Robust Learning from Multiple Sources

Christoph H. Lampert
joint work with Nikola Konstantinov, Dan Alistarh, Elias Frantar, Eugenia Iofinova

Institute of Science and Technology Austria

The Mathematics of Machine Learning Workshop
Bilbao, Oct 27, 2022

Slides available at: `http://cvml.ist.ac.at`

SCAN ME

Institute of Science and Technology Austria (ISTA)

- public research institute, opened in 2009
- located in outskirts of Vienna

Focus on curiosity-driven basic research

- avoiding boundaries between disciplines
- current 70 research groups
  - Computer Science, Mathematics, Physics, Chemistry, Biology, Neuroscience, Earth and Climate Sciences
- ELLIS unit since 2019

We're hiring! (on all levels)

- interns, PhD students, postdocs
- faculty (tenure-track or tenured), ...

More information: `chl@ist.ac.at` or `https://cvml.ist.ac.at`

Topics in Our Research Group

Machine Learning Theory

- Transfer Learning
- Multi-task Learning
- Lifelong/Meta-Learning
- Multi-source/Federated Learning

Models/Algorithms

- Zero-shot Learning
- Continual Learning
- Weakly-supervised Learning
- Trustworthy/Robust Learning

Learning for Computer Vision

- Scene Understanding
- Generative Models
- Abstract Reasoning
- Zero-Shot Learning

Topics in Our Research Group

Machine Learning Theory

- Transfer Learning
- Multi-task Learning
- Lifelong/Meta-Learning
- Multi-source/Federated Learning

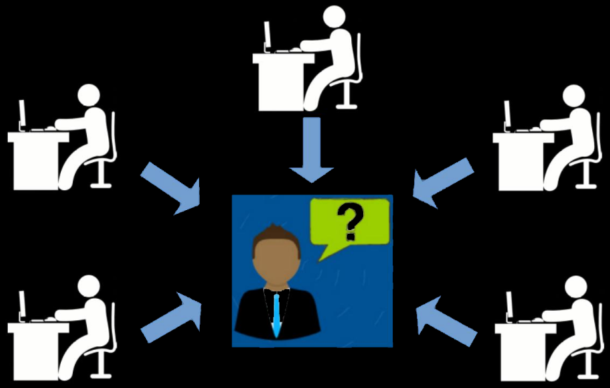
Models/Algorithms

- Zero-shot Learning
- Continual Learning
- Weakly-supervised Learning
- Trustworthy/Robust Learning

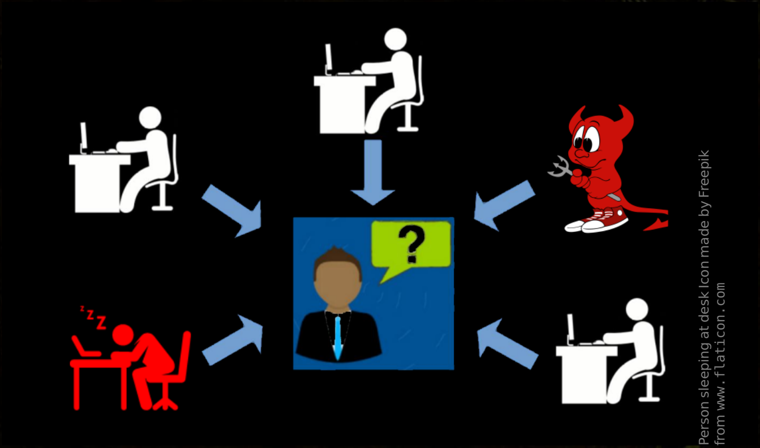
Learning for Computer Vision

- Scene Understanding
- Generative Models
- Abstract Reasoning
- Zero-Shot Learning

Training data from multiple sources



Training data from multiple sources



How much can be learned even if some data is corrupted or manipulated?

Schedule

Overview

Refresher: Statistical Learning Theory

Robust Learning From Untrusted Sources

Robust Fair Learning

Slides available at: `http://cvml.ist.ac.at`

Reminder: Supervised Learning

Setting:

- Inputs: $x \in \mathcal{X}$, e.g. strings, images, vectors, ...
- Outputs: $y \in \mathcal{Y}$. For simplicity: $\mathcal{Y} = \{\pm 1\}$ (binary classification)
- Probability distribution: $p(x, y)$ over $\mathcal{X} \times \mathcal{Y}$, unknown to the learner
- Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. For simplicity: 0/1-loss $\ell(y, \bar{y}) = \mathbb{1}\{y \neq \bar{y}\}$

Abstract Goal:

- find a predictor, $f : \mathcal{X} \to \mathcal{Y}$, such that the expected loss

$$\mathrm{er}(h) = \mathbb{E}_{(x,y)\sim p}[\ell(y, f(x))] = \mathrm{Pr}_{(x,y)\sim p}\{f(x) \neq y\}$$

on *future data* is small.

Learning from data:

- training data: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p$
- hypothesis class: $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$
- learning algorithm $\mathcal{L} : \mathbb{P}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}$, $\mathbb{P}(\cdot) =$ power set
  - input: a training set, $S \subset \mathcal{X} \times \mathcal{Y}$,
  - output: a trained model $\mathcal{L}(S) \in \mathcal{H}$ (= prediction function).

Learning from data:

- training data: $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p$
- hypothesis class: $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$
- learning algorithm $\mathcal{L} : \mathbb{P}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}$, $\mathbb{P}(\cdot) =$ power set
  - input: a training set, $S \subset \mathcal{X} \times \mathcal{Y}$,
  - output: a trained model $\mathcal{L}(S) \in \mathcal{H}$ (= prediction function).

Central question in Statistical Learning Theory:

Is there a universal learning algorithm, such that: $\mathrm{er}(\mathcal{L}(S)) \overset{|S|\to\infty}{\to} \min_{h \in \mathcal{H}} \mathrm{er}(h)$?

Learning from data:

- training data: $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p$
- hypothesis class: $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$
- learning algorithm $\mathcal{L} : \mathbb{P}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}$, $\mathbb{P}(\cdot) =$ power set
  - input: a training set, $S \subset \mathcal{X} \times \mathcal{Y}$,
  - output: a trained model $\mathcal{L}(S) \in \mathcal{H}$ (= prediction function).

Central question in Statistical Learning Theory:

Is there a universal learning algorithm, such that: $\text{er}(\mathcal{L}(S)) \overset{|S|\to\infty}{\to} \min_{h\in\mathcal{H}} \text{er}(h)$?

Classic result: [Vapnik&Chervonenkis, 1971], [Blumer, Ehrenfeucht, Hassler, Warmuth, 1989]

If and only if **VC**$(\mathcal{H}) < \infty$, empirical risk minimization (ERM) does the job:

$$\mathcal{L}(S) \leftarrow \operatorname*{argmin}_{h\in\mathcal{H}} \text{er}_S(h) \qquad \text{for } \text{er}_S(h) := \frac{1}{|S|} \sum_{(x,y)\in S} \mathbb{1}\{f(x) \neq y\}.$$

[V. N. Vapnik, A. Ya. Chervonenkis. "Theory of uniform convergence of frequencies of appearance of attributes to their probabilities and problems of defining optimal solution by empiric data". Theory of Probability and its Applications, 1971]
[A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth. "Learnability and the Vapnik-Chervonenkis Dimension". Journal of the ACM, 1989]

Learning from unreliable/malicious data:

- training set: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$
- but: data has issues: some data points might not really be samples from p

Learning from unreliable/malicious data:

- training set: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$
- but: data has issues: some data points might not really be samples from p
- formally: malicious adversary $\mathfrak{A}$ [Valiant 1985]
  - $\mathfrak{A}$ can manipulate a fraction α of the dataset
  - input: dataset S
  - output: dataset $S' = \mathfrak{A}(S)$
    - $\lceil (1-\alpha)m \rceil$ points are unchanged,
    - $\lfloor \alpha m \rfloor$ are arbitrary
  - $\mathfrak{A}$ can depend on the learning algorithms, true data distribution, etc.

Question: Is ERM still be a universally good learning strategy?

Learning from unreliable/malicious data:

- training set: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$
- but: data has issues: some data points might not really be samples from p
- formally: malicious adversary $\mathfrak{A}$ [Valiant 1985]
  - $\mathfrak{A}$ can manipulate a fraction α of the dataset
  - input: dataset S
  - output: dataset $S' = \mathfrak{A}(S)$
    - $\lceil (1-\alpha)m \rceil$ points are unchanged,
    - $\lfloor \alpha m \rfloor$ are arbitrary
  - $\mathfrak{A}$ can depend on the learning algorithms, true data distribution, etc.

Question: Is ERM still be a universally good learning strategy?

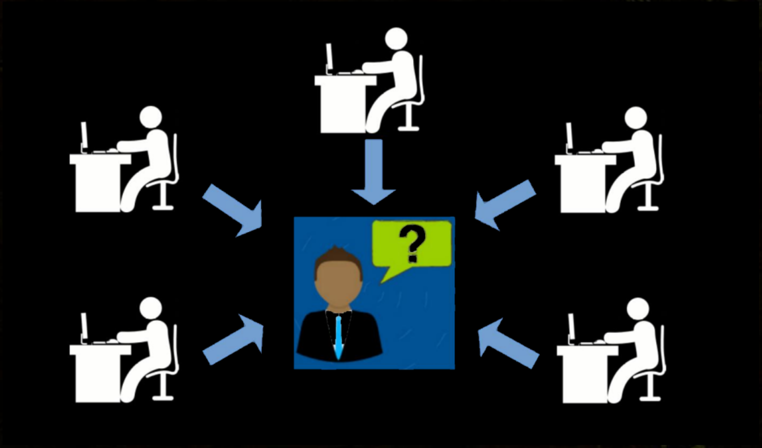
Classic Result: no! [Kerns&Li, 1993]

> No learning algorithm can guarantee an error less than $\frac{\alpha}{1-\alpha}$ on future data!

[L. G. Valiant. "Learning disjunctons of conjunctions". IJCAI 1985]
[M. Kearns, M. Li. "Learning in the presence of malicious errors". SIAM Journal on Computing, 1993]

Learning from Multiple Sources

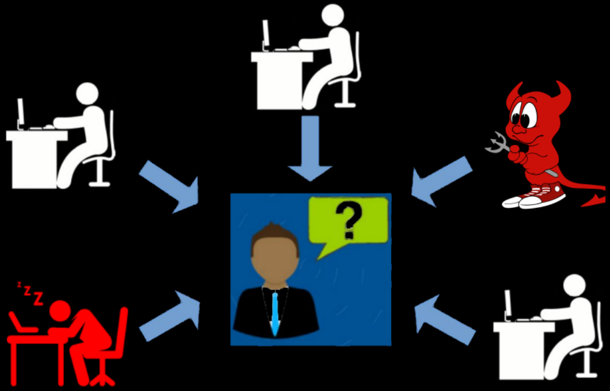
Training data from multiple sources



If all sources are i.i.d. samples from the correct data distribution

$\longrightarrow$ naive strategy "*merge all datasets and train a classifier*" works perfectly

Training data from multiple sources

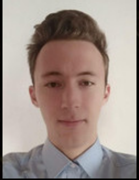


Person sleeping at desk Icon made by Freepik from www.flaticon.com

If some sources are not reliable, naive strategy can fail miserably!

Robust Learning from Unreliable or Malicious Sources

Nikola Konstantinov (ETH Zurich)



Elias Frantar (ISTA)

Dan Alistarh (ISTA)

Disclaimer: "These results have been modified from their original form. They have been edited to fit the screen and the allotted time slot."

[N. Konstantinov, E. Frantar, D. Alistarh, CHL. "*On the Sample Complexity of Adversarial Multi-Source PAC Learning*", ICML 2020]
[N. Konstantinov, CHL. "*Robust Learning from Untrusted Sources*", ICML 2019]

Learning from Multiple Sources

- multiple training sets: $S_1, S_2, \ldots, S_N$
  - each $S_i = \{(x^i_1, y^i_1), \ldots, (x^i_m, y^i_m)\} \overset{i.i.d.}{\sim} p$
- multi-source learning algorithm: $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^{N \times m} \to \mathcal{H}$
  - input: training sets, $S_1, S_2, \ldots, S_N$
  - output: one hypothesis $\mathcal{L}(S_1, \ldots, S_N) \in \mathcal{H}$ (= a trained model).

Learning from Multiple Unreliable/Malicious Sources

- multiple training sets: $S_1, S_2, \dots, S_N$
  - each $S_i = \{(x^i_1, y^i_1), \dots, (x^i_m, y^i_m)\} \overset{i.i.d.}{\sim} p$
- multi-source learning algorithm: $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^{N \times m} \to \mathcal{H}$
  - input: training sets, $S'_1, S'_2, \dots, S'_N = \mathfrak{A}(S_1, \dots, S_N)$
  - output: one hypothesis $\mathcal{L}(S'_1, S'_2, \dots, S'_N) \in \mathcal{H}$ (= a trained model).
- adversary $\mathfrak{A}$
  - input: data sets $S_1, \dots, S_N$
  - output: data sets $S'_1, \dots, S'_N$,
    - $\lceil (1-\alpha)N \rceil$ sources are identical to before,
    - $\lfloor \alpha N \rfloor$ sources are arbitrary.
  - the adversary might know the training algorithm, data distribution, ...

Learning from Multiple Unreliable/Malicious Sources

- multiple training sets: $S_1, S_2, \dots, S_N$
  - each $S_i = \{(x_1^i, y_1^i), \dots, (x_m^i, y_m^i)\} \overset{i.i.d.}{\sim} p$
- multi-source learning algorithm: $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^{N \times m} \to \mathcal{H}$
  - input: training sets, $S_1', S_2', \dots, S_N' = \mathfrak{A}(S_1, \dots, S_N)$
  - output: one hypothesis $\mathcal{L}(S_1', S_2', \dots, S_N') \in \mathcal{H}$ (= a trained model).
- adversary $\mathfrak{A}$
  - input: data sets $S_1, \dots, S_N$
  - output: data sets $S_1', \dots, S_N'$,
    - $\lceil (1-\alpha)N \rceil$ sources are identical to before,
    - $\lfloor \alpha N \rfloor$ sources are arbitrary.
  - the adversary might know the training algorithm, data distribution, ...

Is there a universal learning algorithm, i.e. $\text{er}(\mathcal{L}(S_1', \dots, S_N')) \overset{m\to\infty}{\to} \min_{h\in\mathcal{H}} \text{er}(h)$?

Related Work

Robust learning from a single dataset

- no universal algorithm: minimum guaranteeable error is $\frac{\alpha}{1-\alpha}$ [Kerns and Li, 1993]
- identical to our situation when each dataset consists of a single point, $m = 1$

  $\longrightarrow$ only $N \to \infty$ will probably not suffice to learn arbitrarily well

[M. Kearns and M. Li. "Learning in the presence of malicious errors." SIAM Journal on Computing, 1993],

Related Work

Robust learning from a single dataset

- no universal algorithm: minimum guaranteeable error is $\frac{\alpha}{1-\alpha}$ [Kerns and Li, 1993]
- identical to our situation when each dataset consists of a single point, $m = 1$
  $\longrightarrow$ only $N \to \infty$ will probably not suffice to learn arbitrarily well

Collaborative learning (multiple parties together learn *individual models*)

- universal learning algorithm exists [Blum *et al.*, 2017], [Qiao, 2018]

[M. Kearns and M. Li. "Learning in the presence of malicious errors." SIAM Journal on Computing, 1993], [A. Blum, N. Haghtalab, A. D. Procaccia, and M. Qiao. "Collaborative PAC learning". NeurIPS, 2017], [M. Qiao. "Do outliers ruin collaboration?" ICML, 2018],

Related Work

Robust learning from a single dataset

- no universal algorithm: minimum guaranteeable error is $\frac{\alpha}{1-\alpha}$ [Kerns and Li, 1993]
- identical to our situation when each dataset consists of a single point, $m = 1$
  $\longrightarrow$ only $N \to \infty$ will probably not suffice to learn arbitrarily well

Collaborative learning (multiple parties together learn *individual models*)

- universal learning algorithm exists [Blum *et al.*, 2017], [Qiao, 2018]

Density estimation from untrusted batches

- possible, but not applicable to supervised learning [Qiao and Valiant, 2018], [Jain and Orlitsky, 2020]

[M. Kearns and M. Li. "Learning in the presence of malicious errors." SIAM Journal on Computing, 1993], [A. Blum, N. Haghtalab, A. D. Procaccia, and M. Qiao. "Collaborative PAC learning". NeurIPS, 2017], [M. Qiao. "Do outliers ruin collaboration?" ICML, 2018], [A. Jain and A. Orlitsky. "Optimal robust learning of discrete distributions from batches". ICML, 2020], [M. Qiao, G. Valiant. "Learning discrete distributions from untrusted batches". ITCS, 2018],

Related Work

Robust learning from a single dataset

- no universal algorithm: minimum guaranteeable error is $\frac{\alpha}{1-\alpha}$ [Kerns and Li, 1993]
- identical to our situation when each dataset consists of a single point, $m = 1$
  $\longrightarrow$ only $N \to \infty$ will probably not suffice to learn arbitrarily well

Collaborative learning (multiple parties together learn *individual models*)

- universal learning algorithm exists [Blum *et al.*, 2017], [Qiao, 2018]

Density estimation from untrusted batches

- possible, but not applicable to supervised learning [Qiao and Valiant, 2018], [Jain and Orlitsky, 2020]

Byzantine-robust distributed optimization

- specific solutions for gradient-based optimization [Yin *et al.*, 2018], [Alistarh *et al.*, 2018]
- results focus on convergence analysis

[M. Kearns and M. Li. "Learning in the presence of malicious errors." SIAM Journal on Computing, 1993], [A. Blum, N. Haghtalab, A. D. Procaccia, and M. Qiao. "Collaborative PAC learning". NeurIPS, 2017], [M. Qiao. "Do outliers ruin collaboration?" ICML, 2018], [A. Jain and A. Orlitsky. "Optimal robust learning of discrete distributions from batches". ICML, 2020], [M. Qiao, G. Valiant. "Learning discrete distributions from untrusted batches". ITCS, 2018], D. Yin, Y. Chen, K. Ramchandran, P. Bartlett. "Byzantine-robust distributed learning: Towards optimal statistical rates". ICML, 2018], [D. Alistarh, Z. Allen-Zhu, J. Li. "Byzantine stochastic gradient descent". NeurIPS, 2018].

Our Result

Theorem [N. Konstantinov, E. Frantar, D. Alistarh, CHL. ICML 2020]

There exists a learning algorithm, $\mathcal{L}$, such that with high probability:

$$\mathrm{er}(\,\mathcal{L}(S'_1,\dots,S'_N)\,) \;\le\; \min_{h\in\mathcal{H}} \mathrm{er}(h) + \underbrace{\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{(1-\alpha)Nm}} + \alpha\frac{1}{\sqrt{m}}\Big)}_{\to 0 \text{ for } m=|S|\to\infty},$$

with $S'_1,\dots,S'_N = \mathcal{A}(S_1,\dots,S_N)$ for any adversary $\mathfrak{A}$ with $\alpha < \frac{1}{2}$.

($\widetilde{\mathcal{O}}$-notation hides constant and logarithmic factors)

Big Picture

Question: why is learning easier from multiple sources than from a single one?

Answer: it's not. But the task for the adversary is harder!

- single source: no restrictions how to manipulate the data
- multi-source: manipulation must adhere to the source structure

Algorithm idea: exploit law of large numbers

1. majority of datasets are unperturbed
2. for $m \to \infty$ these start to look more and more similar
3. we can identify (at least) the unperturbed datasets
4. we perform ERM on the union of only those

Robust multi-source learning algorithm:

- **Input:** datasets $S'_1, \dots, S'_N$
- **Input:** suitable distance measure d between datasets
- **Input:** suitable threshold value θ

- Step 1) identify which sources to trust
  - compute all pairwise distance d_{ij} between datasets $S'_1, \dots, S'_N$
  - for any i: if $d_{ij} < \theta$ for at least $\lfloor \frac{N}{2} \rfloor$ values of $j \neq i$, then $T \leftarrow T \cup \{i\}$

- Step 2) merge data from all sources S'_i with $i \in T$ into a new dataset $\tilde{S}$

- Step 3) minimize training error on $\tilde{S}$

Open choices:

- distance measure d (discussed later), threshold θ (see paper)

All datasets clean

All datasets clean

All datasets clean

All datasets clean

All datasets clean

All datasets clean

All datasets clean

All datasets clean

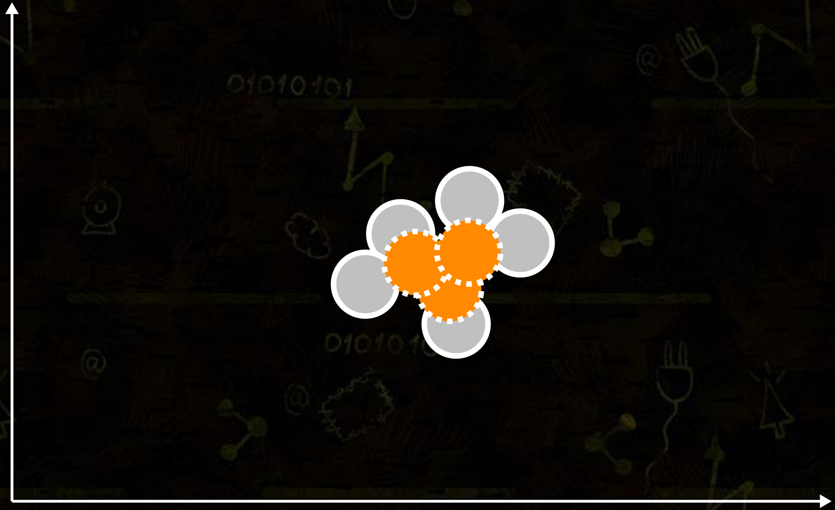
All datasets clean → all datasets included → same as (optimal) naive algorithm

Some datasets manipulated

Some datasets manipulated → manipulated datasets excluded

Consistent manipulations

Consistent manipulations → manipulated datasets excluded



Some datasets manipulated to look like originals

Some datasets manipulated to look like originals → <u>all</u> datasets included.

What properties does the distance measure d need?

What properties does the distance measure d need?

1) 'clean' datasets should get grouped together:

$$S, \hat{S} \sim p \quad \Rightarrow \quad d(S, \hat{S}) \xrightarrow{m \to \infty} 0$$

What properties does the distance measure d need?

1) 'clean' datasets should get grouped together:

$$S, \hat{S} \sim p \quad \Rightarrow \quad d(S, \hat{S}) \overset{m\to\infty}{\longrightarrow} 0$$

2) if manipulated datasets are grouped with the clean ones, they should not hurt the learning step

$$d(S, \hat{S}) \text{ is small} \quad \Rightarrow \quad \mathcal{L}(\hat{S}) \approx \mathcal{L}(S)$$

What properties does the distance measure d need?

1) 'clean' datasets should get grouped together:

$$S, \hat{S} \sim p \quad \Rightarrow \quad d(S, \hat{S}) \xrightarrow{m \to \infty} 0$$

2) if manipulated datasets are grouped with the clean ones, they should not hurt the learning step

$$d(S, \hat{S}) \text{ is small} \quad \Rightarrow \quad \mathcal{L}(\hat{S}) \approx \mathcal{L}(S)$$

Observation:

- many candidate distances do not fulfill both conditions simultaneously:
  - geometric: average Euclidean distance, Chamfer distance, Haussdorf distance, ...
  - probabilistic: Wasserstein distance, total variation, KL-divergence, ...
- discrepancy distance does fulfill the conditions!

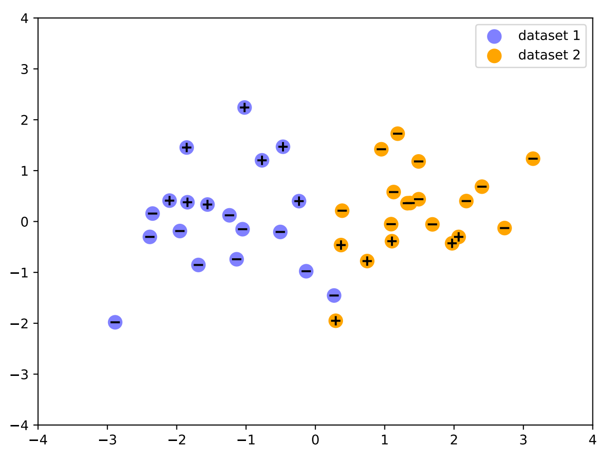
Discrepancy Distance [Mansour *et al.* 2009], [Kifer *et al.* 2004]

For a set of classifiers $\mathcal{H}$ and datasets $S, \hat{S}$, define

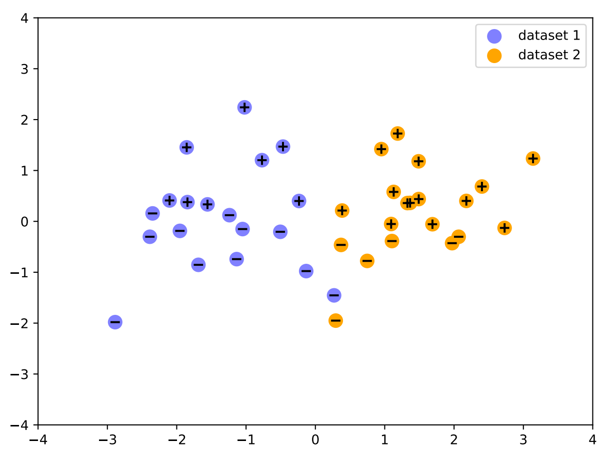
$$\mathrm{disc}(S, \hat{S}) = \max_{h \in \mathcal{H}} \left| \mathsf{er}_S(h) - \mathsf{er}_{\hat{S}}(h) \right| .$$

- maximal amount any classifier, $h \in \mathcal{H}$, can disagree between $S, \hat{S}$
- discrepancy can be estimated by training a classifier itself:
  - $S^{\pm} \leftarrow S$ with all ± 1 labels flipped to their opposites
  - $\tilde{S} \leftarrow S^{\pm} \cup \hat{S}$
  - $\mathrm{disc}(S, \hat{S}) \leftarrow 1 - 2 \min_{h \in \mathcal{H}} \mathsf{er}_{\tilde{S}}(h)$ (minimal training error of any $h \in \mathcal{H}$ on $\tilde{S}$)

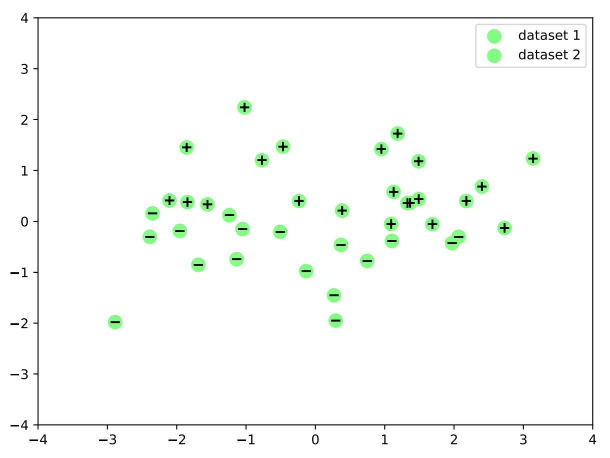
[Y. Mansour, M. Mohri, and A. Rostamizadeh. "Domain adaptation: Learning bounds and algorithms.", COLT 2009]
[D. Kifer, S. Ben-David, J. Gehrke. "Detecting Change in Data Streams", VLDB 2004]



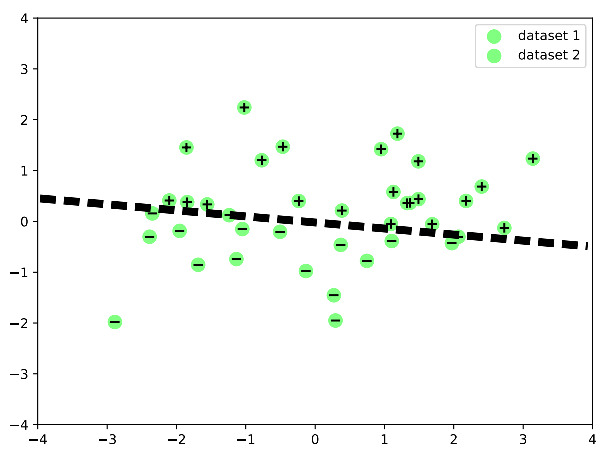
Two datasets, $S, \hat{S}$



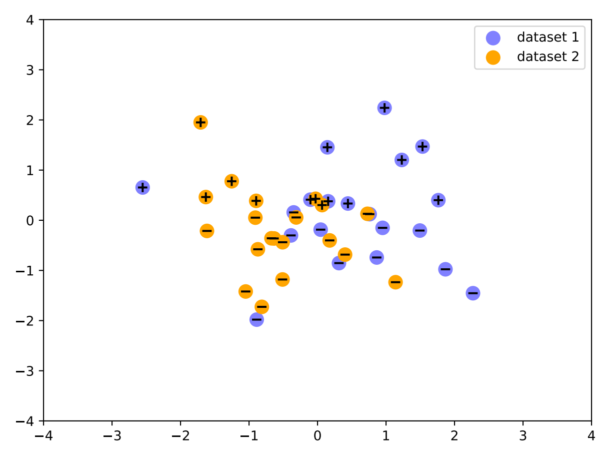
Flip signs of S



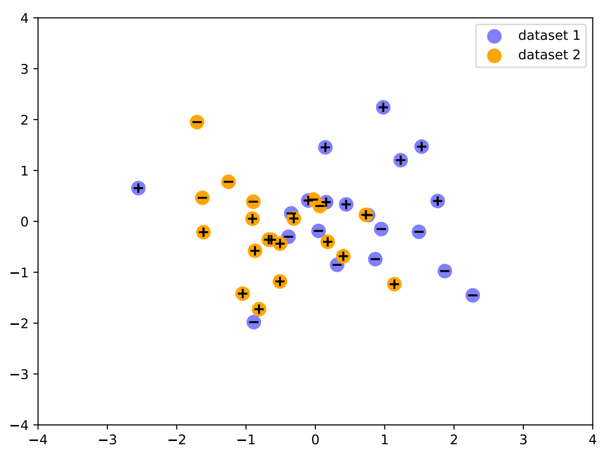
Merge both datasets



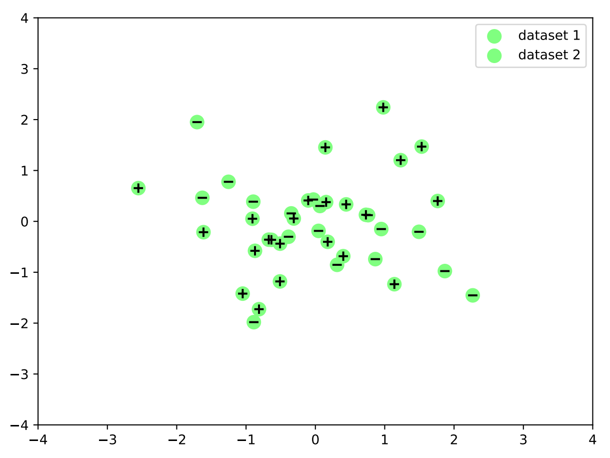
Classifier with small training error → large discrepancy



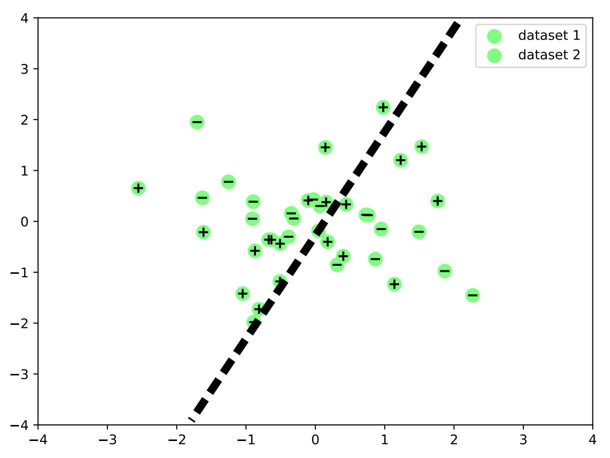
Two datasets, $S, \hat{S}$



Flip signs of S



Merge both datasets



No classifier with small training error → small discrepancy

Observation: discrepancy distance has both property we need

1) Datasets from the same distribution (eventually) gets grouped together
   - for **VC**$(\mathcal{H}) < \infty$, if S and $\hat{S}$ are sampled from the same distribution, then

$$\mathrm{disc}(S, \hat{S}) \to 0 \quad \text{for} \quad |S|, |\hat{S}| \to \infty$$

2) Datasets that are grouped together cannot hurt the learning much

   Consider:
   - training set $S_{\text{trn}} \overset{i.i.d.}{\sim} p$
   - arbitrary set $\hat{S}$, potentially manipulated but with $\mathrm{disc}(S_{\text{trn}}, \hat{S}) \leq \theta$
   - test set $S_{\text{tst}} \overset{i.i.d.}{\sim} p$

Then, for every $h \in \mathcal{H}$:

$$\text{er}_{S_{\text{tst}}}(h) \leq \text{er}_{\hat{S}}(h) + \underbrace{\mathrm{disc}(S_{\text{trn}}, \hat{S})}_{\leq \theta} + \underbrace{\mathrm{disc}(S_{\text{trn}}, S_{\text{tst}})}_{\text{small by prop. 1)}}$$

Observation: discrepancy distance has both property we need

1) Datasets from the same distribution (eventually) gets grouped together
   - for **VC**$(\mathcal{H}) < \infty$, if S and $\hat{S}$ are sampled from the same distribution, then

$$\mathrm{disc}(S, \hat{S}) \to 0 \quad \text{for} \quad |S|, |\hat{S}| \to \infty$$

2) Datasets that are grouped together cannot hurt the learning much

   Consider:
   - training set $S_{\text{trn}} \overset{i.i.d.}{\sim} p$
   - arbitrary set $\hat{S}$, potentially manipulated but with $\mathrm{disc}(S_{\text{trn}}, \hat{S}) \leq \theta$
   - test set $S_{\text{tst}} \overset{i.i.d.}{\sim} p$

Then, for every $h \in \mathcal{H}$:
$$\mathrm{er}_{S_{\text{tst}}}(h) \leq \mathrm{er}_{\hat{S}}(h) + \underbrace{\mathrm{disc}(S_{\text{trn}}, \hat{S})}_{\leq \theta} + \underbrace{\mathrm{disc}(S_{\text{trn}}, S_{\text{tst}})}_{\text{small by prop. 1)}}$$

Open question: how to do this for high-**VC** classes, such as deep networks?

Robust Fair Learning

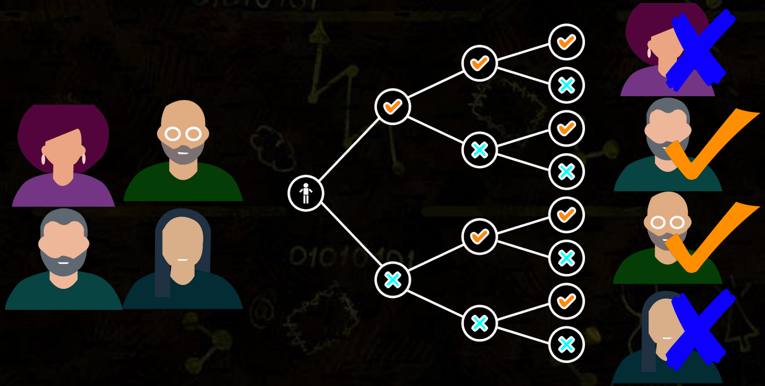
Fairness-Aware Learning from Unreliable or Malicious Data

Nikola Konstantinov (ETH Zurich)

Jen Iofinova (ISTA)

Disclaimer: "These results have been modified from their original form. They have been edited to fit the screen and the allotted time slot."

[N. Konstantinov, CHL. "*Fairness-Aware PAC Learning from Corrupted Data*", JMLR 2022, https://www.jmlr.org/papers/v23/21-1189.html]
[E. Iofinova*, N. Konstantinov*, CHL, "FLEA: Provably Robust Fair Multisource Learning", TMLR 2022, https://openreview.net/forum?id=XsPopigZXV]



How to ensure that a classifier does not discriminate against certain groups?

Setting:

- Inputs: $x \in \mathcal{X}$, e.g. strings, images, vectors, ...
- Protected attribute: $a \in \mathcal{A}$, e.g. gender, age, race, ...
- Outputs: $y \in \mathcal{Y} = \{\pm 1\}$
- Probability distribution: $p(x, a, y)$ over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$
- Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. For simplicity: 0/1-loss $\quad \ell(y, \bar{y}) = \mathbb{1}\{y \neq \bar{y}\}$

Abstract Goal:

- find a prediction function, $f : \mathcal{X} \to \mathcal{Y}$ low expected loss

$$\text{er}(h) = \mathbb{E}_{(x,y)\sim p}\big(\mathbb{1}\{f(x) \neq y\}\big) = \Pr_{(x,y)\sim p}\{f(x) \neq y\}$$

that in addition fulfills some condition of (group) fairness.

Group Fairness:

- demographic parity: "all groups have the same success rate"

$$\forall a, b \in \mathcal{A} \quad \Pr(f(X) = 1 | A = a) = \Pr(f(X) = 1 | A = b)$$

- equality of opportunity: "all groups have the same true positive rate"

$$\forall a, b \in \mathcal{A} \quad \Pr(f(X) = 1 | A = a, Y = 1) = \Pr(f(X) = 1 | A = b, Y = 1)$$

and many others. [Barocas *et al.*, 2019]

Several fairness-aware learning methods exist to enforce these criteria.

[S. Barocas, M. Hardt, A. Narayanan. "Fairness and Machine Learning. Limitations and Opportunities", fairmlbook.org, 2019]

Fair Learning from unreliable/malicious data:

- original training set: $S = \{(x_1, a_1, y_1), \ldots, (x_m, a_m, y_m)\}$
- adversary $\mathfrak{A}$ can manipulate a fraction α of the dataset
- actual training set: $\mathfrak{A}(S)$

Question: Can a fairness-aware learner overcome the manipulation?

Fair Learning from unreliable/malicious data:

- original training set: $S = \{(x_1, a_1, y_1), \dots, (x_m, a_m, y_m)\}$
- adversary $\mathfrak{A}$ can manipulate a fraction α of the dataset
- actual training set: $\mathfrak{A}(S)$

Question: Can a fairness-aware learner overcome the manipulation?

Theorem [Konstantinov & CHL, 2022]

There is an even finite-sized hypothesis classes, $\mathcal{H}$, for which:

- No learning algorithm can guarantee optimal fairness.
- This effect is independent of whether accuracy is also affected or not.
- The smaller the minority group, the stronger the bias.

[N. Konstantinov, CHL. "*Fairness-Aware PAC Learning from Corrupted Data*", JMLR 2022, https://www.jmlr.org/papers/v23/21-1189.html]

Fairness-Aware Learning from Multiple Unreliable Sources

- multiple training sets: $S_1, S_2, \ldots, S_N \subset \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$
- adversary $\mathfrak{A}$ can manipulate $K = \lfloor \alpha N \rfloor$ of the datasets for $\alpha < \frac{1}{2}$
- actual training sets: $\mathfrak{A}(S_1, \ldots, S_N)$

Is there a fairness-aware learning algorithm that overcomes such manipulations?

Fairness-Aware Learning from Multiple Unreliable Sources

- multiple training sets: $S_1, S_2, \ldots, S_N \subset \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$
- adversary $\mathfrak{A}$ can manipulate $K = \lfloor \alpha N \rfloor$ of the datasets for $\alpha < \frac{1}{2}$
- actual training sets: $\mathfrak{A}(S_1, \ldots, S_N)$

Is there a fairness-aware learning algorithm that overcomes such manipulations?

Theorem [Iofinva, Konstantinov & CHL, TMLR 2022 + revision in preparation]

There exists a filtering algorithm, $\mathcal{F}$ that selects at least $\lceil N/2 \rceil$ out of N sources, such that for each source $S \in \mathcal{F}(\mathfrak{A}(S_1, \ldots, S_N))$ it holds with high probability for all $h \in \mathcal{H}$:

$$|\text{er}(h) - \text{er}_S(h)| \leq \widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{m}}\Big), \qquad |\Gamma(h) - \Gamma_S(h)| \leq \widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{m}}\Big)$$

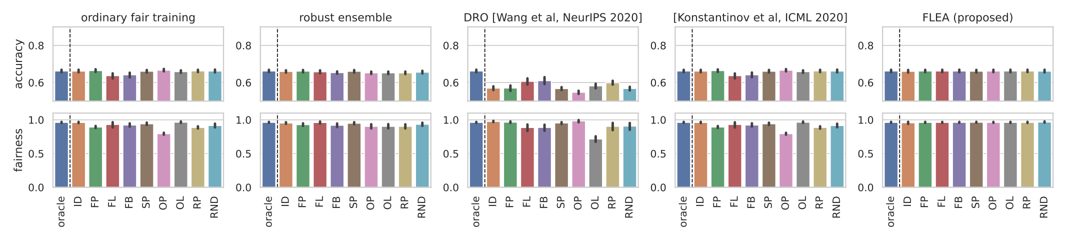
where Γ is a quantitative measure of *demographic parity* fairness.

[E. Iofinova*, N. Konstantinov*, CHL, "FLEA: Provably Robust Fair Multisource Learning", TMLR 2022, https://openreview.net/forum?id=XsPopigZXV]

FLEA (Fair LEarning against Adversaries):

- **Input:** datasets $S'_1, \dots, S'_N$
- **Input:** $\beta \leq \frac{1}{2}$ upper bound on fraction of malignant sources
- **Define:** distance measure $d(S, \hat{S}) = \text{disc}(S, \hat{S}) + \text{disp}(S, \hat{S}) + \text{disb}(S, \hat{S})$
  - $\text{disc}(S, \hat{S})$: discrepancy as before
  - $\text{disp}(S, \hat{S})$: maximal fairness difference of any classifier between S and $\hat{S}$
  - $\text{disb}(S, \hat{S})$: difference in protected group proportions
- Step 1) identify which sources to trust
  - compute all pairwise distance d_{ij} between datasets $S'_1, \dots, S'_N$
  - for any $i = 1, \dots, N$: $\quad q_i \leftarrow \beta\text{-quantile}(d_{i1}, \dots, d_{iN})$
  - $T \leftarrow \{i : q_i \leq \beta\text{-quantile}(q_1, \dots, q_N)\}$
- Step 2) merge data from all sources S'_i with $i \in T$ into a new dataset $\tilde{S}$
- Step 3) train fairness-aware learning algorithm on $\tilde{S}$

Experimental Results (Examples)



bars: different data manipulations, designed to hurt accuracy or fairness. panels: different methods.

- simply training on all data often suboptimal
- other baselines often fail to overcome problems
- FLEA reliably recovers fairness and accuracy

method	COMPAS accuracy	COMPAS fairness
naive	$63.5_{\pm 2.1}$	$78.9_{\pm 2.3}$
robust ensemble	$65.0_{\pm 1.1}$	$88.4_{\pm 2.9}$
DRO (Wang et al., 2020)	$54.5_{\pm 1.2}$	$70.9_{\pm 5.7}$
(Konstantinov et al., 2020)	$63.5_{\pm 2.1}$	$78.9_{\pm 2.3}$
FLEA (proposed)	$65.9_{\pm 1.1}$	$95.3_{\pm 2.3}$
oracle	$66.2_{\pm 1.1}$	$96.2_{\pm 1.3}$

reported values: minimum across data manipulations

More results and ablation studies in [Iofinva, Konstantinov, CHL. 2022]

[E. Iofinova*, N. Konstantinov*, CHL, "FLEA: Provably Robust Fair Multisource Learning", TMLR 2022, https://openreview.net/forum?id=XsPopigZXV]

Summary

Bad news:

- Learning is not robust to bad data.
- This can affect accuracy as well as fairness.

Good news:

- Modern data sets are often not monolithic but collected from multiple sources.
- Multi-source learning **can** be made robust to bad data sources.
- This holds for accuracy as well as fairness.

Thank you!

Thanks to:

Nikola Konstantinov

Jen Iofinova

Elias Frantar

Dan Alistarh

Funding sources: