In all Likelihood*s*: Robust Selection of Pseudo-Labeled Data

Julian Rodemann, Christoph Jansen, Georg Schollmeyer and Thomas Augustin

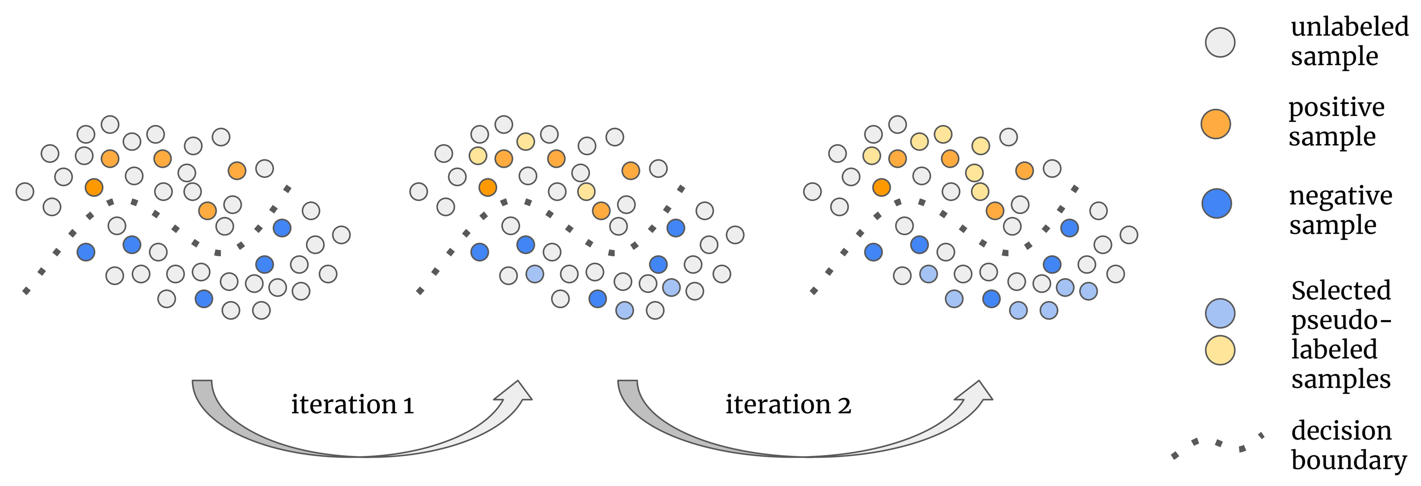
Department of Statistics, Ludwig-Maximilians-Universität München (LMU Munich), Germany

Pseudo-Label Selection ...

Consider labeled data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and unlabeled data $\mathcal{U} = \{(x_i, \mathcal{Y})_i\}_{i=n+1}^m \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^{m-n}$, from the same data generation process ($m > n$).

Standard Pseudo-Labeling (other names: Self-Training, Self-Labeling)

```
Data: D, U
Result: fitted model ŷ*(x)
while stopping criterion not met do
    fit model on labeled data D to obtain prediction function ŷ(x)
    for i ∈ {1, ..., |U|} do
        predict Y ∋ ŷ_i = ŷ(x_i) with x_i from (x_i, Y)_i in U
        compute some selection criterion c(x_i, ŷ_i)
    end
    obtain i* = arg max_i c(x_i, ŷ_i)
    add (x_{i*}, ŷ_{i*}) to labeled data: D ← D ∪ (x_i, ŷ_i)
    update U ← U \ (x_i, Y)_i
end
```



... Is a Decision Problem ...

$(\mathbb{A}, \Theta, u(\cdot))$ is a **decision-theoretic triple** with states of nature $\theta \in \Theta$, action space $\mathbb{A}$, and a utility function $u : \mathbb{A} \times \Theta \to \mathbb{R}$. Here:

- (Compact) parameter space Θ as states of nature
- Action space $\mathbb{A}_{\mathcal{U}} = \{(z, \mathcal{Y}) \mid \exists i \in \{n+1, \ldots, m\} : (z, \mathcal{Y}) = (x_i, \mathcal{Y})_i \in \mathcal{U}\}$, i.e., instances as actions $\mathbb{A}_{\mathcal{U}} \ni a = (z, \mathcal{Y})$
- Given $\mathcal{D}$ and the prediction functional $\hat{y} : \mathcal{X} \to \mathcal{Y}$, we define the utility

$$u \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}$$
$$((z, \mathcal{Y}), \theta) \mapsto u((z, \mathcal{Y}), \theta) = p(\mathcal{D} \cup (z, \hat{y}(z)) \mid \theta, M),$$

which is said to be the pseudo-label likelihood. In the following, for ease of exposition, we will write $\ell(i) := p(i \mid \theta, M) := p(\mathcal{D} \cup (x_i, \hat{y}(x_i)) \mid \theta, M)$.

... With Bayes-Optimal Actions [6] ...

Proposition 1 (Bayes-actions in PLS [6])**.** *In the decision problem above $(\mathbb{A}_{\mathcal{U}}, \Theta, u(\cdot))$, using the pseudo-label likelihood as utility function and an updated prior (i.e., posterior) $\pi(\theta) = p(\theta \mid \mathcal{D})$ on Θ, the standard Bayes criterion $\Phi(\cdot, \pi) \colon \mathcal{U} \to \mathbb{R}$; $a \mapsto \Phi(a, \pi) = \mathbb{E}_\pi(u(a, \theta))$ corresponds to the pseudo posterior predictive $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$.*

... that can be robustified [7] ...

... Not Using Second-Order information

... Using Second-Order Information

Regret-based Updating Rule

Problem: Confirmation Bias

- By design, PLS relies on initial model fit
  - If the initial model generalizes poorly, initial misconceptions can propagate throughout the process [1]
  - This can be due to model misspecification and erroneous label predictions
- Accordingly, we strive for a PLS criterion that is robust with respect to these regrets

We adapt the α-cut updating rule by [2] s.t. 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, $j \in \{1, \ldots, J\}$ 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. 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 2 (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 $\pi \in \Pi_\alpha$.*

Gen. Stochastic Dominance

Problem: Weakly Structured Info

Embed the multi-model-utility into a *preference system* $\mathcal{A}$ ([4]). This allows to

- *harness the entire information* encoded in its cardinal dimensions

while still being able to

- *avoid unjustified assumptions* on the hierarchy of the involved models.

Denote by $\mathcal{N}_{\mathcal{A}}$ the set of all representations ϕ of $\mathcal{A}$ and define a preorder on the pseudo-labeled data $\mathbb{A}_{\mathcal{U}}$ by setting $a_1 \succsim_\pi a_2$ *iff*

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

Select all pseudo-labeled data in $\mathbb{A}_{\mathcal{U}}$ that are *undominated* w.r.t. $\succsim_\pi$ (compare also [5]).

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

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

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

The relations $\succsim_\pi$ and $\succsim_\Pi$ are referred to as **Generalized Stochastic Dominance (GSD)**.

Multi-Model Utility

Problem: Model Selection for PLS

Consider $M_1, \ldots, M_K$, $K < \infty$, different parametric models specified on respective parameter spaces $\Theta_1, \ldots, \Theta_K$. Denote by $\tilde{\Theta} = \times_{k=1}^K \Theta_k$ their Cartesian product and by $f_k : \tilde{\Theta} \to \Theta_k$, $k \in \{1, \ldots, K\}$ the projections from the Cartesian product to each Θ_k. Consider $\mathcal{D}$ and pseudo-labels $\hat{y} \in \mathcal{Y}$ from $\hat{y} : \mathcal{X} \to \mathcal{Y}$ as given. The K-dimensional utility function

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

shall be called multi-model likelihood, where we write $\ell(i, k) = p(i \mid f_k(\theta), M_k) = p(\mathcal{D} \cup (z, \hat{y}(z)) \mid f_k(\theta), M_k)$ with $f_k(\theta) = \theta_k$ the parameter vector of model k.

Reversed Occam's Razor

Nested Case: Consider again $M_1, \ldots, M_K$, $K < \infty$. 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 $\Phi_{\tau,\xi,\pi} \colon \mathbb{A}_{\mathcal{U}} \to \mathbb{R}; a \mapsto$

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

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

Reversed Occam's Razor

```
Data: D, U, set S_{K+1} = A_U, criterion
      value c ∈ {0.5, 1}
Result: D
for k ∈ {K, ..., 1} do
    for i ∈ {1, ..., |U|} do
        predict Y ∋ ŷ_i = ŷ(x_i)
        evaluate E_π(ℓ(i, k))
    end
    select S_k = {(x_i, ŷ_i)_i |
      Φ_{τ,ξ,π}(a) ≥ c, a ∼ i}
    if S_k ∩ S_{k+1} ≠ ∅ : update
      D = D ∪ (S_k ∩ S_{k+1})
    else stop
end
```

Multi-Label Utility

Problem: Accumulation of Errors

- Inherent uncertainty in PLS: pseudo-labeled data are treated as ground truths in subsequent iterations
- Idea: Consider not all other hypothetical labels $\tilde{y}_i \in \mathcal{Y} \setminus \{\hat{y}_i\}$ in PLS

Denote by $\tilde{y}_{i,j} \in \mathcal{Y}$ all possible labels for $(x_i, \mathcal{Y})_i$ with $j \in \{1, \ldots, J\}$ and $J = |\mathcal{Y}|$. We assign utility to each $(x_i, \mathcal{Y})_i$ by the following utility function $u_l \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}$;

$$((z, \mathcal{Y}), \theta) \mapsto \sum_{j=1}^{J} w_j \cdot p(\mathcal{D} \cup (z, \tilde{y}_{i,j}) \mid \theta, M)$$

with weights $w_j \in (0, 1)$ summing up to 1.

Multi-Data-Utility

Problem: Covariate Shift

- PLS criteria render some unlabeled data more likely to be added than others [8]

-> Induces distributional shift of covariates' marginal distribution

- Idea: select pseudo-labeled data that are optimal w.r.t both the *de facto* selected data $\mathcal{D}$ and a *hypothetical i.i.d.* sample $\mathcal{D}'$ that we generate by drawing pseudo-labeled data randomly.

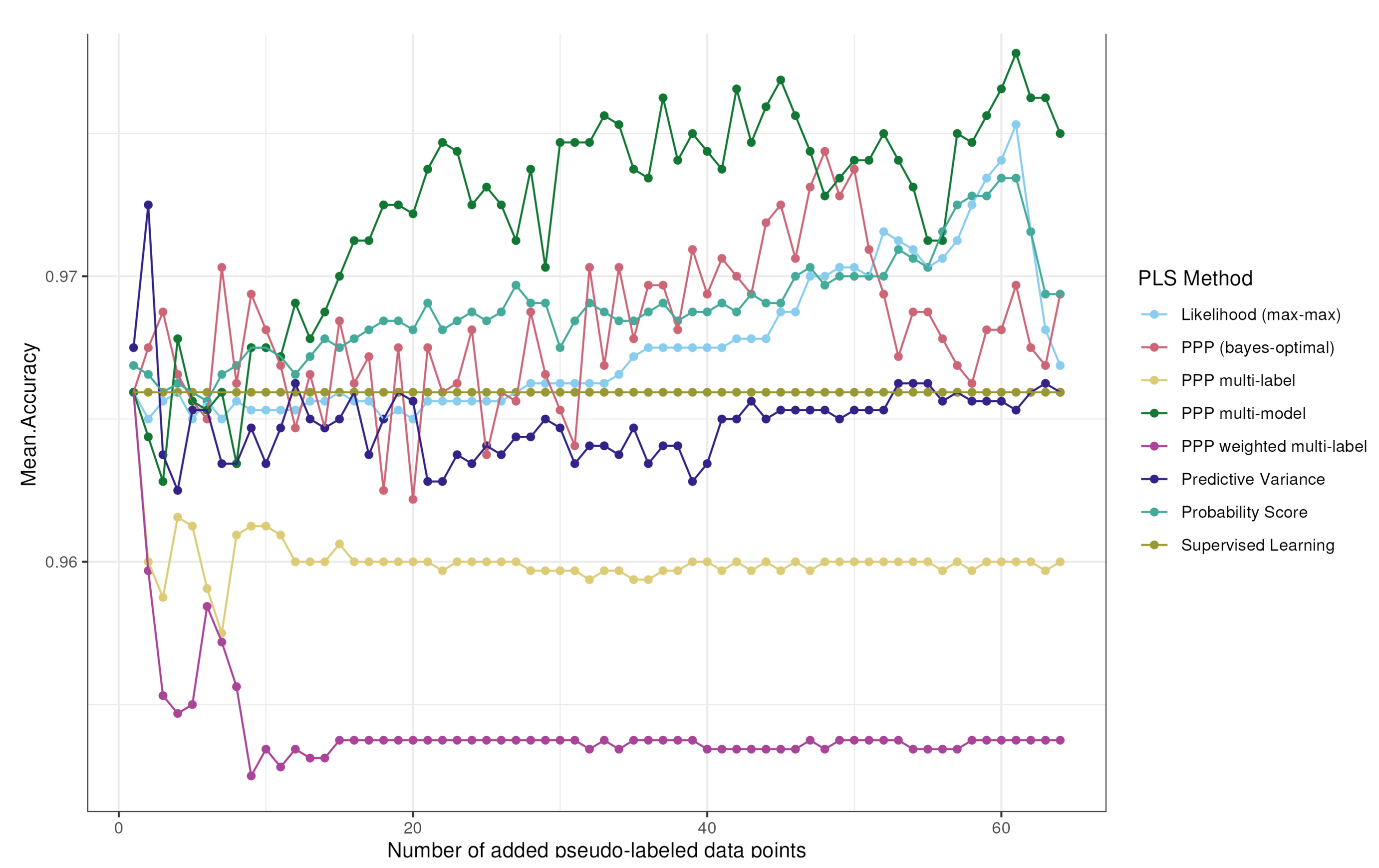
We assign utility to each $(x_i, \mathcal{Y})_i$ given $\mathcal{D}$, $\mathcal{D}'$ and the prediction functional $\hat{y} : \mathcal{X} \to \mathcal{Y}$ by $u_d \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}^2$;

$$((z, \mathcal{Y}), \theta) \mapsto (\ell_{\mathcal{D}}(i), \ell_{\mathcal{D}'}(i))',$$

with $\ell_{\mathcal{D}}(i) = p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta, M)$ and $\ell_{\mathcal{D}'}(i) = p(\mathcal{D}' \cup (x_i, \hat{y}_i) \mid \theta, M)$.

Results

- Extensive results on simulated and real-world data: `github.com/rodemann/robust-pls`
- Below: results from counterfeit banknote classification task [3]
  - PLS with multi-model pseudo-posterior predictive (PPP) outperforms
  - PLS with multi-label PPP underperforms



References

[1] E. Arazo et al. "Pseudo-labeling and confirmation bias in deep semi-supervised learning". In: *International Joint Conference on Neural Networks*. IEEE. 2020, pp. 1–8.

[2] M. Cattaneo. "A continuous updating rule for imprecise probabilities". In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer. 2014, pp. 426–435.

[3] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017.

[4] C. Jansen, G. Schollmeyer, and T. Augustin. "Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences". In: *International Journal of Approximate Reasoning* 98 (2018), pp. 112–131.

[5] C. Jansen et al. "Robust statistical comparison of random variables with locally varying scale of measurement". In: *Uncertainty in Artificial Intelligence (UAI)*. PMLR. 2023 (to appear).

[6] J. Rodemann et al. "Approximately Bayes-optimal pseudo-label selection". In: *Uncertainty in Artificial Intelligence (UAI)*. PMLR. 2023 (to appear).

[7] J. Rodemann et al. "In all Likelihoods: Robust Selection of Pseudo-Labeled Data". In: *International Symposium on Imprecise Probabilities Theories and Applications (ISIPTA)*. PMLR. 2023.

[8] J. Rodemann et al. "Not all data are created equal: Lessons from sampling theory for adaptive machine learning". *Poster presented at IMS International Conference on Statistics and Data Science (IMS-ICSDS)*. 2022.