AI Colloquium@KAIST

Sep. 22, 2020

Robust Machine Learning for Reliable Deployment

Masashi Sugiyama

RIKEN Center for Advanced Intelligence Project/
The University of Tokyo

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

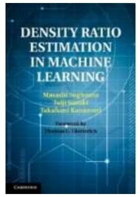
About Myself

- **My jobs:**
  - Director: RIKEN AIP
  - Professor: University of Tokyo
  - Consultant: several local startups

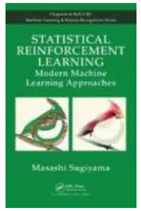
- **Interests: Machine learning (ML)**
  - Weakly-supervised learning,
  - Robust learning,
  - Transfer learning,
  - Density ratio estimation,
  - Reinforcement learning,
  - Variational inference…
- **Academic activities:**
  - PC Chairs for NeurIPS2015, AISTATS2019, ACML2010/2020…

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

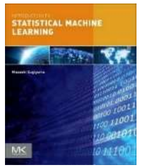
Sugiyama, Suzuki & Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012



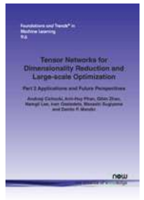
Sugiyama, Statistical Reinforcement Learning, Chapman and Hall/CRC, 2015



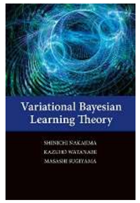
Sugiyama, Introduction to Statistical Machine Learning, Morgan Kaufmann, 2015



Cichocki, Phan, Zhao, Lee, Oseledets, Sugiyama & Mandic, Tensor Networks for Dimensionality Reduction and Large-Scale Optimizations, Now, 2017

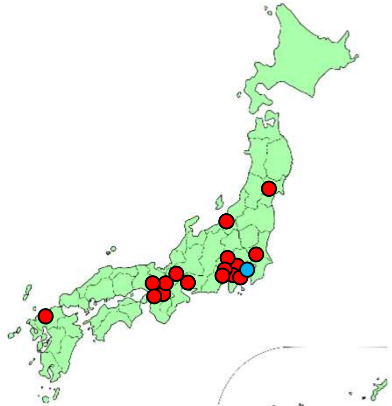


Nakajima, Watanabe & Sugiyama, Variational Bayesian Learning Theory, Cambridge University Press, 2019



RIKEN Center for Advanced Intelligence Project (AIP)

- 10-year national project in Japan (2016-2025):
  - Develop next-generation AI technology (learning and optimization theory, etc.)
  - Accelerate scientific research (material, cancer, stem cells, genomics, etc.)
  - Solve socially critical problems (natural disaster, elderly healthcare, etc.)
  - Study of ethical, legal and social issues of AI (ethical guideline, privacy protection, etc.)
  - Human resource development (150+ researchers, 200+ students, 150+ interns, 300+ visiting scientists, 40+ industry projects)



Contents

ML Conferences

- ICML: International Conference on Machine Learning (since 1980)
  - Conference on learning from data.
  - Top statistical ML conference since around 2000.

- NeurIPS: Neural Information Processing Systems (since 1987)
  - Originally neuro-inspired AI conference.
  - Top statistical ML conference since around 2000.
  - Neuro/cognitive science papers are also accepted.

Conference Statistics

- Rapid increase in size:

ICML	2013	2014	2015	2016	2017	2018	2019	2020
Participants	900	1200	1600	3000+	2400	5000	6200	???
Submitted papers	1204	1238	1037	1327	1701	2473	3424	4990
Accepted papers	283	310	270	322	433	618	773	1088

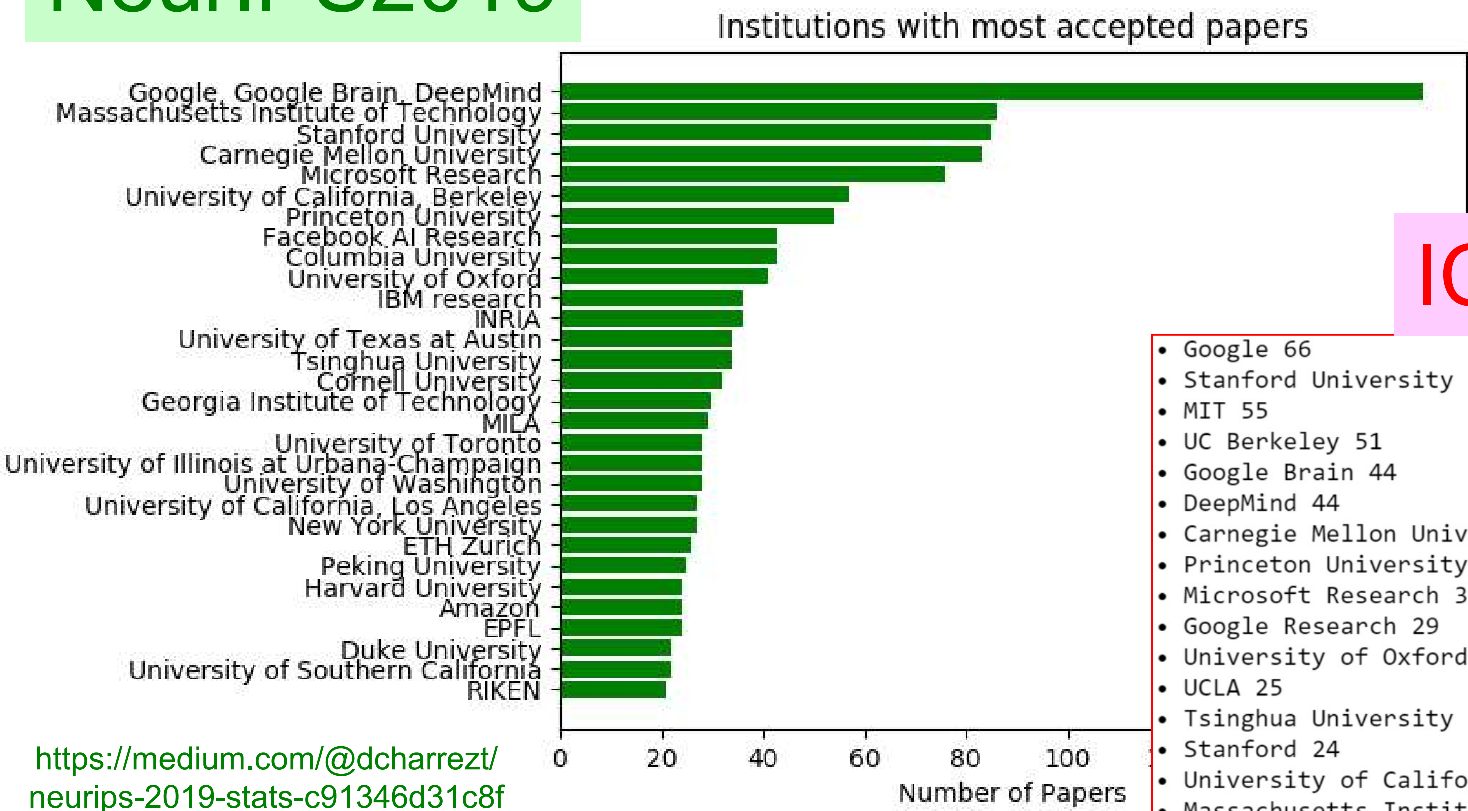
NeurIPS	2013	2014	2015	2016	2017	2018	2019	2020
Participants	1200	2400	3800	6000+	7500+	8000+	13000+	???
Submitted papers	1420	1678	1838	2500	3240	4856	6743	9467
Accepted papers	360	414	403	568	678	1011	1428	???

- Company sponsoring is very active:
  - Around 2000: US-based IT giants
  - Around 2010: Worldwide IT giants
  - Recently: Diverse companies from startups to giants and from IT to various non-IT

Recent Trends

- North-American companies and universities dominate.

NeurIPS2019



https://medium.com/@dcharrezt/neurips-2019-stats-c91346d31c8f

ICML2020

- Google 66
- Stanford University 58
- MIT 55
- UC Berkeley 51
- Google Brain 44
- DeepMind 44
- Carnegie Mellon University 37
- Princeton University 35
- Microsoft Research 35
- Google Research 29
- University of Oxford 25
- UCLA 25
- Tsinghua University 24
- Stanford 24
- University of California, Berkeley 22
- Massachusetts Institute of Technology 22
- Harvard University 21
- Duke University 21
- University of Washington 20
- University of Pennsylvania 20
- Facebook AI Research 20
- Cornell University 20
- RIKEN 18

https://twitter.com/SergeyI49013776/status/1267768532529557504

NeurIPS2015 vs. 2019

- 2015: ML technology was the main concern.
  - Futuristic technologies such as AlphaGo, autonomous driving cars, and chat robots emerged.
  - Expectation for more advanced ML technologies.
  - US-based companies dominated AI business.

- 2019: Social impact of ML is a serious concern.
  - Social issues: privacy, fairness, explanability,…
  - ML-driven science: chemistry, biology, medicine,...
  - US and Chinese companies are competing.
  - Minority support: Women, Black, LatinX, Queer,…

Contents

Today's Topic: Robust Machine Learning

- In real-world applications, it becomes increasingly important to consider robustness:
  - Noise: sensor error, human error
  - Insufficient information: weak supervision
  - Bias: sample selection bias, changing environments
  - Attack: adversarial noise, distribution shift
- In this lecture, I will give an overview of our recent advances in robust machine learning.

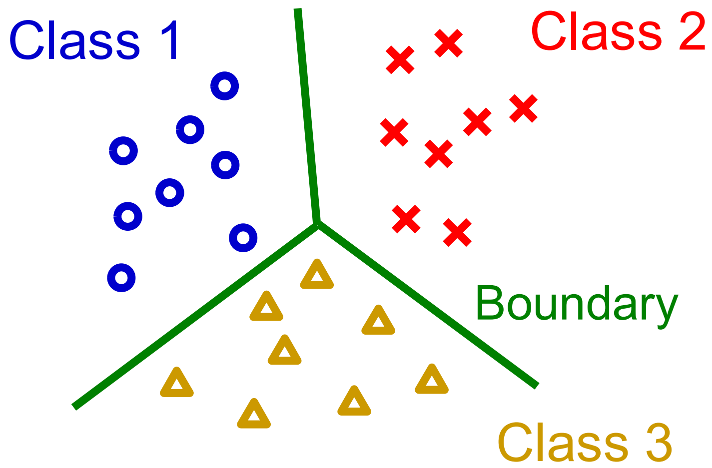
http://www.ms.k.u-tokyo.ac.jp/sugi/publications.html

Contents

1. Trend in ML Research
2. Robust Machine Learning
   - A) Noisy label learning
   - B) Weakly supervised learning
   - C) Bias in training data
   - D) Noise in test input
3. Future ML Research

Ordinary Classification

- Clean training data: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$



$\boldsymbol{x} \in \mathbb{R}^d$: Input pattern

$y \in \{1, \ldots, c\}$: Clean class label (not necessarily separable)

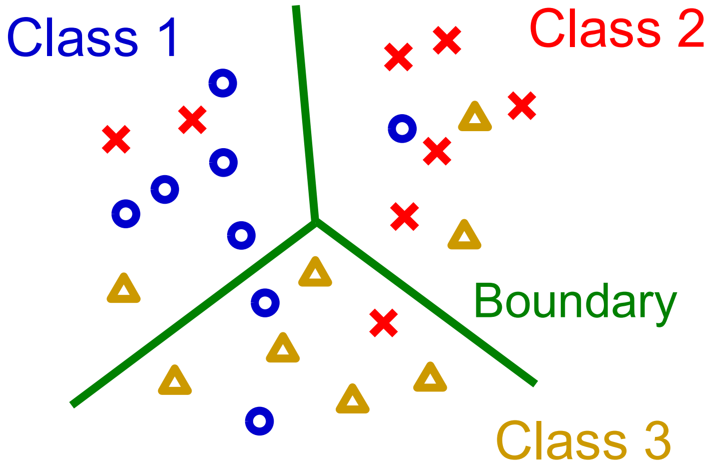
- Training error minimization is statistically consistent and work well:

$$\frac{1}{n}\sum_{i=1}^{n} \ell\Big(y_i, \boldsymbol{g}(\boldsymbol{x}_i)\Big)$$

$\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^c$: Classifier

Noisy Classification

- Noisy training data: $\{(\boldsymbol{x}_i, \widetilde{y}_i)\}_{i=1}^{n}$



$\boldsymbol{x} \in \mathbb{R}^d$: Input pattern

$\widetilde{y} \in \{1, \ldots, c\}$: Noisy class label (clean labels are corrupted)

- Training error minimization is no longer consistent and does not work well:

$$\frac{1}{n}\sum_{i=1}^{n} \ell\Big(\widetilde{y}_i, \boldsymbol{g}(\boldsymbol{x}_i)\Big)$$

$\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^c$: Classifier

Standard Approaches

- Unsupervised outlier removal:
  - Substantially difficult
- Robust loss, regularization:
  - Not robust enough
- We want to go beyond the limitations of existing approaches!
  - Noise transition correction
  - Noiseless sample selection
  - Model capacity control

Noise Transition Correction

- Noise transition matrix $\boldsymbol{T}$:
  - Flipping probability from y to $\widetilde{y}$.

$$\boldsymbol{T}^\top = \begin{array}{c} y \\ \begin{bmatrix} 1 & 0.1 & 0.5 \\ 0 & 0.8 & 0.5 \\ 0 & 0.1 & 0 \end{bmatrix} \widetilde{y} \end{array}$$

- Major approaches: Patrini et al. (CVPR2017)
  - Loss correction by $\boldsymbol{T}^{-1}$ to eliminate noise.
  - Classifier correction 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)
  - Learn $\boldsymbol{T}$ and a classifier simultaneously.
    Xia, Liu, Wang, Han, Gong, Niu & Sugiyama (NeurIPS2019)
  - Decompose $\boldsymbol{T}$ into simpler components.
    Yao, Liu, Han, Gong, Deng, Niu, Sugiyama & Tao (arXiv2020)
  - Extension to input-dependent noise $\boldsymbol{T}(\boldsymbol{x})$.
    Xia, Liu, Han, Wang, Gong, Liu, Niu, Tao & Sugiyama (arXiv2020)

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.

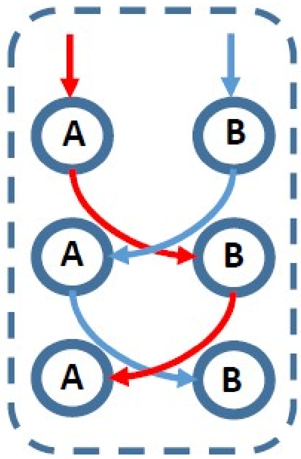
Han, Yao, Yu, Niu, Xu, Hu, Tsang & Sugiyama (NeurIPS2018)

- 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)



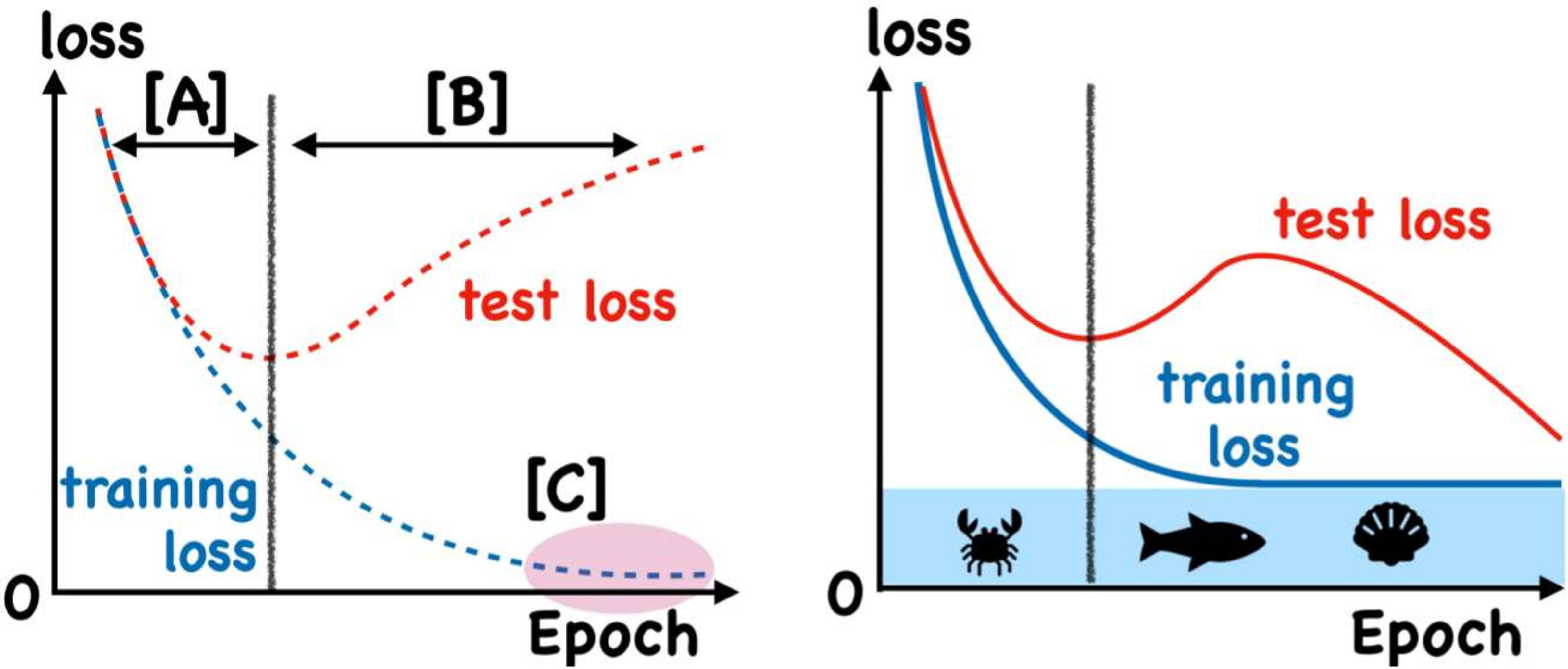
No theory but very robust in experiments:

- Works well even if 50% labels are randomly flipped.

Flooding

- Neural nets tend to overfit.
- "Flooding" the training error prevents overfitting.

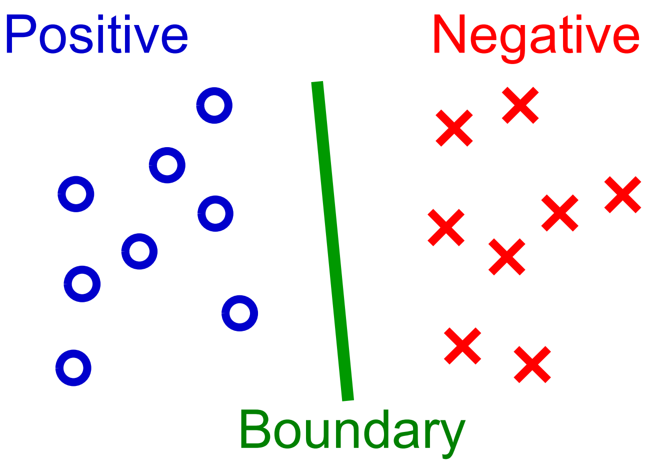
Ishida, Yamane, Sakai, Niu & Sugiyama (ICML2020)



Contents

Weakly Supervised Learning

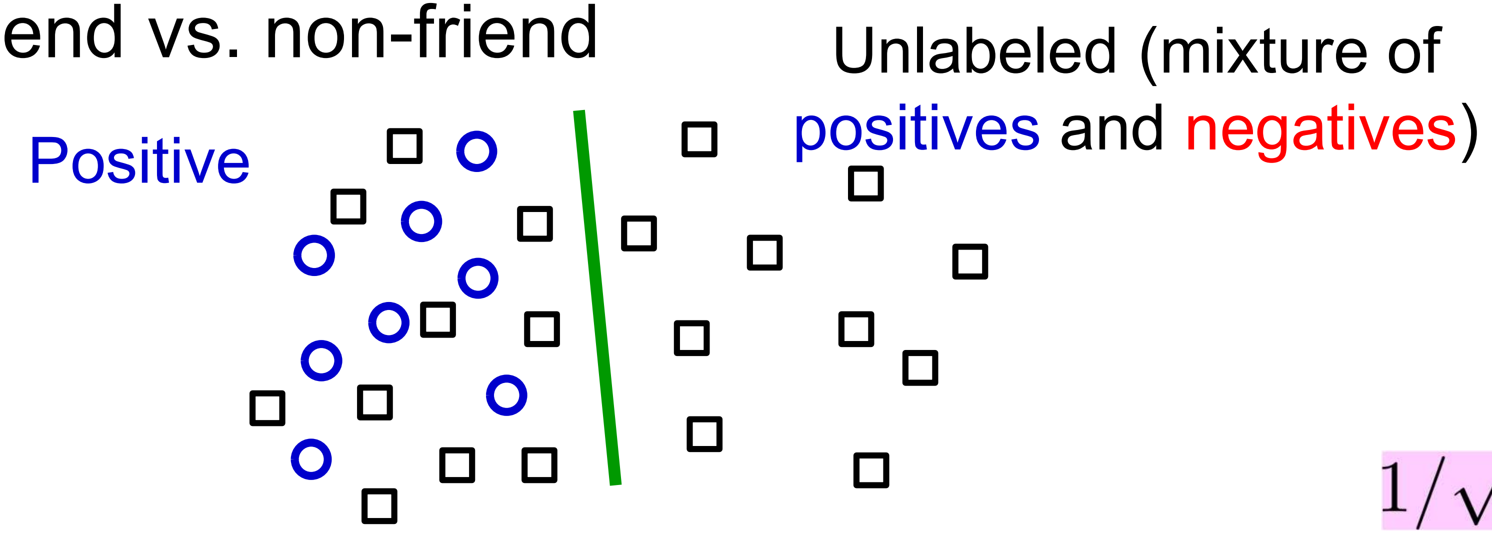
- Ordinary supervised learning requires fully labeled data (input-output pairs).
- But collecting fully labeled data can be expensive in practice.
- Can we utilize "weakly" labeled data?
  - No negative data
  - Positive confidence data
  - Similar/dissimilar data
  - Complementary data
  - Partial-label data



P: Positive, N: Negative, U: Unlabeled

PU Classification

- Only positive and unlabeled data is available; negative data is completely missing:
  - Click vs. non-click
  - Friend vs. non-friend



$1/\sqrt{n}$

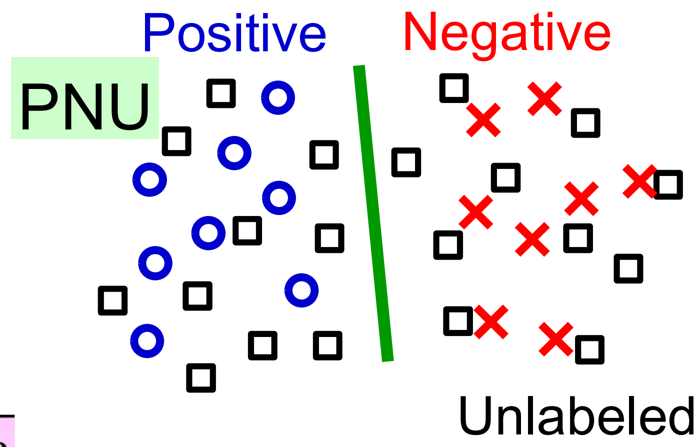
- From PU data, PN classifiers are trainable!

du Plessis, Niu & Sugiyama (NIPS2014, ICML2015)
Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016)
Kiryo, du Plessis, Niu & Sugiyama (NIPS2017)
Hsieh, Niu & Sugiyama (ICML2019)

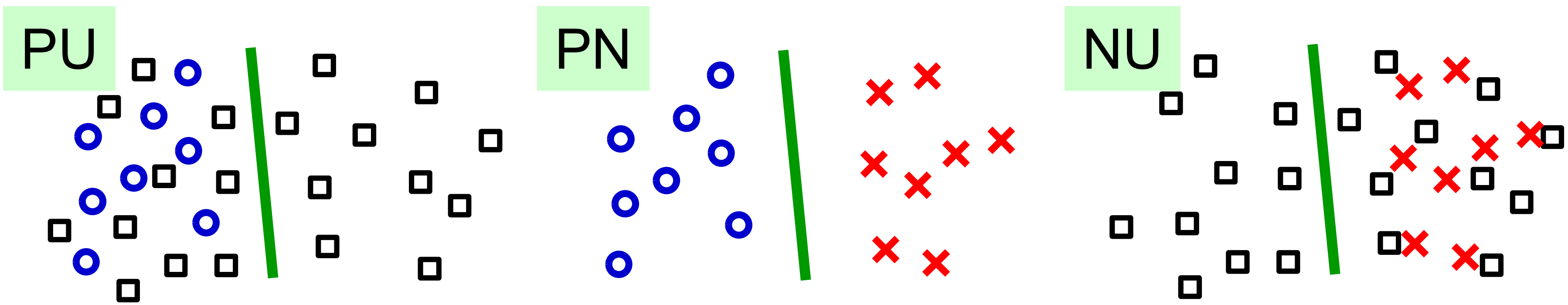
PNU Classification (Semi-Supervised Classification)

Sakai, du Plessis, Niu & Sugiyama (ICML2017)
Sakai, Niu & Sugiyama (MLJ2018)

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



$1/\sqrt{n}$

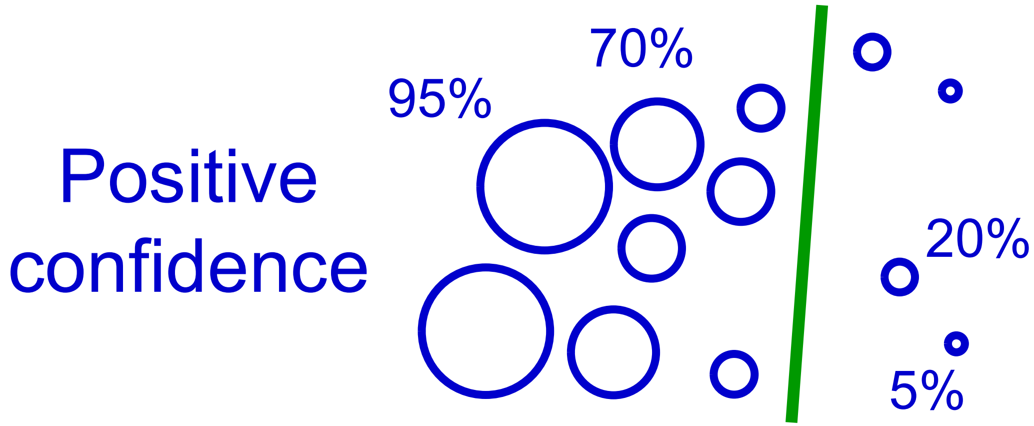


Pconf Classification

Ishida, Niu & Sugiyama (NeurIPS2018)

- Only P data is available, not U data:
  - Data from rival companies cannot be obtained.
  - Only positive results are reported (publication bias).
- "Only-P learning" is unsupervised.
- From Positive-confidence data, PN classifiers are trainable!

$1/\sqrt{n}$

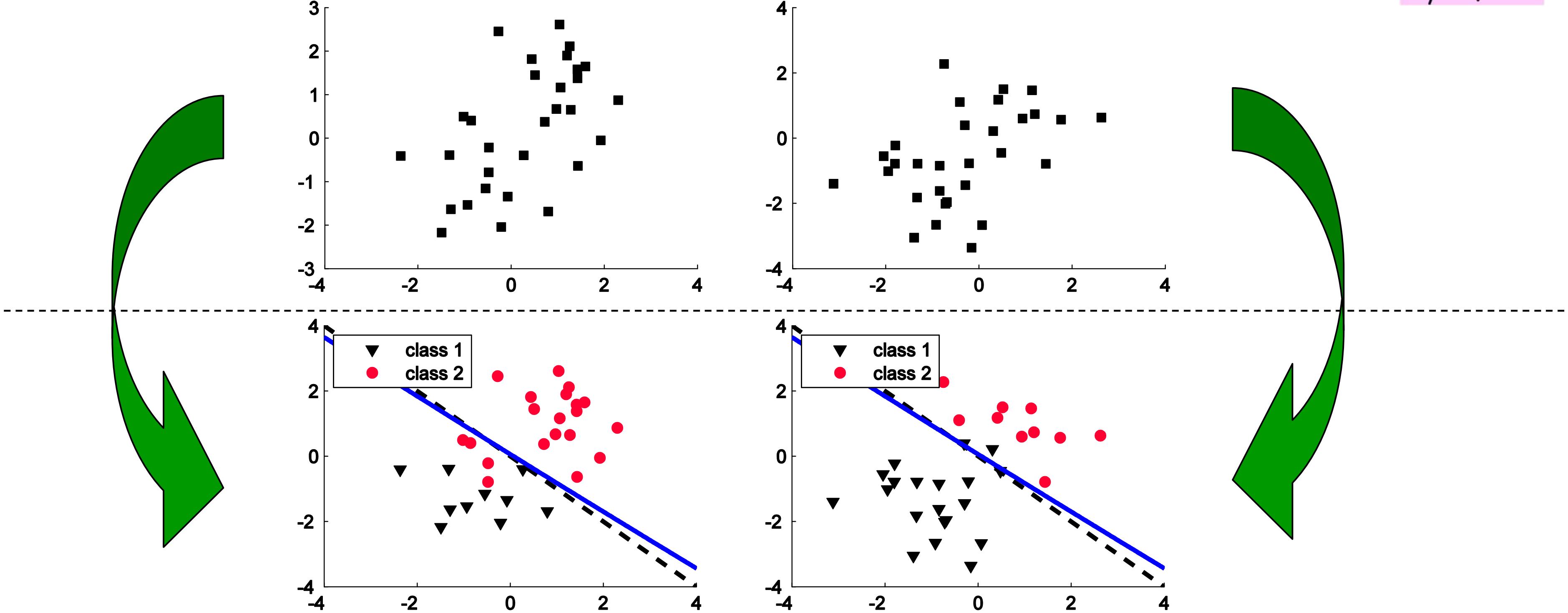


UU Classification

du Plessis, Niu & Sugiyama (TAAI2013)
Lu, Niu, Menon & Sugiyama (ICLR2019)
Charoenphakdee, Lee & Sugiyama (ICML2019)
Lu, Zhang, Niu & Sugiyama (AISTATS2020)

- From two sets of unlabeled data with different class priors, PN classifiers are trainable!

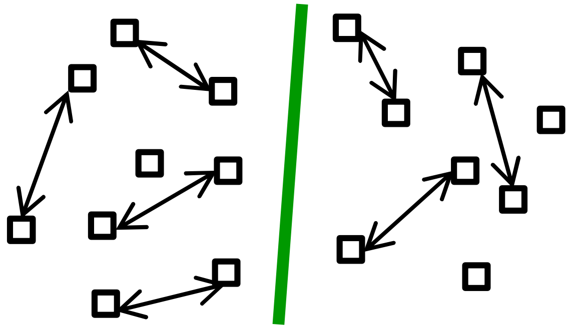
$1/\sqrt{n}$



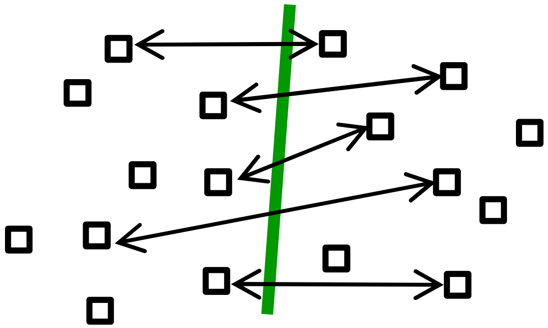
SU Classification

Bao, Niu & Sugiyama (ICML2018)

- Delicate classification (money, religion…):
  - Highly hesitant to directly answer questions.
  - Less reluctant to just say “same as him/her”.
- From similar data pairs and unlabeled data, PN classifiers are trainable!



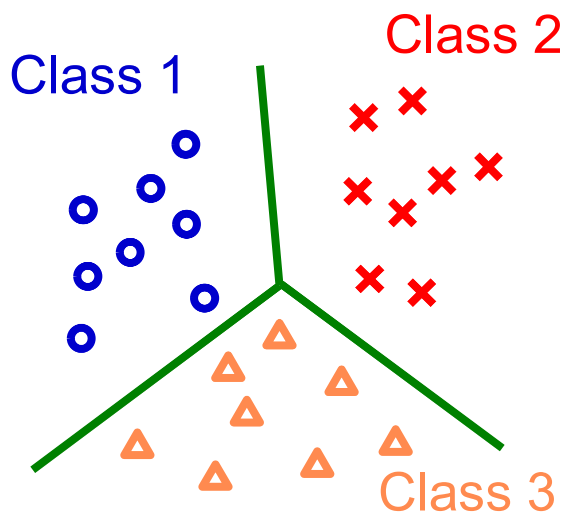
- Learning from dissimilar data pairs is also possible. $1/\sqrt{n}$
  - SDU classification is also possible.



Shimada, Bao, Sato & Sugiyama (arXiv2019)
Dan, Bao & Sugiyama (arXiv2020)

Complementary Classification

- **Complementary label**:
  a class the pattern does not belong to.
  - E.g., “not class 1”.
  - Cheaper than ordinary labels.



- Classifiers can be trained only from complementary labels.
  - Unbiased risk estimation

    Ishida, Niu & Sugiyama (NIPS2017)
    Ishida, Niu, Menon & Sugiyama (ICML2019)

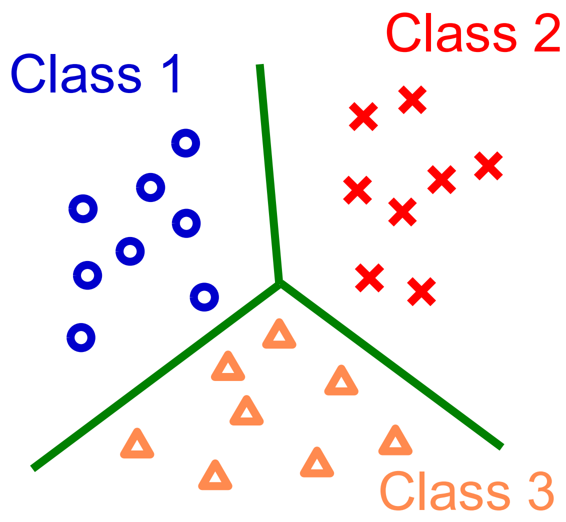
    $1/\sqrt{n}$
  - Multiple complementary labels

    Feng, Kaneko, Han, Niu, An & Sugiyama (ICML2020)
  - Beyond unbiased risk estimation

    Chou, Niu, Lin & Sugiyama (ICML2020)

Partial-Label Classification

- Partial label: Nguyen and Caruana (KDD2008)
  a subset of labels containing the true one
  - "Either 1 or 2"
  - Cheaper than ordinary labels



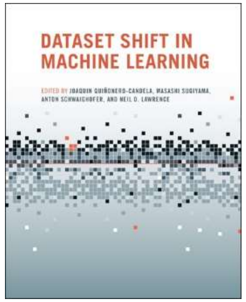
- Classifiers can be trained only from partial labels. $1/\sqrt{n}$
  - Progressive identification of correct labels.
    Lv, Xu, Feng, Niu, Geng & Sugiyama (ICML2020)
  - Explicit modeling of partial label generation.
    Feng, Lv, Han, Xu, Niu, Geng, An & Sugiyama (submitted)

Contents

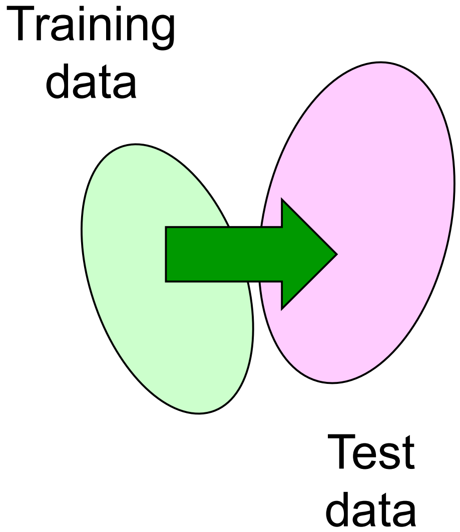
1. Trend in ML Research
2. Robust Machine Learning
   - A) Noisy label learning
   - B) Weakly supervised learning
   - C) Bias in training data
   - D) Noise in test input
3. Future ML Research

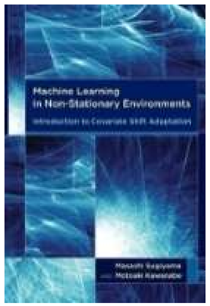
Bias in Training Data

Quiñonero-Candela, Sugiyama, Schwaighofer & Lawrence (MIT Press 2009)



- Training and test data often have different distributions, due to
  - changing environments,
  - sample selection bias.
- Transfer learning:
  - Match the distributions so that training data resemble test data.





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

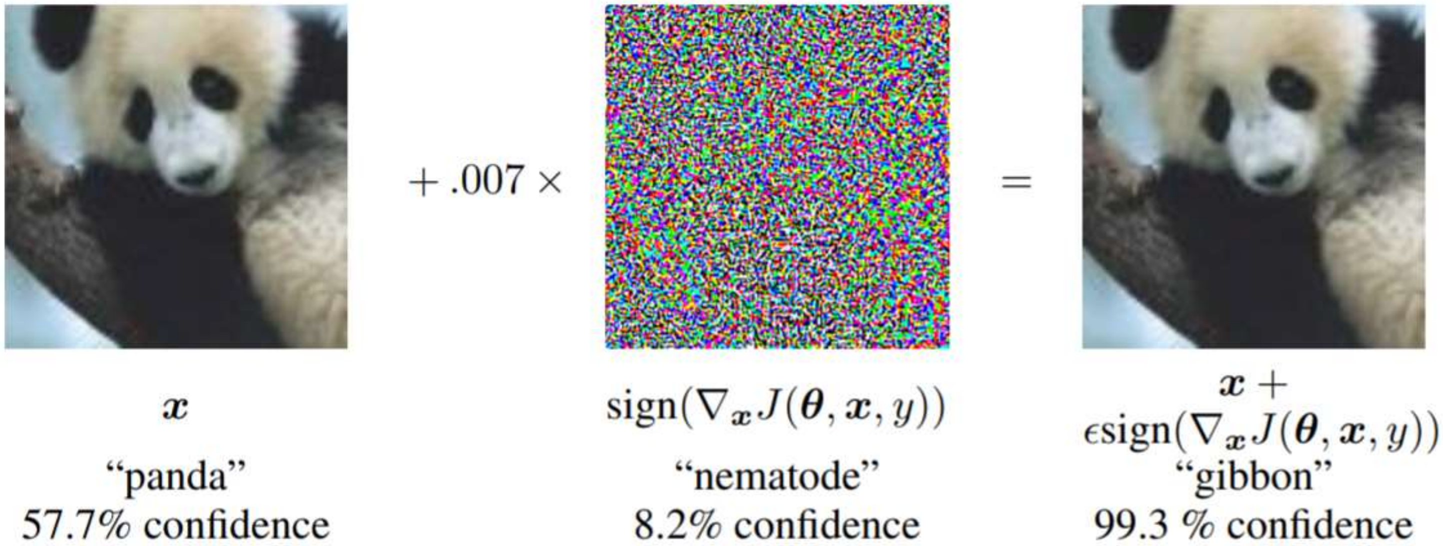
Unsupervised Transfer Learning

- Given training input-output and test input, match the training and test distributions:
  - Better discrepancy measures for matching:
    Kuroki, Charoenphakdee, Bao, Honda, Sato & Sugiyama (AAAI2019)
    Lee, Charoenphakdee, Kuroki & Sugiyama (arXiv2019)
  - Handling noisy labels in the source domain:
    Liu, Lu, Han, Niu, Zhang & Sugiyama (arXiv2019)
  - Transferring data generation mechanism:
    Teshima, Sato & Sugiyama (ICML2020)
  - Simultaneous learning of a classifier and importance weights:
    Zhang, Yamane, Lu & Sugiyama (submitted)
    Fang, Lu, Niu & Sugiyama (arXiv2020)

Contents

Noise in Test Input

- Neural nets are vulnerable to small perturbations in test input. Goodfellow et al. (ICLR2015)



- We want to be robust to such perturbations:
  - Defense to pointwise adversarial attack.
  - Robust to adversarial distribution shift.
  - Rejection of adversarial data.

Defense to Pointwise Attack

- Stabilize output of the neural net:

$$\forall \epsilon, \left(\|\epsilon\|_2 < c \; \Rightarrow \; t_X = \underset{i}{\operatorname{argmax}} \{ F\,(X + \epsilon)_i \} \right)$$

- Lipchitz-margin training:

Tsuzuku, Sato & Sugiyama (NeurIPS2018)

  - Compute the Lipchitz constant for each layer and for the entire network:

$$\|F(X) - F(X + \epsilon)\|_2 \leq L_F \|\epsilon\|_2$$

  - Train the neural net to have large prediction margins:

$$\forall i \neq t_X, (F_{t_X} \geq F_i + \sqrt{2} c L_F)$$

  - Robustness is theoretically guaranteed.

Distributionally Robust Learning

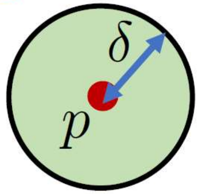
- Consider the worst-case test distribution when only training input-output is given:

$$\boxed{\min_{\theta} \sup_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(g_\theta(x), y)]}$$

$$\mathcal{Q}_p = \{q \mid \mathrm{D}_f(q \| p) \leq \delta\}$$

"f-divergence ball"

[Bagnell 2005, Ben-Tal+ 2013, Namkoong+ 2016, 2017]



- However, a naïve minimax approach does not work well:
  - Proved to be non-robust for classification.
    Hu, Niu, Sato & Sugiyama (ICML2018)
  - Elucidated the condition for loss calibration.
    Bao, Scott & Sugiyama (COLT2020)
  - New formulation for being not too conservative.
    Zhang, Xu, Han, Niu, Cui, Sugiyama & Kankanhalli (ICML2020)

Classification with Reject Option

Ni, Charoenphakdee, Honda & Sugiyama (NeurIPS2019)

- In severe applications, better to reject difficult test inputs and ask human to predict instead.
- Approach 1: Reject low-confidence prediction
  - Existing methods have limitation in loss functions (e.g., logistic loss), resulting in weak performance.
  - New rejection criteria for general losses with theoretical convergence guarantee.
- Approach 2: Train the classifier and rejector
  - Existing methods only focus on binary problems.
  - This approach was proved not converge to the optimal solution generally in multi-class cases.

Contents

1. Trend in ML Research
2. Robust Machine Learning
   - A) Noisy label learning
   - B) Weakly supervised learning
   - C) Bias in training data
   - D) Noise in test input
3. Future ML Research

Summary of Robust ML

- Nowadays, ML systems are deployed in various societal problems, where reliability is extremely important:
  - Robustness to expectable situations:
    - Model the corruption process explicitly and correct the solution.
  - Robustness to unexpected situations:
    - Consider worst-case robustness,
    - Include human support.
  - Somewhere in the middle would be practically more important.

Summary of General AI

- Many companies are interested in AI:
  - IT, finance, manufacturing, material, IT, education, medicine, electricity,…
- AI-driven science is becoming norm:
  - Physics, astronomy, chemistry, material, medicine, biology, informatics, control,…
- Social impact of AI is a serious concern:
  - Privacy, fairness, explanability,…

Future of ML Research

- Current ML achieves human-level performance for elementary tasks such as image understanding, speech recognition, and language translation:
  - Many standard jobs may be replaced by AI.
  - However, highly creative jobs and low-level jobs will never be taken over by AI.
- There are still challenges in ML research:
  - ML from less data, further robustness, time-series analysis, automatic ML, sequential decision making, life-long learning,…

Past and Future of AI Research

Logical AI

- 1960's: Inference and search
- 1980's: Expert systems and knowledge bases

Neuro-inspired AI

- 1960's: Single-layer perceptrons
- 1990's: Multi-layer perceptrons

Statistical ML based AI

- 2000's: Frequentist statistics, convex optimization, Bayesian statistics
- 2010's: Deep learning

Future AI

Need young talents!