電子情報通信学会IBISML研究会　　2020年10月20日
オーガナイズセッション「ロバスト機械学習」

不確実性下での機械学習手法：ロバスト機械学習法の紹介

杉山　将

理化学研究所／東京大学

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

ロバスト機械学習

- 機械学習の実応用では、様々な要因に対する耐性（ロバスト性）が重要：
  - 雑音：センサー誤差、ヒューマンエラー
  - 不完全情報：弱教師情報
  - バイアス：標本選択、環境変化
  - 攻撃：敵対的雑音、分布シフト
- 本講演では、（私が関わっている）ロバスト機械学習の最近の研究成果をご紹介します

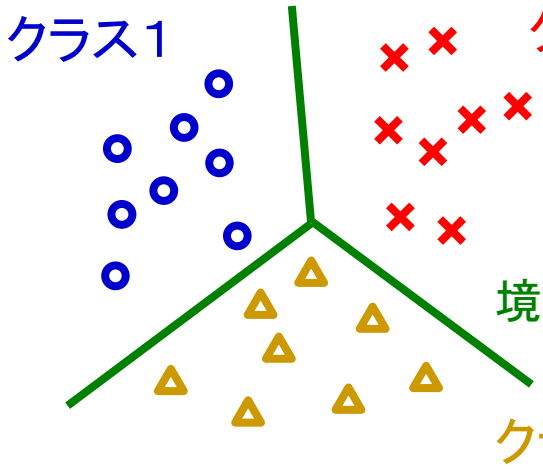
http://www.ms.k.u-tokyo.ac.jp/sugi/publications.html

発表の流れ

1. ラベル雑音下での分類
2. 弱教師付き学習
3. 転移学習
4. 敵対的攻撃

通常の分類

- ノイズのない訓練データ: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$



$\boldsymbol{x} \in \mathbb{R}^d$: 入力パターン

$y \in \{1, \ldots, c\}$: クラスラベル

- 訓練誤差最小化は統計的な一致性を持つ:

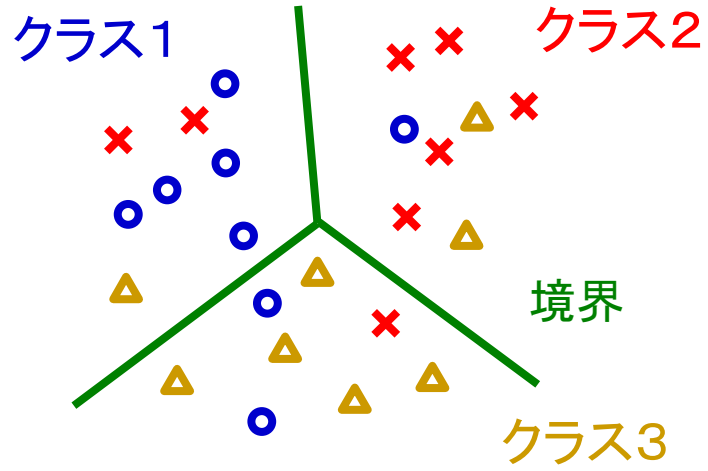
$$\frac{1}{n} \sum_{i=1}^{n} \ell\Big(y_i, \boldsymbol{g}(\boldsymbol{x}_i)\Big)$$

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

$\ell(y, \boldsymbol{g}(\boldsymbol{x})) \in \mathbb{R}$: 損失

ラベル雑音下での分類

- ラベル雑音を含む訓練データ: $\{(\boldsymbol{x}_i, \widetilde{y}_i)\}_{i=1}^n$



$\boldsymbol{x} \in \mathbb{R}^d$: 入力パターン

$\widetilde{y} \in \{1, \ldots, c\}$: 雑音を含むクラスラベル

- 訓練誤差最小化は一致性を持たず、実際にもうまくいかない事が多い:

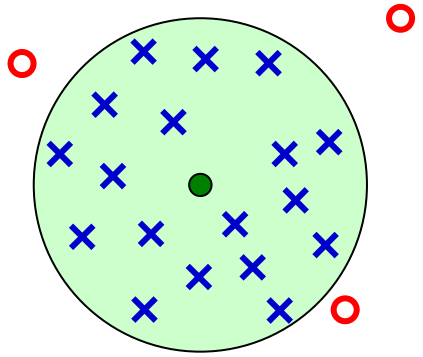
$$\frac{1}{n}\sum_{i=1}^{n} \ell\left(\widetilde{y}_i, \boldsymbol{g}(\boldsymbol{x}_i)\right)$$

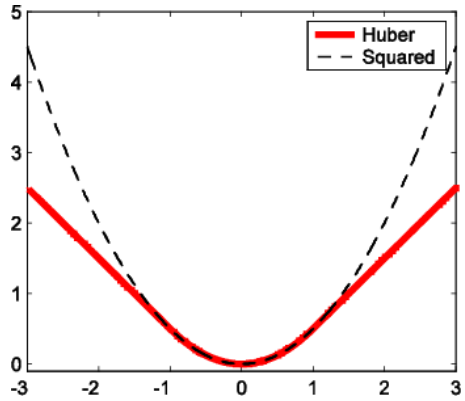
$\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^c$: 分類器

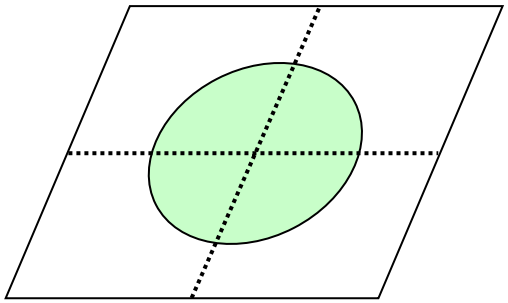
$\ell(y, \boldsymbol{g}(\boldsymbol{x})) \in \mathbb{R}$: 損失

ラベル雑音に対する従来のアプローチ

- 教師なし外れ値除去:
  - そもそも分類よりも難しい
- ロバスト損失、正則化:
  - 実用的だが、ロバスト性はそれほど高くない
- 新しいアプローチが必要！
  - 雑音遷移補正
  - 低雑音標本選択
  - モデル複雑さ制御

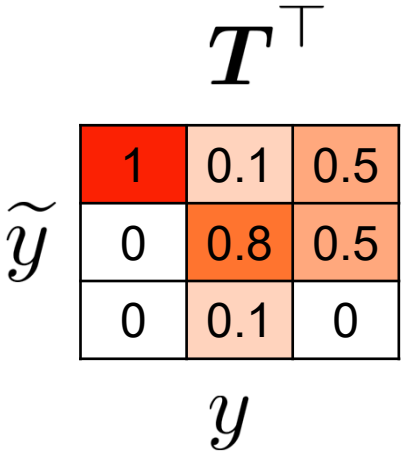






(1-A)雑音遷移補正

- 雑音遷移行列 T : y が $\widetilde{y}$ に変わる確率
- 補正法： Patrini et al. (CVPR2017)
  - 損失補正 : T^{-1} で損失の雑音を除去
  - 分類器補正 : $T^{\top}$ で分類器に雑音を付加



- 雑音ありデータだけから T が推定できないか？
  - ヒトの認知バイアスを取り込む :
    Han, Yao, Niu, Zhou, Tsang, Zhang & Sugiyama (NeurIPS2018)
  - 雑音遷移行列と分類器を同時学習 :
    Xia, Liu, Wang, Han, Gong, Niu & Sugiyama (NeurIPS2019)
  - 雑音遷移行列を分解して精度良く推定 :
    Yao, Liu, Han, Gong, Deng, Niu, Sugiyama & Tao (NeurIPS2020)
  - 入力依存の雑音遷移関数 $T(x)$ への拡張 :
    Xia, Liu, Han, Wang, Gong, Liu, Niu, Tao & Sugiyama (NeurIPS2020)

(1-B) 共教示

■ ニューラルネットの記憶能力：

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

- 確率的降下学習は雑音なしデータに早く記憶
- しかし，単純な早期終了ではうまくいかない

■ 2つのニューラルネットを用いた共教示：

- 誤差の小さいデータを選んで教え合う

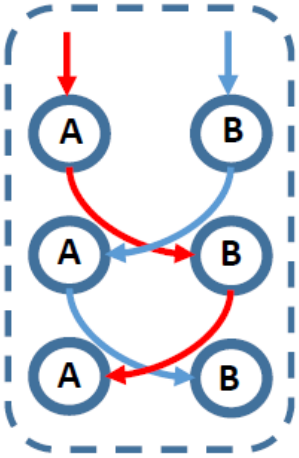
Han, Yao, Yu, Niu, Xu, Hu, Tsang & Sugiyama (NeurIPS2018)

- 出力が合致しないデータだけを教える

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

- 誤差の大きいデータに対して勾配上昇

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



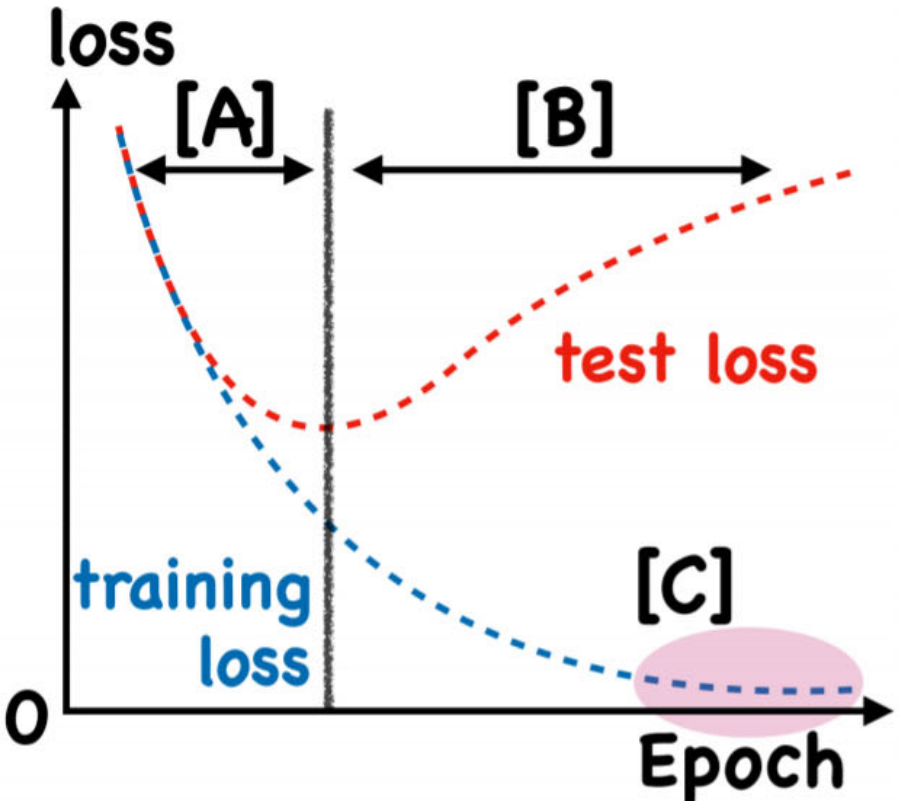
■ 理論はないが，実験的には超ロバスト：

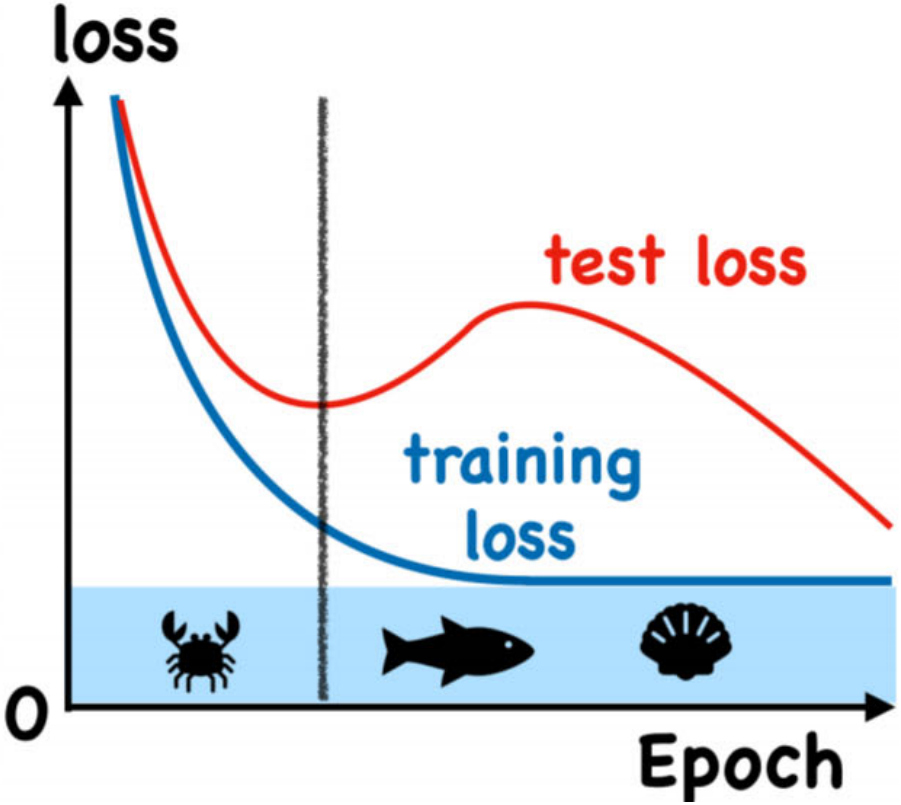
- 50%のラベルをランダムに変えてもうまく学習できる！

(1-C) 洪水法

- ニューラルネットは雑音に過適合しやすい？
- 訓練誤差を「洪水」させて過適合を抑制：
  - 二重降下が誘発される？

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



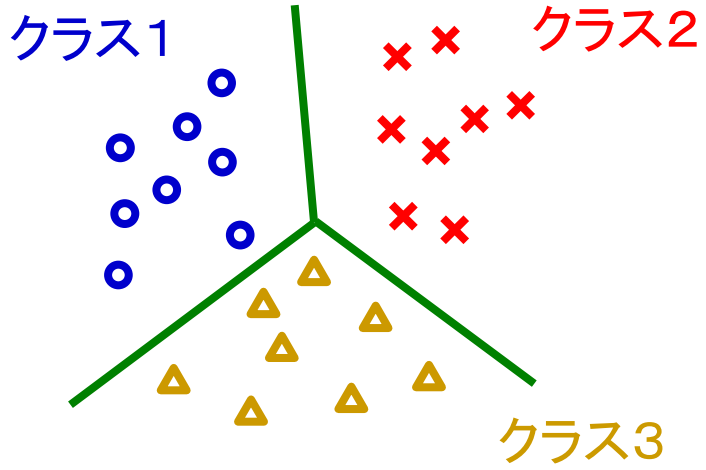


発表の流れ

1. ラベル雑音下での分類
2. 弱教師付き学習
3. 転移学習
4. 敵対的攻撃

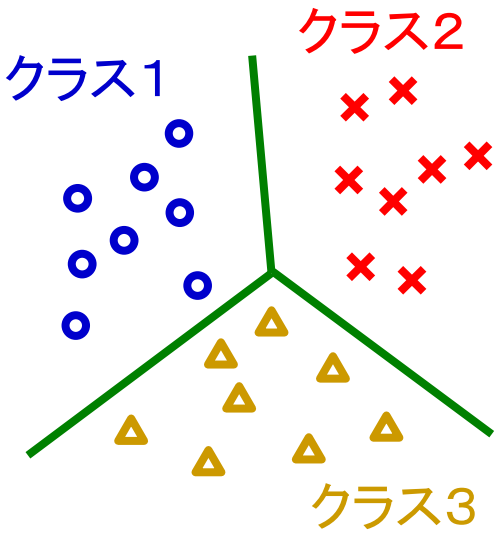
弱教師付き学習

- 通常の教師付き学習にはラベル付き訓練データが必要:
  - しかし，データのラベル付けにはコストがかかる
- 低コストで集められる弱教師付きデータを使えないか？
  - 補ラベル分類
  - 部分ラベル分類
  - 二値分類に対する様々な弱教師付き分類



(2-A) 補ラベル分類

- 補ラベル：パターンが属さないクラスを示すラベル
  - 例：「クラス1に属さない」
  - 通常のラベルより簡単に集められる
    （クラスを適当に選べば大抵は間違い）



- 補ラベル付き訓練データだけから
  でも分類器の学習が可能！ $1/\sqrt{n}$
  - 不偏リスク推定の枠組み

    Ishida, Niu & Sugiyama (NIPS2017)
    Ishida, Niu, Menon & Sugiyama (ICML2019)
  - 複数補ラベル分類への拡張

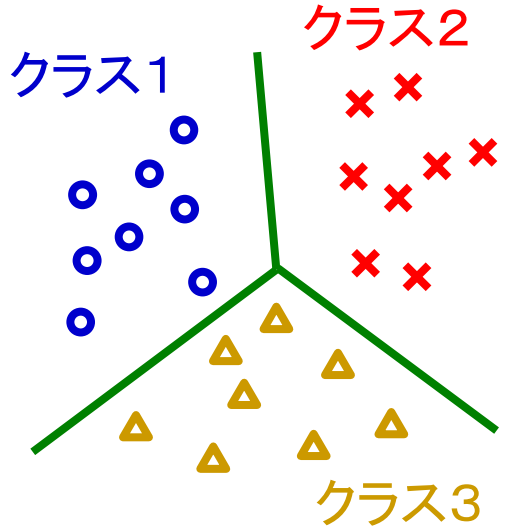
    Feng, Kaneko, Han, Niu, An & Sugiyama (ICML2020)
  - 不偏リスク推定を超えた新しい定式化

    Chou, Niu, Lin & Sugiyama (ICML2020)

(2-B) 部分ラベル分類

Nguyen and Caruana (KDD2008)

- **部分ラベル**：真のクラスを含むラベルのサブセット
  - 例：「クラス1か2に属する」
  - 通常のラベルより簡単に集められる
    （正しいラベルを絞り込まなくてもよい）



- **部分ラベル付き訓練データだけからでも分類器の学習が可能！** $1/\sqrt{n}$
  - 分類器と正しいラベルを反復推定

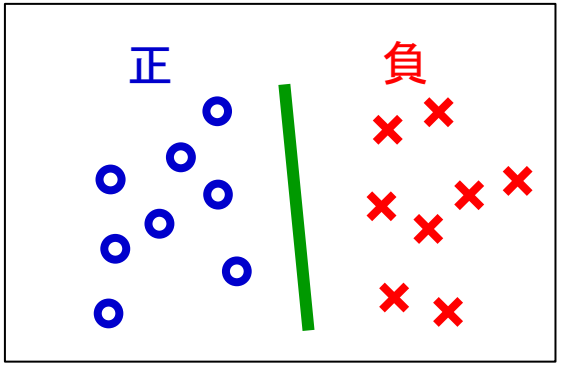
Lv, Xu, Feng, Niu, Geng & Sugiyama (ICML2020)

  - 部分ラベルの生成プロセスの精緻化

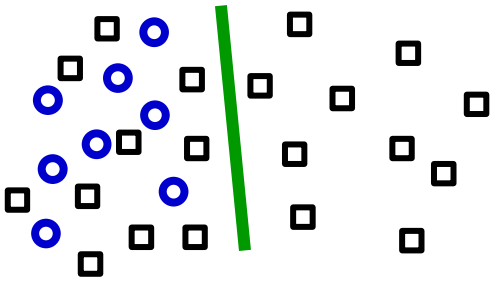
Feng, Lv, Han, Xu, Niu, Geng, An & Sugiyama (NeurIPS2020)

(2-C) 二値分類に対する研究

■ 様々な弱教師付きデータから二値分類が可能！ $1/\sqrt{n}$

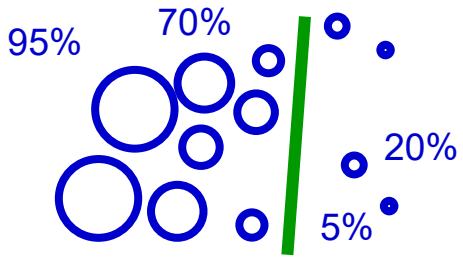


正ラベルなし分類



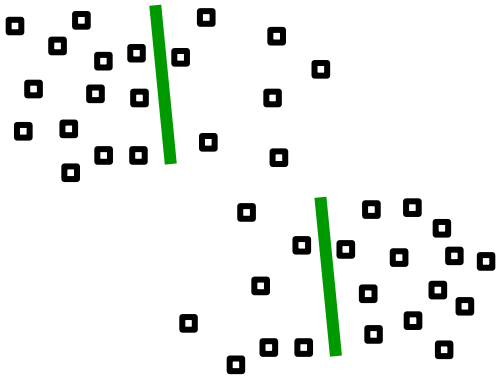
du Plessis, Niu & Sugiyama (NIPS2014, ICML2015)
Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016)
Kiryo, du Plessis, Niu & Sugiyama (NIPS2017)
Hsieh, Niu & Sugiyama (ICML2019)

正信頼度学習



Ishida, Niu & Sugiyama (NeurIPS2018)

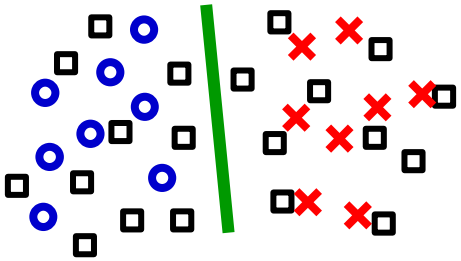
ラベルなしラベルなし分類



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

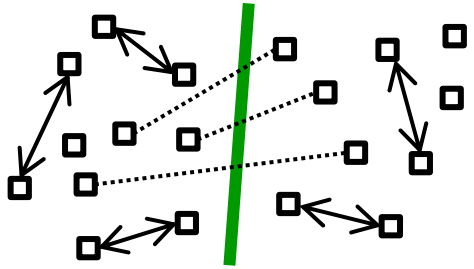
Sugiyama, Sakai, Ishida, Nan, Bao & Niu,
Machine Learning from Weak Supervision,
MIT Press, 2021?

正負ラベルなし分類



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

類似非類似ラベルなし分類



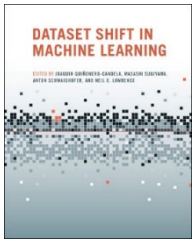
Bao, Niu & Sugiyama (ICML2018)
Shimada, Bao, Sato & Sugiyama (arXiv2019)
Dan, Bao & Sugiyama (arXiv2020)

発表の流れ

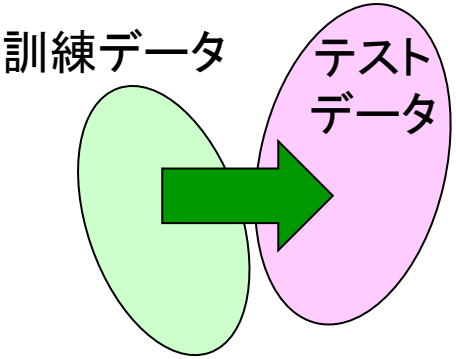
1. ラベル雑音下での分類
2. 弱教師付き学習
3. 転移学習
4. 敵対的攻撃

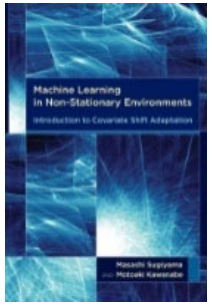
訓練データのバイアス

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



- 訓練データとテストデータの分布が異なると，標準的な機械学習法はうまくいかない：
  - 環境の変化
  - 標本選択バイアス
- 転移学習，ドメイン適応：
  - 訓練データの分布をテストデータに合わせる





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

教師なし転移学習

- ラベル付き訓練データと、ラベルなしテストデータからの転移学習：
  - 分布適合のための距離尺度について
    Kuroki, Charoenphakdee, Bao, Honda, Sato & Sugiyama (AAAI2019)
    Lee, Charoenphakdee, Kuroki & Sugiyama (arXiv2019)
  - 訓練データのラベル雑音への対応
    Liu, Lu, Han, Niu, Zhang & Sugiyama (arXiv2019)
  - データ生成メカニズムの転移
    Teshima, Sato & Sugiyama (ICML2020)
  - 転移に用いる重要度重みと分類器を同時学習
    Zhang, Yamane, Lu & Sugiyama (ACML2020)
    Fang, Lu, Niu & Sugiyama (NeurIPS2020)
  - ラベルなしテストデータがない・不完全な場合への対応
    Ishii, Takenouchi & Sugiyama (ACML2019, WACV2020)

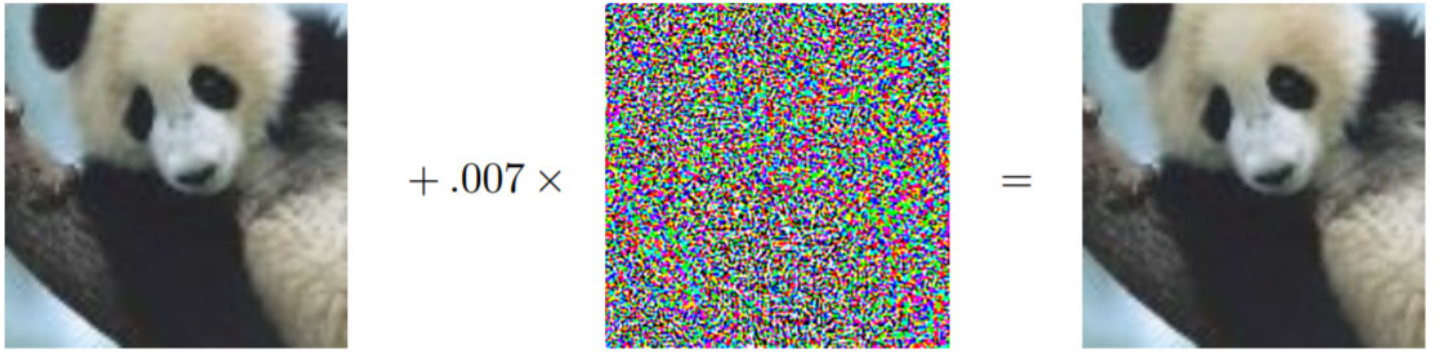
発表の流れ

1. ラベル雑音下での分類
2. 弱教師付き学習
3. 転移学習
4. 敵対的攻撃

テスト入力への雑音

- ニューラルネットはテスト入力の微小な摂動に弱い

Goodfellow et al. (ICLR2015)



- そのような摂動に対するロバスト性を改善したい:
  - 敵対的入力摂動への対応
  - 敵対的分布シフトへの対応
  - 敵対的入力の棄却

敵対的入力摂動への対応

- ニューラルネットの出力を安定化：

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

- リプシッツ・マージン学習：

Tsuzuku, Sato & Sugiyama (NeurIPS2018)

  - ニューラルネットの各層のリプシッツ定数を動的に計算

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

  - 予測マージンが大きくなるようにニューラルネットを学習

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

- 敵対的入力摂動に対するロバスト性を理論保証：
  - ただし，予測精度とのトレードオフがある

敵対的分布シフトへの対応

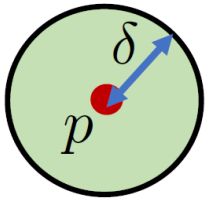
- 最悪のテスト分布を想定して学習：

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

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

“f-divergence ball”

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



- しかし，minimax定式化は必ずしもうまくいかない：
  - 分類ではロバストにならない

    Hu, Niu, Sato & Sugiyama (ICML2018)
  - 損失関数の選択について

    Bao, Scott & Sugiyama (COLT2020)
  - 保守的すぎない新しい定式化

    Zhang, Xu, Han, Niu, Cui, Sugiyama & Kankanhalli (ICML2020)

棄却付き分類

Ni, Charoenphakdee, Honda & Sugiyama (NeurIPS2019)

- 予測に自信がないとき，自動分類をあきらめて，人間のエキスパートに分類してもらう:
  - 医療診断など，誤分類にリスクがある問題で有効
- アプローチ1：予測信頼度が低い場合に棄却
  - 従来はロジスティック損失を使う必要があり，分類性能があまり良くない
  - 棄却基準を一般化し，より広いクラスの損失に対して理論保証をもって棄却できる
- アプローチ2：分類器と棄却器を同時学習
  - 二値分類に対して優れている Cortes (ALT2016, NeurIPS2016)
  - しかし，多値分類では実用的でないことを証明

発表の流れ

1. ラベル雑音下での分類
2. 弱教師付き学習
3. 転移学習
4. 敵対的攻撃

まとめ

- 機械学習システムの信頼性向上は不可欠
  - 想定できる悪状況に対するロバスト性
    - 悪状況をモデル化し悪影響を補正：
      悪状況のモデル化誤差をどう抑えるかが課題
  - 想定できない悪状況に対するロバスト性：
    - 最悪ケースを考えて補正：
      保守的になりすぎないようにすることが課題
    - 人間のサポートに頼る：
      自動運転など実時間応用では使えない？
  - 実用的にはその中間くらいが重要そう？
- ヒトとの相性が良くなればAIの信頼性は増す？
  - 脳の学習機構，認知バイアスなどを取り込む
  - 社会常識，文化などをAIに反映