection criterion $c(x_i, \hat{y}_i)$
  end
  obtain $i^* = \arg\max_i \, c(x_i, \hat{y}_i)$
  add $(x_{i^*}, \hat{y}_{i^*})$ to labeled data: $\mathcal{D} \leftarrow \mathcal{D} \cup (x_i, \hat{y}_i)$
  update $\mathcal{U} \leftarrow \mathcal{U} \setminus (x_i, \mathcal{Y})_i$
end

[1] Other names: Self-Training, Self-Labeling.

Contents

PLS Is a Decision Problem

Definition (PLS as Decision Problem)

Consider the decision-theoretic triple $(\mathbb{A}_{\mathcal{U}}, \Theta, u(\cdot))$ with

- an action space $\mathbb{A}_{\mathcal{U}}$ of unlabeled data to be selected,
- a space of unknown states of nature (parameters) Θ
- and a utility function $u : \mathbb{A}_{\mathcal{U}} \times \Theta \longrightarrow \mathbb{R}$.

Notably, this decision-theoretic embedding entails optimistic superset learning as special case corresponding to max-max-actions (Hüllermeier, Destercke, and Couso 2019; Rodemann, Kreiss, Hüllermeier, and Augustin 2022)

Contents

Bayesian(,) PL(ea)S(e!)

Proposition (Rodemann, Goschenhofer, Dorigatti, Nagler, and Augustin 2023)

In the decision problem $(\mathbb{A}_\mathcal{U}, \Theta, u(\cdot))$ *with pseudo-label likelihood* $u := p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta)$ *as utility and an updated prior* $\pi(\theta) = p(\theta \mid \mathcal{D})$ *on* Θ*, the standard Bayes criterion*

$$\Phi(\cdot, \pi) \colon \mathbb{A}_\mathcal{U} \longrightarrow \mathbb{R}$$
$$a \mapsto \Phi(a, \pi) = \mathbb{E}_\pi(u(a, \cdot))$$

corresponds to the pseudo posterior predictive $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$.

Bayesian(,) PL(ea)S(e!)

Proposition (tl;dr)

Our Bayes criterion is the **pseudo posterior predictive (PPP)** $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$ *if the likelihood* $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta)$ *is our utility.*

Bayesian(,) PL(ea)S(e!)

Problem: $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$ is expensive to evaluate! $\longrightarrow$ Approximate it (Rodemann, Goschenhofer, Dorigatti, Nagler, and Augustin 2023)

$$p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D}) \approx \underbrace{\underbrace{\ell_{\mathcal{D} \cup (x_i, \hat{y}_i)}(\tilde{\theta})}_{\text{Likelihood of pseudo-sample in light of fitted parameter}} \underbrace{- \frac{1}{2} \log |I(\tilde{\theta})|}_{\text{Flatness of likelihood at this fitted parameter}}}_{\text{uninformative case}} \underbrace{+ \log \pi(\tilde{\theta})}_{\text{Prior at fitted parameter}},$$

where $\tilde{\theta} \approx \arg\max_\theta \ell_{\mathcal{D} \cup (x_i, \hat{y}_i)}(\theta)$

Contents

In all Likelihoods: Robust PLS by multi-objective utility

Definition (Multi-Objective Likelihood Utility)

Consider labeled data $\mathcal{D}$ and pseudo-labels $\hat{y} \in \mathcal{Y}$ from $\hat{y} : \mathcal{X} \to \mathcal{Y}$ as given. The K-dimensional utility function

$$u \colon \mathbb{A}_{\mathcal{U}} \times \tilde{\Theta} \to \mathbb{R}^K$$
$$((x_i, \mathcal{Y}), \theta) \mapsto (\ell(i,1), \dots, \ell(i,K))'$$

shall be called **multi-objective** likelihood.
For instance, with any $M_1, \dots, M_K$, $K < \infty$, different parametric models specified on respective parameter spaces $\Theta_1, \dots, \Theta_K$[2] we can set $\ell(i,k) := p(i \mid f_k(\theta), M_k) = p(\mathcal{D} \cup (z, \hat{y}(z)) \mid f_k(\theta), M_k)$ with $\theta_k \in \Theta_k$.

[2]Further denote by $\tilde{\Theta} = \times_{k=1}^{K} \Theta_k$ their Cartesian product and by $f_k : \tilde{\Theta} \to \Theta_k$, $k \in \{1, \dots, K\}$ the projections from the Cartesian product to each Θ_k.

Contents

Contents

Contents

Contents

Contents

Contents

Generalized Stochastic Dominance

- Embed the multi-objective utility into a *preference system* $\mathscr{A}$ (Jansen, Schollmeyer, and Augustin 2018)
- Denote by $\mathcal{N}_{\mathscr{A}}$ the set of all representations ϕ of $\mathscr{A}$ and define a preorder on the pseudo-labeled data $\mathbb{A}_{\mathcal{U}}$ by setting $a_1 \gtrsim_\pi a_2$ *iff*

$$\forall \phi : \quad \mathbb{E}_\pi(\phi \circ u(a_1, \cdot)) \geq \mathbb{E}_\pi(\phi \circ u(a_2, \cdot))$$

- Then select all pseudo-labeled data in $\mathbb{A}_{\mathcal{U}}$ that are *undominated* w.r.t. $\gtrsim_\pi$

Generalized Stochastic Dominance

- Good News: Under credal prior info Π we can generalize $\gtrsim_\pi$ to $\gtrsim_\Pi$ by setting

$$a_1 \gtrsim_\Pi a_2 \quad :\text{iff} \quad \forall \pi \in \Pi : \; a_1 \gtrsim_\pi a_2$$

  and select all pseudo-labeled data in $\mathbb{A}_\mathcal{U}$ that are *undominated* w.r.t. $\gtrsim_\Pi$
- The relations $\gtrsim_\pi$ and $\gtrsim_\Pi$ are referred to as **Generalized Stochastic Dominance (GSD)** (Jansen, Schollmeyer, Blocher, Rodemann, and Augustin 2023)

Contents

Regret-Based Updating Rule

- By design, PLS relies on initial model fit
- If the initial model generalizes poorly, initial misconceptions can propagate throughout the process (Arazo, Ortego, Albert, O'Connor, and McGuinness 2020)
- Main reasons: model misspecification and/or erroneous label predictions
- Accordingly, we strive for a PLS criterion that is robust with respect to these regrets

Regret-Based Updating Rule

We adapt the α-cut updating rule by (Cattaneo 2014) such that the posterior credal set is

$$\Pi_\alpha = \{\pi \in \Pi \mid m(\ell_{h,h}, \pi) \geq \alpha \cdot \sup_{j,k} m(\ell_{j,k}, \pi)\}$$

with Π a prior credal set, $m(\ell, \pi) = \int_\Theta \ell(\theta)\pi(\theta)d\theta$ the marginal likelihood, $j \in \{1, \ldots, J\}$ for $J = |\mathcal{Y}|$ labels, and $k \in \{1, \ldots, K\}$ for models $M_1, \ldots, M_K$. Denote by $\tilde{u}_{j,k}(\theta, a^*)$ the utility of $a^* \hat{=} i^*$ with prediction $\tilde{y}_{i^*,j}$ under model M_k.

Regret-Based Updating Rule

Defining

$$r(\theta, a^*) = \frac{\sup_{j,k} \tilde{u}_{j,k}(\theta, a^*)}{\tilde{u}_{h,h}(\theta, a^*)}$$

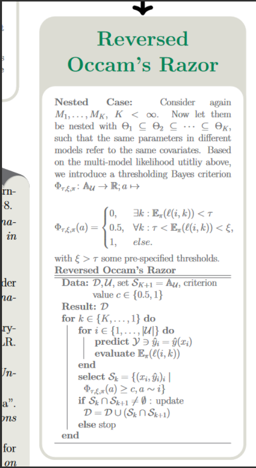
as the myopic regret, we get

Proposition (Myopic Regret-Guarantee of α-Cuts)

Bayes-optimal selections a^* of pseudo-labeled data under the above α-cut updating rule have expected total regret $\mathbb{E}_\pi(r(\theta, a^*)) \leq \frac{1}{\alpha}$ for any posterior $\pi \in \Pi_\alpha$.

Contents

Why You Should Visit Our Poster (1)



Reversed Occam's Razor

Nested Case: Consider again $M_1, \dots, M_K$, $K < \infty$. Now let them be nested with $\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K$, such that the same parameters in different models refer to the same covariates. Based on the multi-model likelihood utitliy above, we introduce a thresholding Bayes criterion $\Phi_{\tau,\xi,\pi} : \mathbb{A}_{\mathcal{U}} \to \mathbb{R}; a \mapsto$

$$\Phi_{\tau,\xi,\pi}(a) = \begin{cases} 0, & \exists k : \mathbb{E}_\pi(\ell(i,k)) < \tau \\ 0.5, & \forall k : \tau < \mathbb{E}_\pi(\ell(i,k)) < \xi, \\ 1, & else. \end{cases}$$

with $\xi > \tau$ some pre-specified thresholds.

Reversed Occam's Razor

Data: $\mathcal{D}, \mathcal{U}$, set $\mathcal{S}_{K+1} = \mathbb{A}_{\mathcal{U}}$, criterion value $c \in \{0.5, 1\}$
Result: $\mathcal{D}$
for $k \in \{K, \dots, 1\}$ **do**
  for $i \in \{1, \dots, |\mathcal{U}|\}$ **do**
    predict $\mathcal{Y} \ni \hat{y}_i = \hat{y}(x_i)$
    evaluate $\mathbb{E}_\pi(\ell(i,k))$
  end
  select $\mathcal{S}_k = \{(x_i, \hat{y}_i)_i \mid \Phi_{\tau,\xi,\pi}(a) \geq c, a \sim i\}$
  if $\mathcal{S}_k \cap \mathcal{S}_{k+1} \neq \emptyset$: update $\mathcal{D} = \mathcal{D} \cup (\mathcal{S}_k \cap \mathcal{S}_{k+1})$
  else stop
end

Figure: How does this all relate to Occam's razor?

Why You Should Visit Our Poster (2)

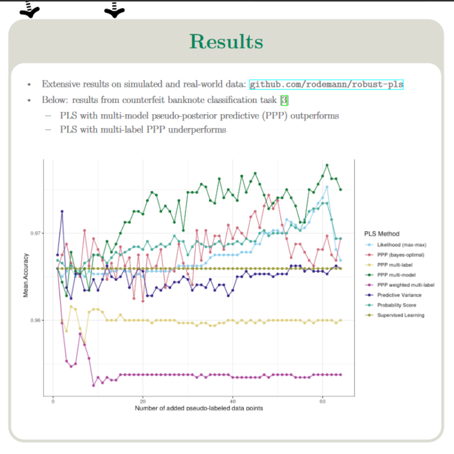


Figure: How does this work in practice?

Literature I

- Arazo, Eric et al. (2020). “Pseudo-labeling and confirmation bias in deep semi-supervised learning”. In: *2020 International Joint Conference on Neural Networks*. IEEE, pp. 1–8.
- Cattaneo, Marco EGV (2014). “A continuous updating rule for imprecise probabilities”. In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, pp. 426–435.
- Hüllermeier, Eyke, Sébastien Destercke, and Ines Couso (2019). “Learning from imprecise data: adjustments of optimistic and pessimistic variants”. In: *International Conference on Scalable Uncertainty Management (SUM)*. Springer, pp. 266–279.
- Jansen, Christoph, Georg Schollmeyer, and Thomas Augustin (2018). “Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences”. In: *International Journal of Approximate Reasoning* 98, pp. 112–131.

Literature II

- Jansen, Christoph et al. (2023). "Robust statistical comparison of random variables with locally varying scale of measurement". In: *Uncertainty in Artificial Intelligence (UAI)*. PMLR.
- Rodemann, Julian et al. (2022). "Levelwise Data Disambiguation by Cautious Superset Learning". In: *International Conference on Scalable Uncertainty Management (SUM)*. Springer, pp. 263–276.
- Rodemann, Julian et al. (2023). "Approximately Bayes-optimal pseudo-label selection". In: *Uncertainty in Artificial Intelligence (UAI)*. PMLR.