YSLS: Pseudo-Label Selection: Some Insights From Decision Theory

joint work with Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, Thomas Augustin, Christoph Jansen, Georg Schollmeyer

Julian Rodemann

April 18, 2023

Contents

Contents

Weakly 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 generation process, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the categorical target space.
- Aim: Use unlabeled data for training
- Applications
  - image classification
  - genomics
  - ranking search results `https://ai.googleblog.com/2021/07/from-vision-to-language-semi-supervised.html`

Contents

Pseudo-Labeling

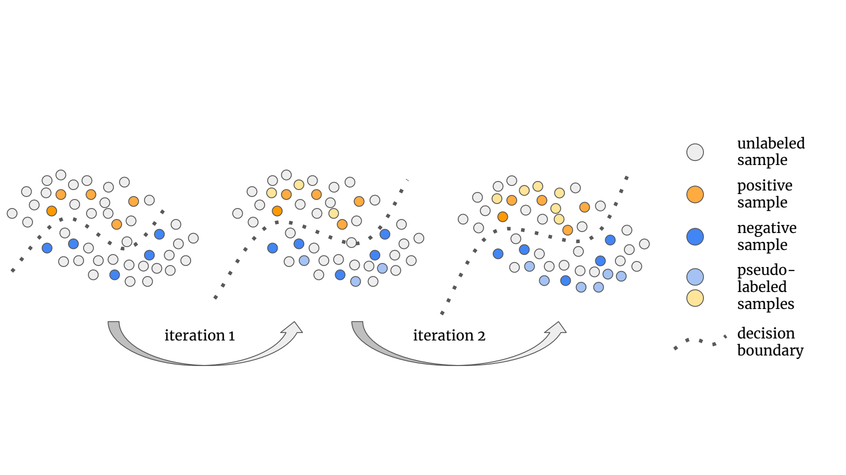

Figure: Sketch of Pseudo-Labeling for Binary Classification. Credits: Jann G.

Contents

PLS is a decision problem! [5]

Definition (PLS as Decision Problem)

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

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

Contents

Why Bayesian?

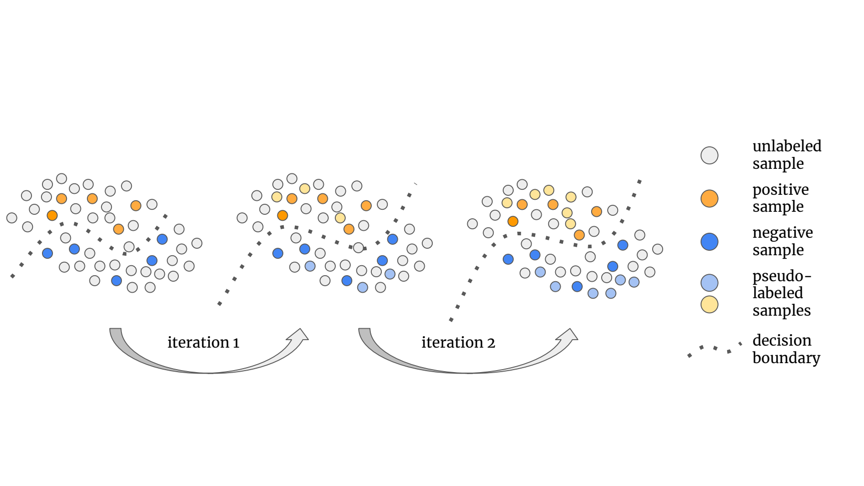

Figure: Sketch of Pseudo-Labeling for Binary Classification. Credits: Jann G.

Bayesian PLS! [5]

Theorem

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

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

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

Bayesian PLS!

Theorem (tl;dr)

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

Bayesian PLS!

"First, be Bayesian. Then worry about how well you're doing it."

– Philipp Hennig

Bayesian PLS!

Problem: $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$ is expensive to evaluate! $\longrightarrow$ Approximate it!

Contents

Approximate Bayes Optimal PLS [5]

Selection Criterion:

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

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

Contents

Results (Uninformative Prior)

Results on Simulated Data with $q = 60$

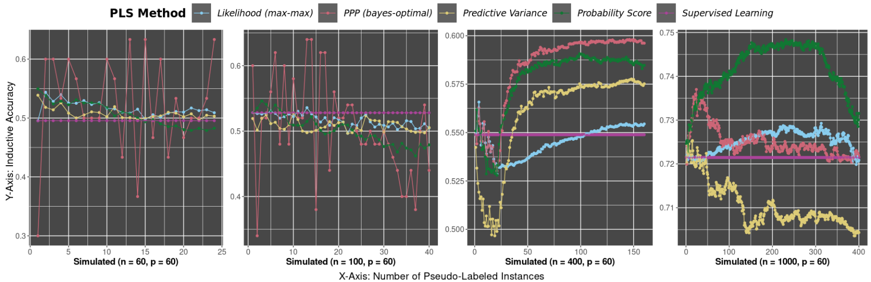

Figure: Complete Results on Simulated Data for $q = 60$. $R = 100$; $\frac{n_{unlabeled}}{n_{train}} = 0.8$.

Results (Uninformative Prior)

Results on Real Data

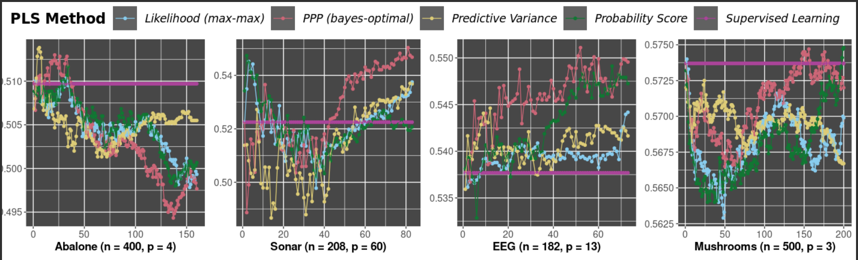

Figure: Results from 8 classification tasks based on real-world data [2] in descending difficulty (measured by supervised test accuracy), where p denotes the number of features here and the share of unlabeled data is 0.8. Accuracy averaged over 100 repetitions.

Results (Informative Prior)

Results on Simulated Data

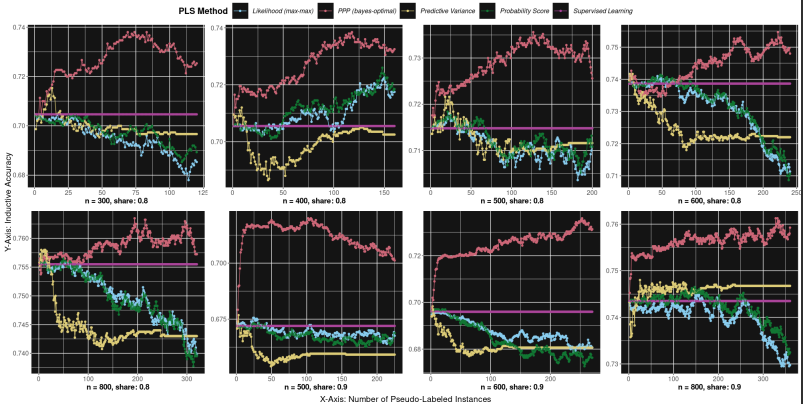

Figure: Results of PPP with informative priors on simulated data with different shares of unlabeled data. Accuracy averaged over 100 repetitions.

Contents

Extensions

- Extensions: Decision-theoretic embedding paves the way for various extensions [6]
  - multi-objective utility accounting for
    - model selection
    - covariate shift
    - accumulation of errors
    - ...
  - Generalized Bayes via Credal Sets
    - α-cut updating

Extensions

- Consider any $M_1, \ldots, M_K$, $K < \infty$, different parametric models specified on respective parameter spaces $\Theta_1, \ldots, \Theta_K$.[1]
- We can easily extend the pseudo-label likelihood utility (definition **??**) to account for several models, inducing a multiobjective decision problem.

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

Extensions

Definition (Multi-Model 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

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

shall be called multi-model likelihood. 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 $\theta_k \in \Theta_k$ for brevity.

Extensions: Generalized Bayes

- Idea: (convex) set of priors

$$\Pi \subseteq \{\pi(\theta) \mid \pi(\cdot) \text{ a probabilty measure on } (\Theta, \sigma(\Theta))\}$$

with Θ compact as above and $\sigma(\cdot)$ an appropriate σ-algebra.

- Γ-maximin, e.g. [7, 1, 3, 8, 4]: $\underline{\mathbb{E}}_{\Pi}(u(a,\theta)) = \inf_{\pi \in \Pi} \mathbb{E}(u(a,\theta))$
- How to update Π?

$$\{\pi \in \Pi \mid m(\pi) \geq \alpha \cdot \max_{\pi} m(\pi)\}$$

with $m(\ell, \pi) = \int_{\Theta} \ell(\theta)\pi(\theta)d\theta$ the marginal likelihood.

Contents

Discussion

- Approximate Bayes optimal PLS ...
  - ... is more robust towards the initial fit than classical PLS
  - ... can be applied to any kind of predictive model whose likelihood and Fisher-information are accessible
  - ... allows to include prior information
  - ... does not require an *i.i.d.* assumption

Discussion

- Limitations
  - With $|\mathcal{U}| = m$ unlabeled data points and no stopping criterion, $m + (m-1) + \cdots + 1 = \frac{m^2+m}{2}$ PPPs have to approximated.
  - Overfitting scenarios might be hard to identify

Contents

Literature I

[1] James O. Berger. *Statistical decision theory and Bayesian analysis.* 2nd. Springer, Berlin., 1985.

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. `http://archive.ics.uci.edu/ml`. 2017.

[3] Itzhak Gilboa and David Schmeidler. “Maxmin expected utility with non-unique prior”. In: *Journal of Mathematical Economics* 18.2 (1989), pp. 141–153.

[4] Peijun Guo and Hideo Tanaka. “Decision making with interval probabilities”. In: *European Journal of Operational Research* 203.2 (2010), pp. 444–454.

Literature II

[5] Julian Rodemann, Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, and Thomas Augustin. *Bayesian PLS! Approximate Bayes Optimal Pseudo-Label Selection (PLS)*. arXiv preprint `https://arxiv.org/pdf/2302.08883v2.pdf`. 2023.

[6] Julian Rodemann, Christoph Jansen, Georg Schollmeyer, and Thomas Augustin. *In all Likelihoods: How to Reliably Select Pseudo-Labeled Data for Self-Training in Semi-Supervised Learning*. under review. 2023.

[7] Teddy Seidenfeld. "A contrast between two decision rules for use with (convex) sets of probabilities: Γ-maximin versus E-admissibility". In: *Synthese* 140.1/2 (2004), pp. 69–88.

[8] Matthias C.M. Troffaes. "Decision making under uncertainty using imprecise probabilities". In: *International Journal of Approximate Reasoning* 45.1 (2007), pp. 17–29.