Data Conversations Seminar
at IBM Bangalore

Jan. 24, 2023

Improving Robustness in Data Centric Machine Learning

Masashi Sugiyama

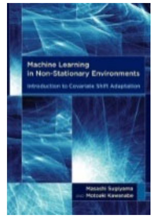
RIKEN Center for Advanced Intelligence Project/
The University of Tokyo

http://www.ms.k.u-tokyo.ac.jp/sugi/

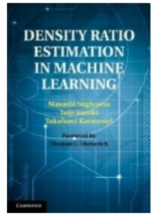
About Myself

- **Masashi Sugiyama**:
  - Director: RIKEN AIP, Japan
  - Professor: University of Tokyo, Japan
  - Consultant: several local startups
- **Interests: Machine learning (ML)**
  - ML theory & algorithm →
  - ML applications (signal, image, language, brain, robot, mobility, advertisement, biology, medicine, education…)
- **Academic activities:**
  - Program Chairs for NeurIPS2015, AISTATS2019, ACML2010/2020…

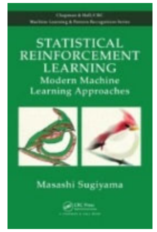
Sugiyama & Kawanabe, Machine Learning in Non-Stationary Environments, MIT Press, 2012



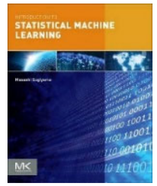
Sugiyama, Suzuki & Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012



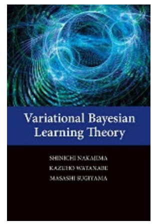
Sugiyama, Statistical Reinforcement Learning, Chapman and Hall/CRC, 2015



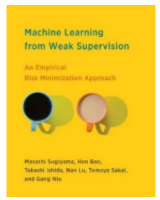
Sugiyama, Introduction to Statistical Machine Learning, Morgan Kaufmann, 2015



Nakajima, Watanabe & Sugiyama, Variational Bayesian Learning Theory, Cambridge University Press, 2019



Sugiyama, Bao, Ishida, Lu, Sakai & Niu. Machine Learning from Weak Supervision, MIT Press, 2022.



What is “RIKEN”?

- Name in Japanese: 理化学研究所
  - Pronounced as: rikagaku kenkyusho
  - Meaning: Physics and Chemistry Research Institute
- Acronym in Japanese: 理研 (RIKEN)

What is RIKEN-AIP?

MEXT Advanced Intelligence Project (2016-2025):

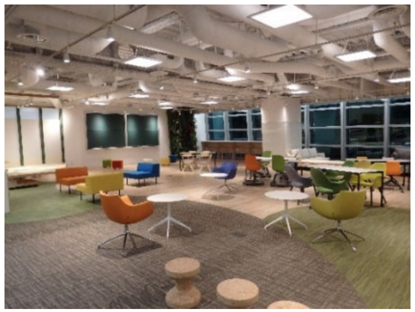
- 130 employed researchers (36% international, 23% female)
- 200 visiting researchers, 100 domestic students
- 140 international interns (total)

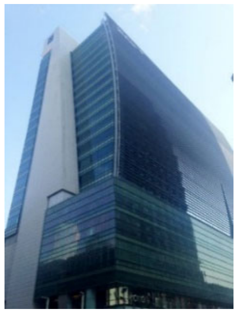
MEXT
MINISTRY OF EDUCATION, CULTURE, SPORTS, SCIENCE AND TECHNOLOGY-JAPAN

Missions:

- Develop new AI technology (ML, Opt, math)
- Accelerate scientific research (cancer, material, genomics)
- Solve socially critical problems (disaster, elderly healthcare)
- Study of ELSI in AI (ethical guidelines, personal data)
- Human resource development (researchers, engineers)

Main office in the heart of Tokyo



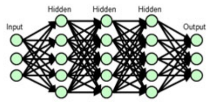


Distributed offices across Japan

Selected Research

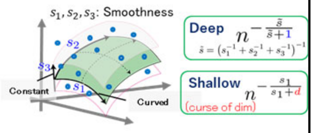
Developing New AI Technology

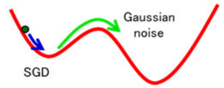
- **Theory of deep learning**:
  - Better prediction than shallow learning
  - No curse of dimensionality
  - Global optimization

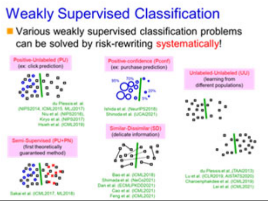


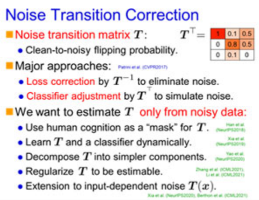
$$\mathbb{E}[\|\hat{f}_T - f^*\|_{L_2}^2] \leq \epsilon_M + O(T^{-\frac{2r\beta}{2r\beta+1}})$$

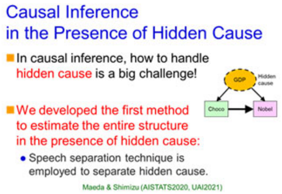
- **Developing new methods**:
  - Weakly supervised learning
  - Noise robust learning
  - Causal inference







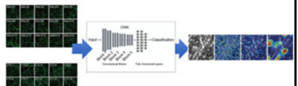




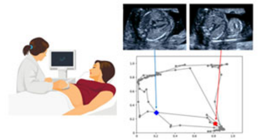
Accelerating Scientific Research

- **Medical science**:
  - Prostate/pancreatic cancer detection
  - ALS early diagnosis
  - Fetal heart screening
  - Colonoscopy

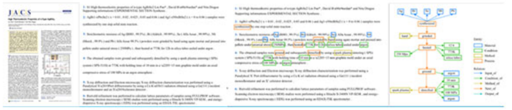
*Yoichiro Yamamoto, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nature Communications* 10:5642, 2019.



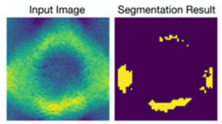
Imamura et al., Ann Neurology, 2021

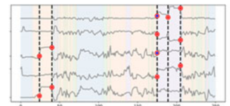


- **Material science**:
  - Database creation with text mining



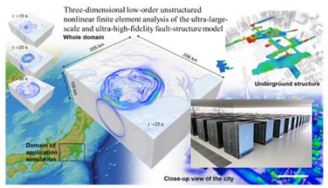
- **Data-driven science**:
  - Selective inference for reliability evaluation

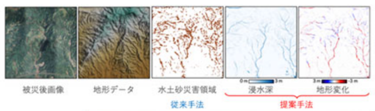




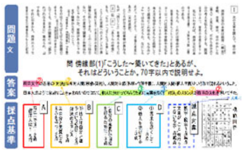
Solving Socially Critical Problems

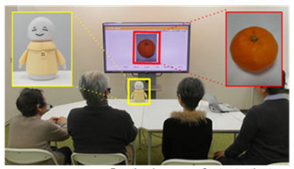
- **Natural disaster**:
  - Fugaku-based earthquake simulation
  - Remote sensing disaster analysis

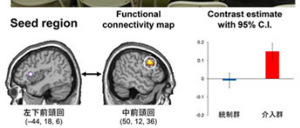




- **Elderly healthcare**:
  - Chat-robot-guided cognitive function improvement
- **Education**:
  - Automatic essay evaluation
  - Interactive essay writing support



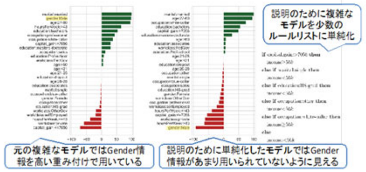


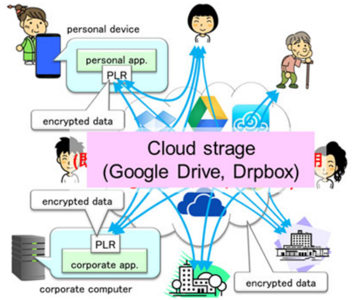


Studying AI-ELSI

- **AI Ethical guidelines**:
  - Japanese Society for AI, Ministry of Internal Affairs and Communications, Cabinet Office
  - IEEE, G20, OECD

- **Personal data management**:
  - Individual-based accessibility control system
- **AI security and reliability**:
  - Adversarial attack/defense
  - Fairness faking/guarantee





Contents

Robust Machine Learning

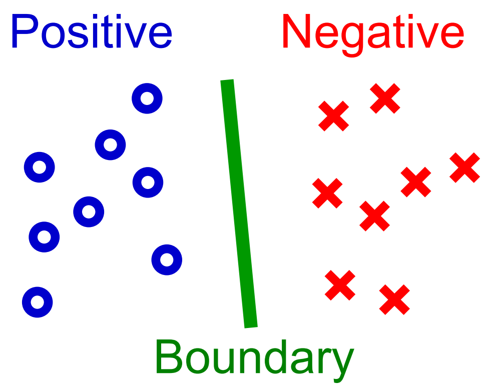
- Goal: Develop novel ML theories and algorithms that enable reliable learning from limited information.
  - Insufficient information: weak supervision.
  - Data bias: changing environments, privacy.
  - Label noise: human error, sensor error.
  - Attack: adversarial noise, distribution shift.

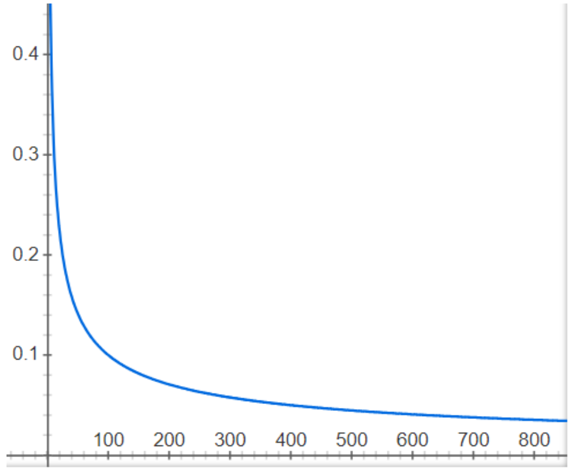
Contents

ML from Limited Data

- ML from big labeled data is successful.
  - Speech, image, language, advertisement,…
  - Estimation error of the boundary decreases in order $1/\sqrt{n}$.

n : Number of labeled samples



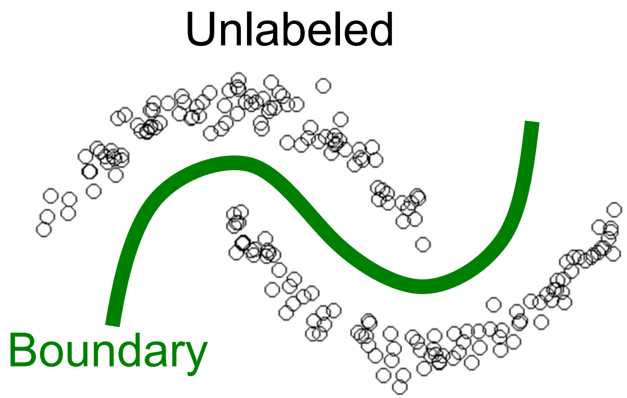


- However, there are various applications where big labeled data is not available.
  - Medicine, disaster, robots, brain, …

Alternatives to Supervised Classification

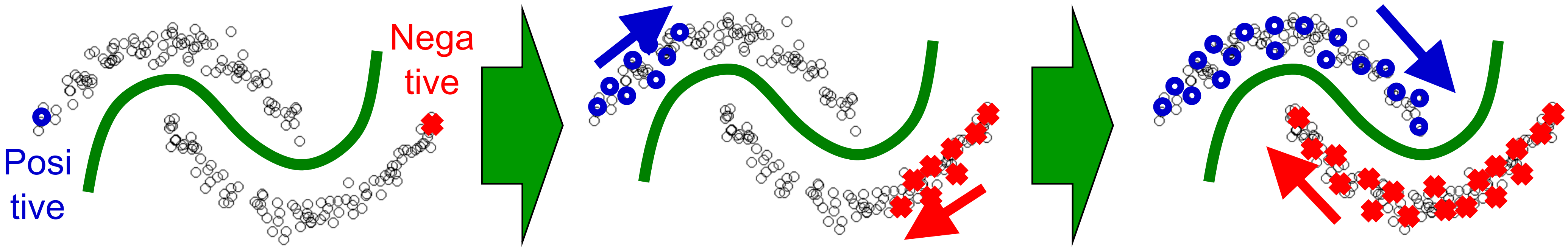
Unsupervised classification:

- No label is used.
- Essentially clustering.
- No guarantee for prediction.



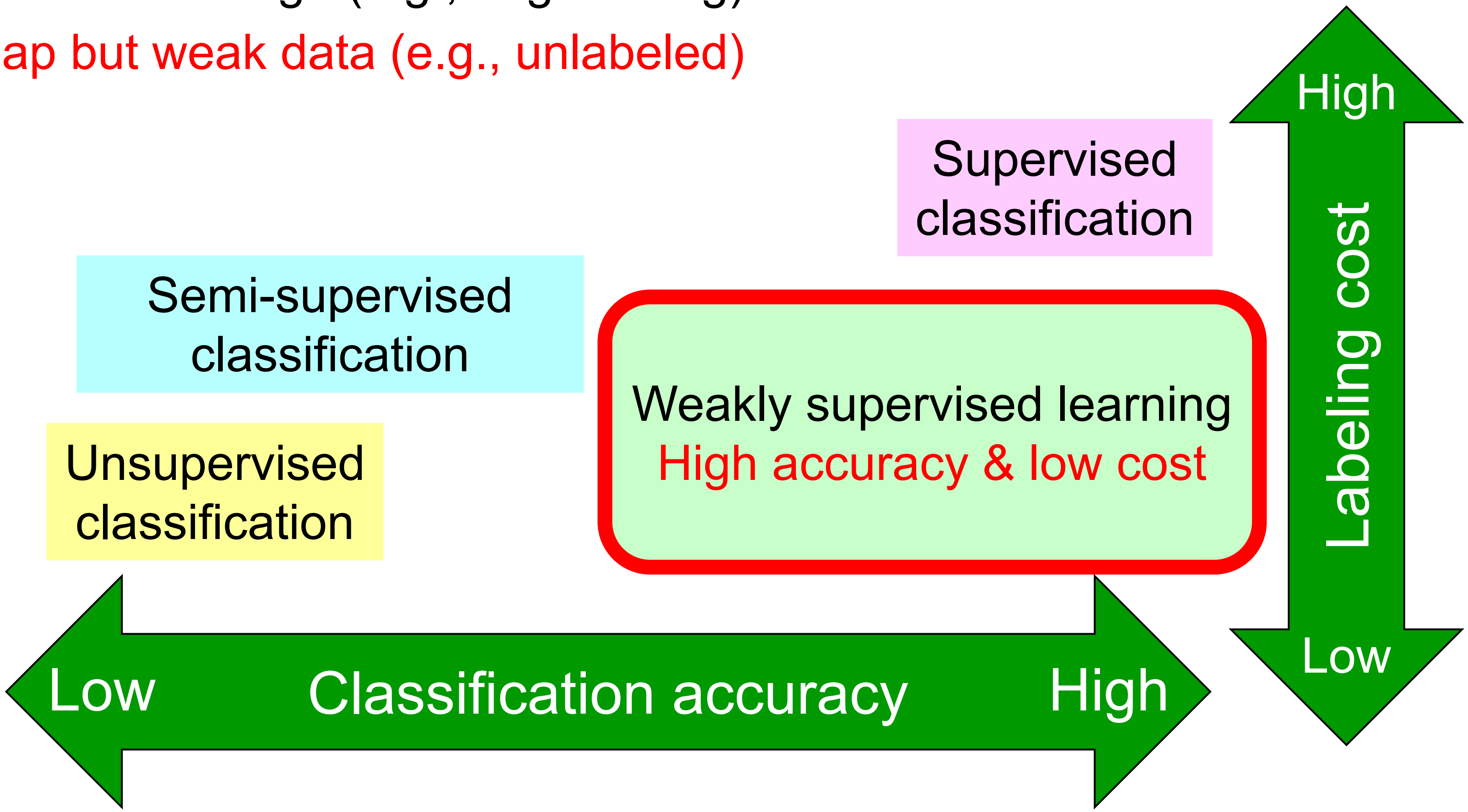
Semi-supervised classification:

- Additionally use a small amount of labeled data.
- Propagate labels along clusters.
- No guarantee for prediction.



Weakly Supervised Learning

- Coping with labeling cost:
  - Improve data collection (e.g., crowdsourcing)
  - Use a simulator to generate pseudo data (e.g., physics, chemistry, robotics, etc.)
  - Use domain knowledge (e.g., engineering)
  - Use cheap but weak data (e.g., unlabeled)



Positive-Unlabeled Classification

- **Given:** Positive and unlabeled samples

$$\{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}|y=+1)$$

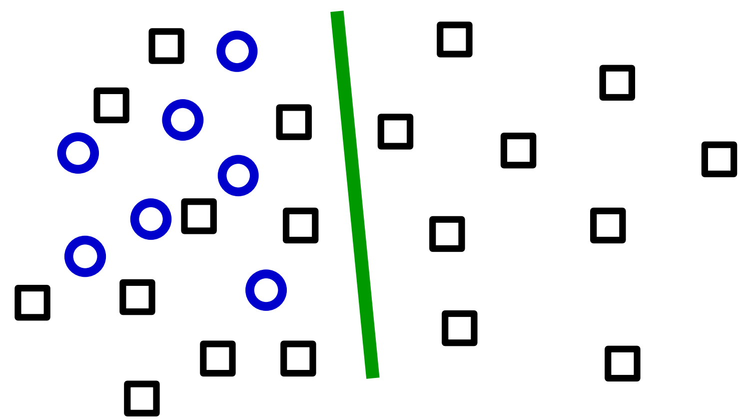
$$\{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

- **Goal:** Obtain a PN classifier

Example: Ad-click prediction

- Clicked ad: User likes it → P
- Unclicked ad: User dislikes it or User likes it but doesn't have time to click it → U (=P or N)

Positive



Unlabeled (mixture of positives and negatives)

Solution (Sketch)

- Given: Positive and unlabeled data

du Plessis, Niu & Sugiyama (NIPS2014, ICML2015)

$$\{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}|y=+1) \qquad \{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

- Decomposition of the classification risk:

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\Big[\ell\Big(yf(\boldsymbol{x})\Big)\Big]$$

ℓ : loss

$\pi = p(y=+1)$: Class prior (assumed known)

$$= \underbrace{\pi\mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big]}_{\text{Risk for positive data}} + \underbrace{(1-\pi)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]}_{\text{Risk for negative data}}$$

- Eliminate the expectation over negative data as

$$\mathbb{E}_{p(\boldsymbol{x})}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big] - \pi\mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

$$p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$$

- Unbiased risk estimation:

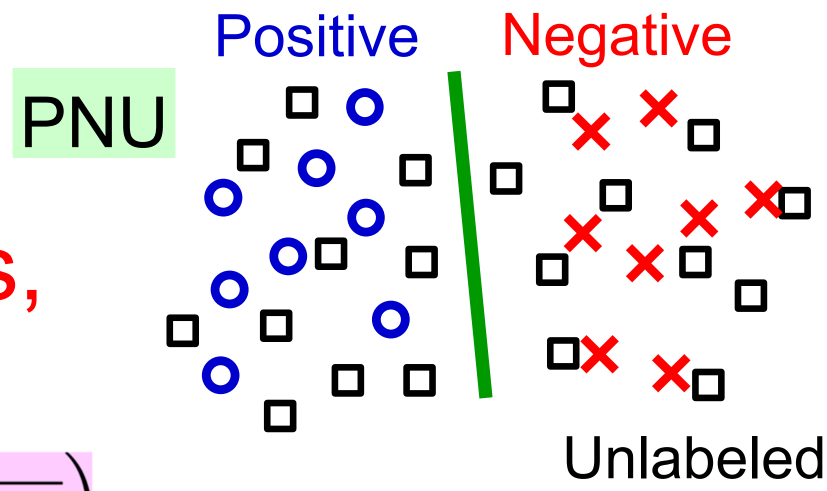
$$\mathcal{O}_p\Big(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{U}}}\Big)$$

$$\widehat{R}_{\mathrm{PU}}(f) = \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(f(\boldsymbol{x}_i^{\mathrm{P}})\Big) + \frac{1}{n_{\mathrm{U}}}\sum_{j=1}^{n_{\mathrm{U}}}\ell\Big(-f(\boldsymbol{x}_j^{\mathrm{U}})\Big) - \frac{\pi}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\Big(-f(\boldsymbol{x}_i^{\mathrm{P}})\Big)$$

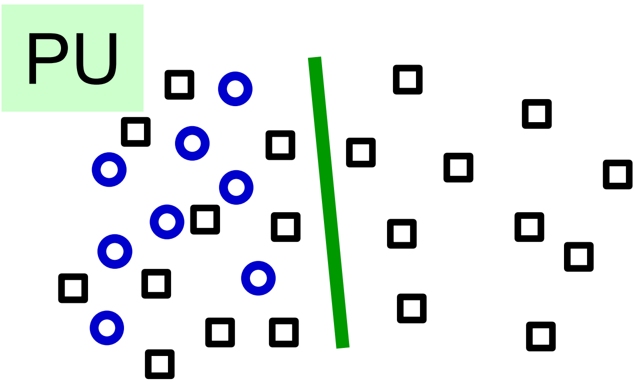
Positive-Negative-Unlabeled Classification (Semi-Supervised Classification)

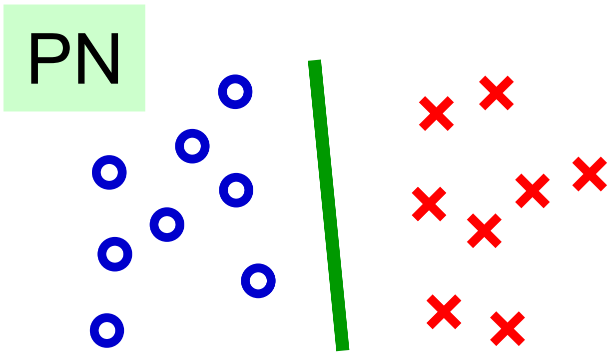
Sakai, du Plessis, Niu & Sugiyama (ICML2017)

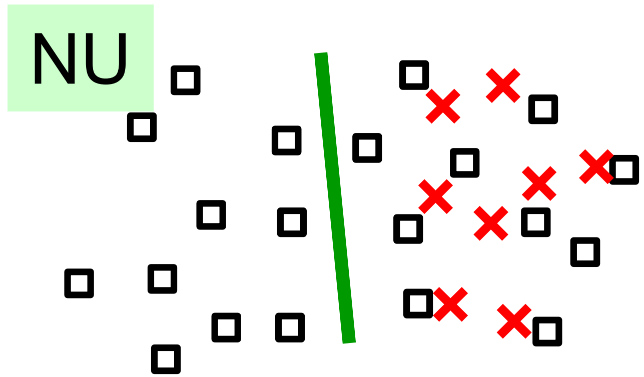
- Let's decompose PNU into PU, PN, and NU:
  - Each is solvable.
  - Let's combine them!
- Without cluster assumptions, PN classifiers are trainable!



$$\mathcal{O}_p\left(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{N}}} + 1/\sqrt{n_{\mathrm{U}}}\right)$$



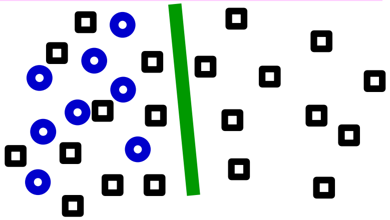




Various Extensions

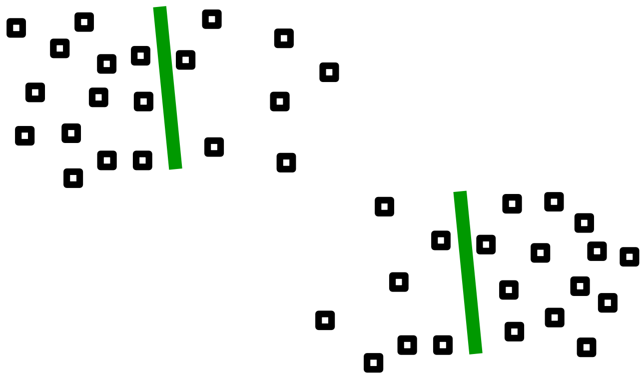
- Learning from weakly supervised data is possible in many different forms!

Positive-Unlabeled



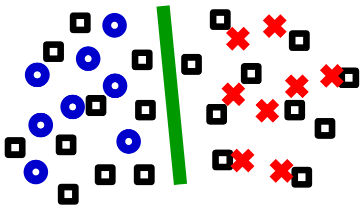
du Plessis et al. (NIPS2014, ICML2015, MLJ2017)
Niu et al. (NIPS2016),, Kiryo et al. (NIPS2017)
Hsieh et al. (ICML2019)

Unlabeled-Unlabeled



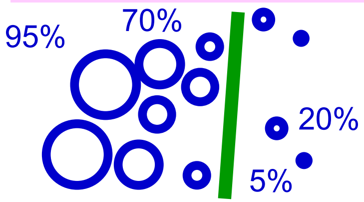
du Plessis et al.,(TAAI2013)
Lu et al. (ICLR2019, AISTATS2020)
Charoenphakdee et al. (ICML2019)
Lei et al. (ICML2021)

Semi-Supervised



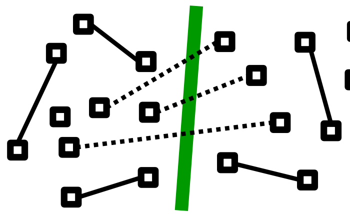
Sakai et al. (ICML2017, ML2018)

Positive-confidence



Ishida et al. (NeurIPS2018)
Shinoda et al. (IJCAI2021)

Similar-Dissimilar

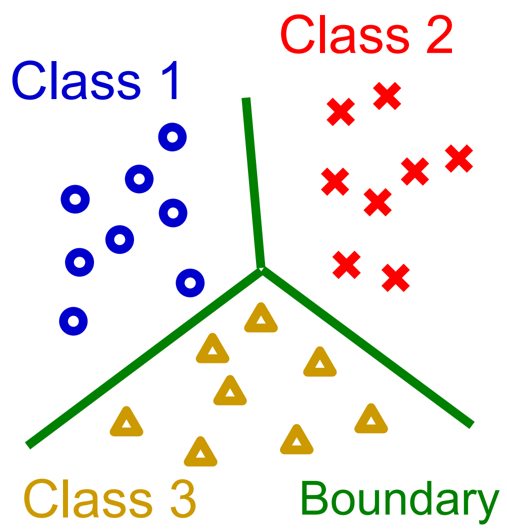


Bao et al. (ICML2018)
Shimada et al. (NeCo2021)
Dan et al. (ECMLPKDD2021)
Cao et al. (ICML2021)
Feng et al. (ICML2021)

- All are loss-correction based and consistent. $\mathcal{O}_p\left(1/\sqrt{n}\right)$
- Any loss, classifier, and optimizer can be used.

Multiclass Methods

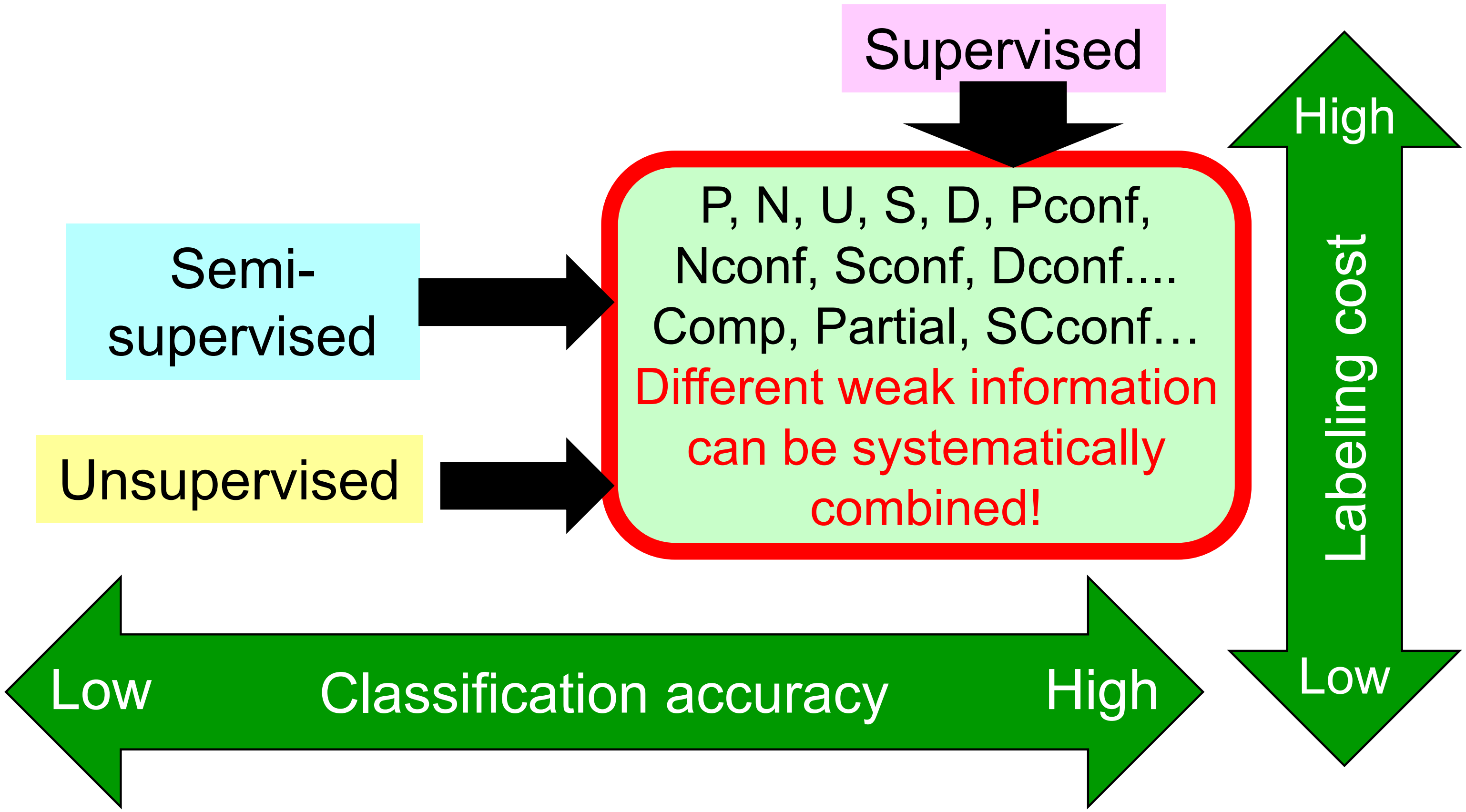
- Labeling patterns in multi-class problems is extremely painful.



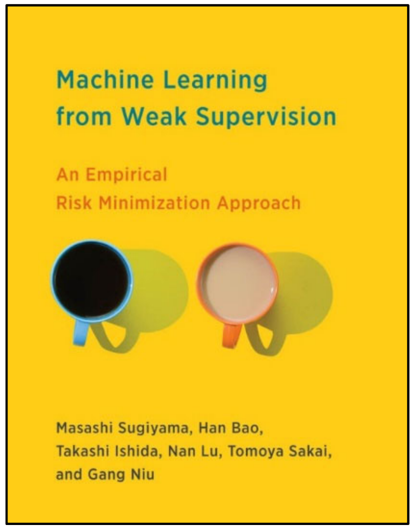
- Multi-class weak-labels:
  - Complementary labels: Specify a class that a pattern does not belong to ("not 1"). Ishida et al. (NIPS2017, ICML2019) Chou et al. (ICML2020)
  - Partial labels: Specify a subset of classes that contains the correct one ("1 or 2"). Feng et al. (ICML2020, NeurIPS2020) Lv et al. (ICML2020)
  - Single-class confidence: Cao et al. (arXiv2021) One-class data with full confidence ("1 with 60%, 2 with 30%, and 3 with 10%")
- Systematic loss correction is possible! $\mathcal{O}_p\left(1/\sqrt{n}\right)$

Summary: Weakly Supervised Learning

- We developed an empirical risk minimization framework for weakly supervised learning:
  - Any loss, classifier, and optimizer can be used.
  - Statistical consistency with optimal convergence.



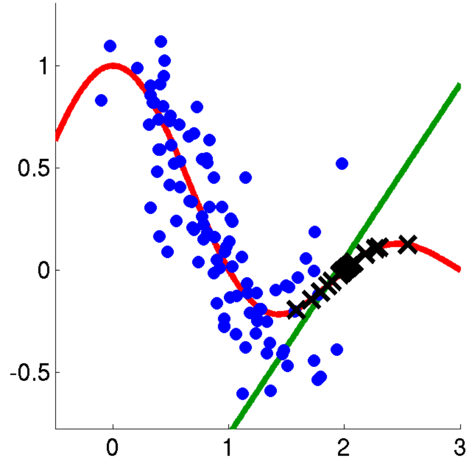
Sugiyama, Bao, Ishida, Lu, Sakai & Niu, Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach. MIT Press, August 2022.



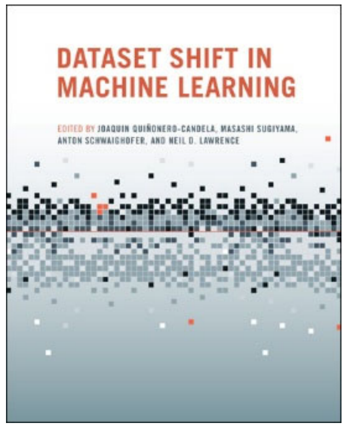
Contents

Transfer Learning

- Training and test data often have different distributions, due to
  - changing environments,
  - sample selection bias (privacy).



- Transfer learning (domain adaptation):
  - Train a test-domain predictor using training data from different domains.



Quiñonero-Candela et al. (MIT Press 2009)

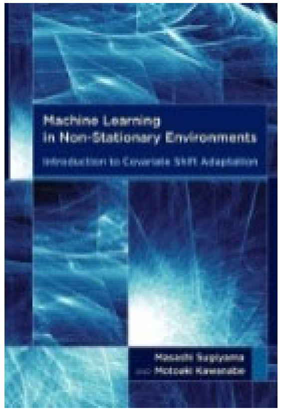
Classical Approach for Transfer Learning

- Two-step adaptation:
  1. Importance weight estimation:

$$\widehat{w} = \mathop{\mathrm{argmin}}_{w} \widehat{\mathbb{E}}_{p_{\mathrm{tr}}(\boldsymbol{x},y)} \left[D\left(w(\boldsymbol{x},y), \frac{p_{\mathrm{te}}(\boldsymbol{x},y)}{p_{\mathrm{tr}}(\boldsymbol{x},y)} \right) \right]$$

  2. Weighted predictor training:

$$\widehat{f} = \mathop{\mathrm{argmin}}_{f} \widehat{\mathbb{E}}_{p_{\mathrm{tr}}(\boldsymbol{x},y)} [\widehat{w}(\boldsymbol{x},y)\ell(f(\boldsymbol{x}),y)]$$



Sugiyama & Kawanabe
(MIT Press 2012)

- However, estimation error in Step 1 is not taken into account in Step 2.
  - We want to integrate these two steps!

Joint Weight-Predictor Optimization

- Covariate shift: Only input distributions change.

$$p_{\mathrm{tr}}(\boldsymbol{x}) \neq p_{\mathrm{te}}(\boldsymbol{x}) \qquad p_{\mathrm{tr}}(y|\boldsymbol{x}) = p_{\mathrm{te}}(y|\boldsymbol{x})$$

Shimodaira (JSPI2000)

- Suppose we are given
  - Labeled training data: $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y)$
  - Unlabeled test data: $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{te}}(\boldsymbol{x})$
- Minimize a risk upper bound jointly w.r.t. weight w and predictor f: $J_{\ell_{\mathrm{tr}}}(f, w) \geq R_{\ell_{\mathrm{te}}}(f)^2$

Zhang et al. (ACML2020, SNCS2021)

$$\widehat{f} = \operatorname*{argmin}_{f} \min_{w \geq 0} \widehat{J}_{\ell_{\mathrm{tr}}}(f, w)$$

$\widehat{J}_\ell$: Empirical approximation of J_ℓ

$$R_\ell(f) = \mathbb{E}_{p_{\mathrm{te}}(\boldsymbol{x},y)}[\ell(f(\boldsymbol{x}), y)]$$

$$\ell_{\mathrm{te}} \leq 1, \ell_{\mathrm{tr}} \geq \ell_{\mathrm{te}}$$

  - Theoretical guarantee:

$$R_{\ell_{\mathrm{te}}}(\widehat{f}) \leq \sqrt{2} \min_{f} R_{\ell_{\mathrm{te}}}(f) + \mathcal{O}_p(n_{\mathrm{tr}}^{-1/4} + n_{\mathrm{te}}^{-1/4})$$

Dynamic Importance Weighting

- General changing distributions: $p_{\mathrm{tr}}(\boldsymbol{x}, y) \neq p_{\mathrm{te}}(\boldsymbol{x}, y)$
- Suppose we are given
  - Labeled training data: $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y)$
  - Labeled test data: $\{(\boldsymbol{x}_i^{\mathrm{te}}, y_i^{\mathrm{te}})\}_{i=1}^{n_{\mathrm{te}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{te}}(\boldsymbol{x}, y)$
- For each mini-batch $\{(\bar{\boldsymbol{x}}_i^{\mathrm{tr}}, \bar{y}_i^{\mathrm{tr}})\}_{i=1}^{\bar{n}_{\mathrm{tr}}}, \{(\bar{\boldsymbol{x}}_i^{\mathrm{te}}, \bar{y}_i^{\mathrm{te}})\}_{i=1}^{\bar{n}_{\mathrm{te}}}$, importance weights are estimated by matching losses by kernel mean matching:

Fang et al. (NeurIPS2020)

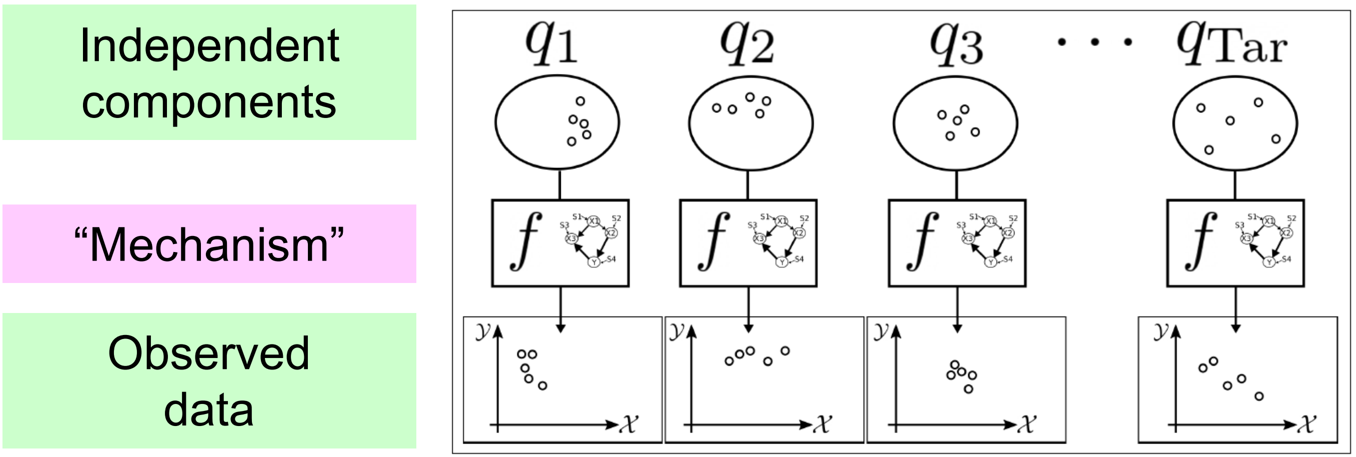
Huang et al. (NeurIPS2007)

$$\frac{1}{\bar{n}_{\mathrm{tr}}} \sum_{i=1}^{\bar{n}_{\mathrm{tr}}} r_i \ell(f(\bar{\boldsymbol{x}}_i^{\mathrm{tr}}), \bar{y}_i^{\mathrm{tr}}) \approx \frac{1}{\bar{n}_{\mathrm{te}}} \sum_{j=1}^{\bar{n}_{\mathrm{te}}} \ell(f(\bar{\boldsymbol{x}}_j^{\mathrm{te}}), \bar{y}_j^{\mathrm{te}})$$

- Extremely simple, but highly powerful!

Summary

- In transfer learning with importance weighting, simultaneously performing importance estimation and predictor training is promising.
- What should we do if training and test distributions look very different?
  - Mechanism transfer! Teshima, Sato & Sugiyama (ICML2020)



Bai, Zhang, Zhao, Sugiyama & Zhou (NeurIPS2022)

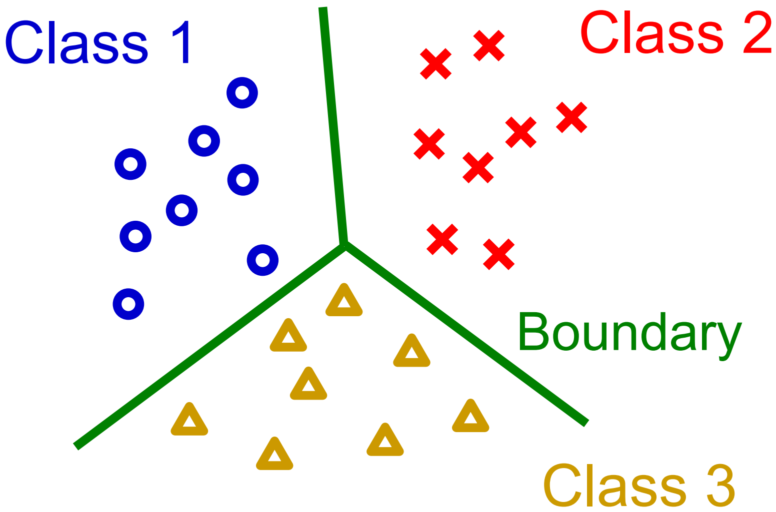
- Current challenge: Continuous distribution change

Contents

1. Introduction of RIKEN-AIP
2. Robust Machine Learning
   A) Weakly Supervised Learning
   B) Transfer Learning
   C) Noise-Robust Learning
3. Summary

Supervised Classification

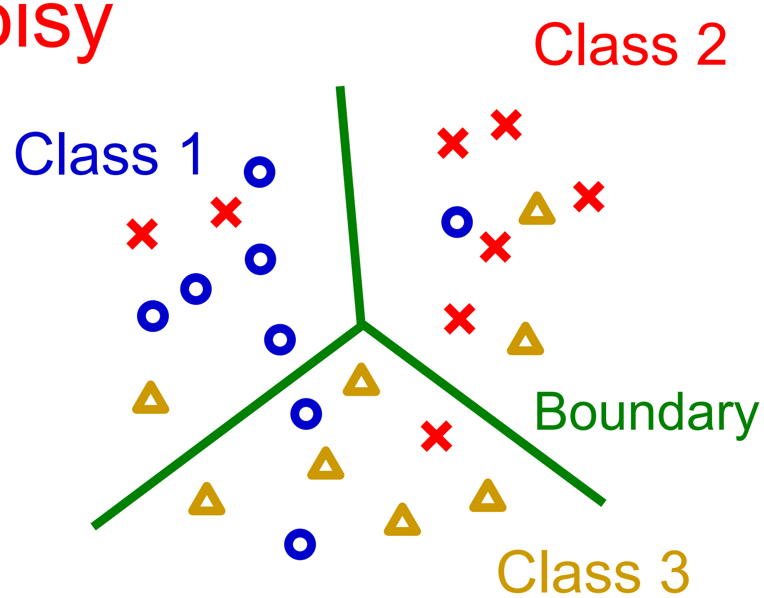
- Supervised classification with clean labels:



Training error minimization is statistically consistent and work well in practice.

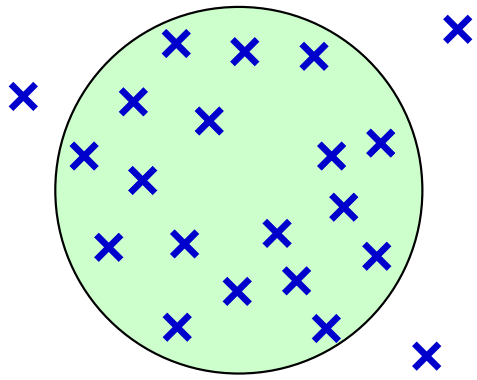
- However, real-world labels are noisy possibly due to human error:

Training error minimization is no longer consistent and does not work well in practice.

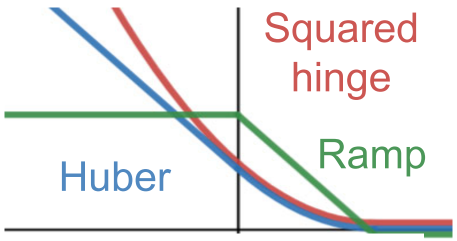


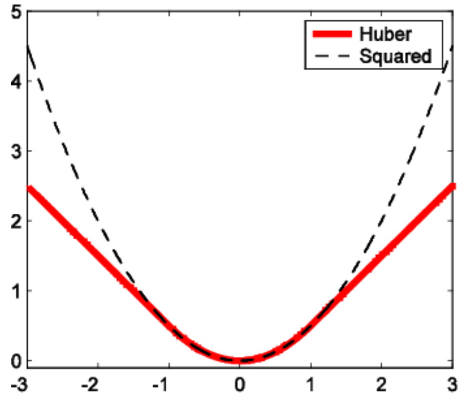
Classical Approaches

- **Unsupervised outlier removal:**
  - Substantially more difficult than classification.

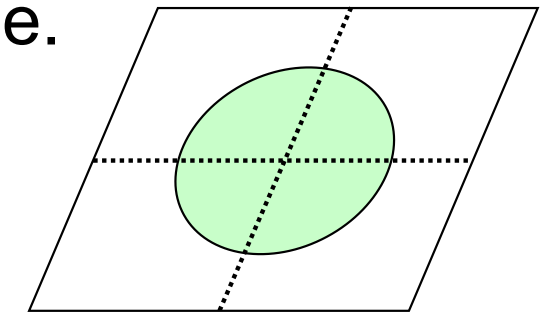


- **Robust loss:**
  - Works well for regression, but limited effectiveness for classification.

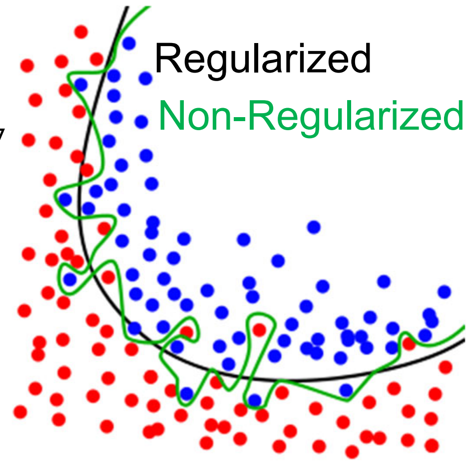




- **Regularization:**
  - Effective in suppressing overfitting, but too smooth for strong noise.



ℓ_2-regularization



https://en.wikipedia.org/wiki/Overfitting

- **Need new approaches!**

Noise Transition Correction

- **Noise transition matrix $\boldsymbol{T}$:**
  - Clean-to-noisy flipping probability.

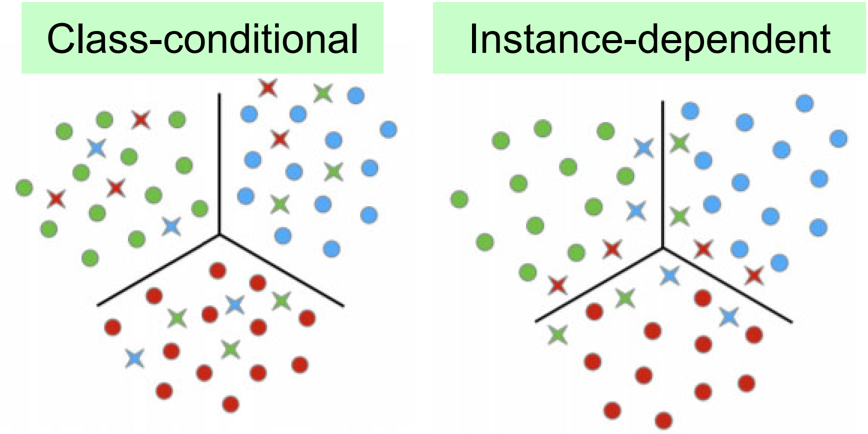
$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

- **Major approaches:** Patrini et al. (CVPR2017)
  - Loss correction by $\boldsymbol{T}^{-1}$ to eliminate noise.
  - Classifier adjustment by $\boldsymbol{T}^{\top}$ to simulate noise.

- **We want to estimate $\boldsymbol{T}$ only from noisy data:**
  - Use human cognition as a "mask" for $\boldsymbol{T}$. Han, Yao, Niu, Zhou, Tsang, Zhang & Sugiyama (NeurIPS2018)
  - Reduce estimation error of $\boldsymbol{T}$. Xia, Liu, Wang, Han, Gong, Niu & Sugiyama (NeurIPS2019) Yao, Liu, Han, Gong, Deng, Niu, Sugiyama & Tao (NeurIPS2020)
  - Learn $\boldsymbol{T}$ and classifier simultaneously. Zhang, Niu & Sugiyama (ICML2021)
  - Estimate $\boldsymbol{T}$ under weaker conditions. Li, Liu, Han, Niu & Sugiyama (ICML2021)

Beyond Class-Conditional Noise

- Real-world noise may be instance-dependent:
  - Ex.: Noise is large near the boundary.



- Instance-dependent noise: $T_{y,\bar{y}}(\boldsymbol{x}) = \bar{p}(\bar{y}|y, \boldsymbol{x})$
  - Extremely challenging to estimate the noise transition matrix function!

- Various heuristic solutions:
  - Parts-based estimation.
  - Use of additional confidence scores.
  - Manifold regularization.

Xia, Liu, Han, Wang, Gong, Liu, Niu, Tao & Sugiyama (NeurIPS2020)

Berthon, Han,Niu, Liu & Sugiyama (ICML2021)

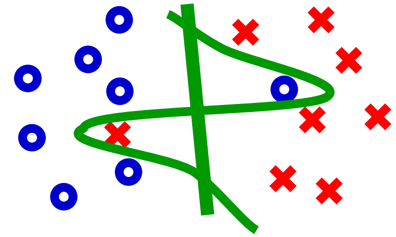
Cheng, Liu, Ning, Wang, Han, Niu, Gao & Sugiyama (CVPR2022)

Co-teaching

Memorization of neural nets:

Arpit et al. (ICML2017)
Zhang et al. (ICLR2017)

- Stochastic gradient descent fits clean data faster.
- However, naïve early stopping does not work well.



“Co-teaching” between two neural nets:

- Teach small-loss data each other.

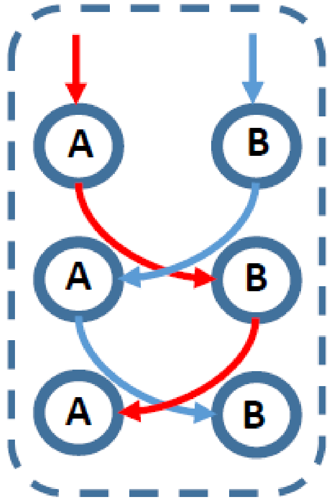
  Han, Yao, Yu, Niu, Xu, Hu, Tsang & Sugiyama (NeurIPS2018)

- Teach only disagreed data.

  Yu, Han, Yao, Niu, Tsang & Sugiyama (ICML2019)

- Gradient ascent for large-loss data.

  Han, Niu, Yu, Yao, Xu, Tsang & Sugiyama (ICML2020)



No theory but very robust in experiments:

- Works well even if 50% random label flipping!

Contents

1. Introduction of RIKEN-AIP
2. Robust Machine Learning
   A) Weakly Supervised Learning
   B) Transfer Learning
   C) Noise-Robust Learning
3. Summary

Challenges in Reliable ML

- Reliability for expectable situations:
  - Model the corruption process explicitly and correct the solution.
    - How to handle modeling error?
- Reliability for unexpected situations:
  - Consider worst-case robustness ("min-max").
    - How to make it less conservative?
  - Include human support ("rejection").
    - How to handle real-time applications?
- Exploring somewhere in the middle would be practically more useful:
  - Use partial knowledge of the corruption process.