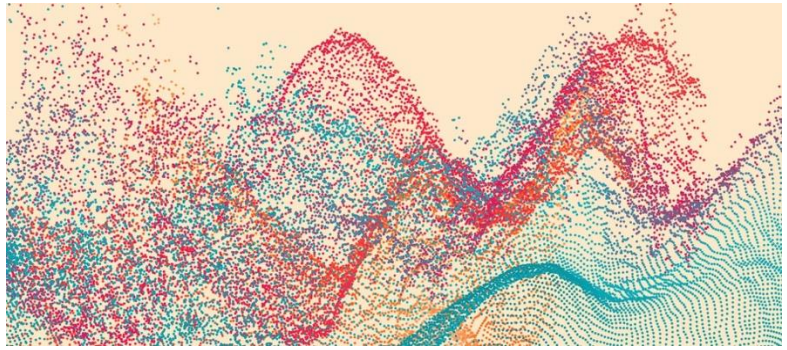
Attribute-Efficient Learning of Sparse Halfspaces with Constant Malicious Noise Rate

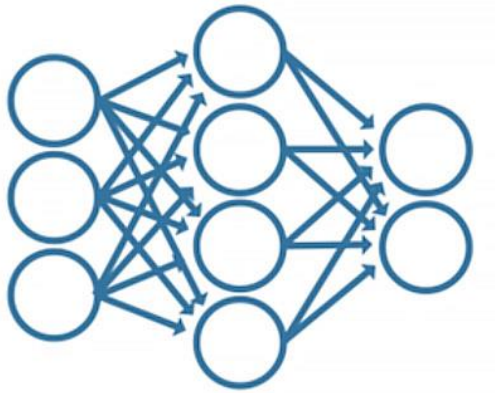
By **Shiwei Zeng**

Joint work with Jie Shen

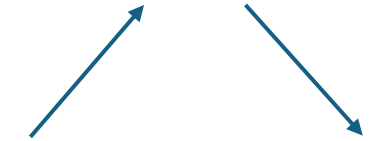
Data in Machine Learning Pipeline







Extremely noisy!!

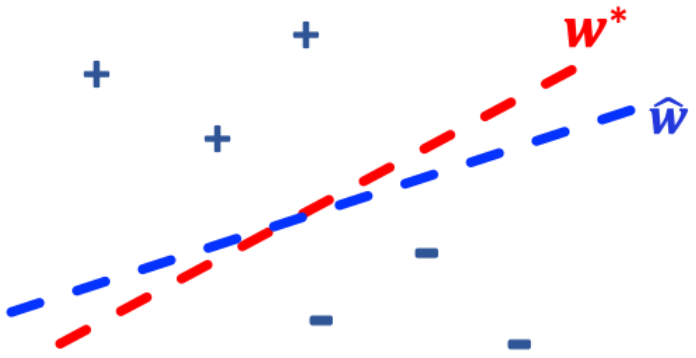


Classification

Hinge Loss Minimization

$$\hat{w} \leftarrow \arg\min L(w, S) = \sum_{(x,y)\in S} \max(0, 1 - yx \cdot w)$$

- Convex minimization
- Shows incredible **noise tolerance**
- PAC learning:



Disagreement $\leq \epsilon$ w.p. $1 - \delta$

Advantages of Hinge Loss Minimization

- Isotropic logconcave: $\eta = O(\epsilon)$
- Mixture of logconcaves + margin: $\boldsymbol{\eta = O(1)}$

- This work: **attribute-efficiency**!

$$\hat{w} \leftarrow \arg \min_{\|w\|_2 \leq 1, \|w\|_1 \leq \sqrt{s}} L(w, S)$$

- Same noise tolerance with $\textbf{\#sample} = \boldsymbol{O\left(s^2 \log^5 d \log^2 \frac{1}{\epsilon\delta}\right)}$

Robustness of Attribute-Efficient Algorithm

$$\hat{w} \leftarrow \arg \min_{\|w\|_2 \le 1, \|w\|_1 \le \sqrt{s}} L(w, S)$$

- A new gradient analysis: contradiction aligning some directions
  - Dense case: using w^* for interior and $\hat{w}_\perp$ for the boundary case
  - Sparsity admitted constraints: KKT conditions changed!
  - $g + \lambda_1 \hat{z} + \lambda_2 \hat{w} = 0$, for some $\lambda_1, \lambda_2 \ge 0$
  - Boundary: $w^* - \hat{w}\langle w^*, \kappa \rangle$, where $\kappa = \frac{\lambda_1 \hat{z} + \lambda_2 \hat{w}}{\sqrt{s}\lambda_1 + \lambda_2}$
- Semi-definite programming for outlier removal

Sample Complexity

- The dense pancake condition:

$$\Pr_{(x,y)\sim D}\left(\Pr_{(x',y')\sim S}\left((x',y') \in P_w^{\tau}(x,y)\right) \geq \rho\right) \geq 1-\beta$$

- How many samples are enough for all in $\mathcal{W} = \{w: \|w\|_2 \leq 1, \|w\|_1 \leq \sqrt{s}\}$?
- τ-cover $\mathcal{W}$, where $\tau = \tilde{O}(\sigma) \approx 1/\sqrt{d}$
- Geometric covering: sample blowup!
- Function space covering: a covering number of $\exp(s \log^4 d)$

Sample Lower Bound

Learning of s-sparse halfspaces under the mixture of logconcaves and margin conditions requires at least $\Omega\left(s \cdot \left(\log\frac{1}{\epsilon} + \log\frac{d}{s}\right)\right)$ samples

- Packing number of s-sparse halfspaces: $\left(\frac{1}{\epsilon}\right)^{\Omega(s)} + \frac{d}{s}^{\Omega(s)}$
- Under the new condition, for a significant fraction of s-sparse halfspaces:
  - Norm distance still implies disagreement rate,
  - i.e. $\|u - v\| \leq \Pr_{x\sim D}\left(\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x)\right)$ up to constant multiplicative factors
- Taking the log of packing number yields the lower bound

Thank you!