Minimax Approach to Supervised Learning

P. Kaloshin, A. Savinov, M. Kolos, M. Vinogradov

Information and Coding Theory, Skoltech

March 22, 2019

https://github.com/KaloshinPE/MEM_detector

Skoltech
Skolkovo Institute of Science and Technology

Abstract

Problem. Given a task of predicting Y from X, a loss function $\mathcal{L}$, and a set of probability distributions Γ on (X, Y), what is the optimal decision rule minimizing the worstcase expected loss over Γ?

Results. We were able to:

- ☑ Learn optimal prediction rule
- ☑ Compare performance of several classifiers

Supervised Learning

Given

$$(\{x_i\}, \{y_i\})$$

predict

$$P(\mathbf{X}, \mathbf{Y})$$

Problem: too expensive in high dimensions.

Solution. Use *empirical risk minimization* (ERM):

$$\arg\min_{\phi} \max_{P(X,Y)} \mathbf{E}[\mathcal{L}(Y, \phi(X))]$$

Key Idea

New optimization task

$$\arg\min_{\phi\in\Phi}\max_{P(X,Y)}\mathbf{E}[\mathcal{L}(Y,\phi(X))]\Rightarrow\arg\min_{\phi}\max_{P\in\Gamma(\hat{P})}\mathbf{E}[\mathcal{L}(Y,\phi(X))]$$

How it helps?

New optimization task

$$\arg\min_{\phi} \max_{P \in \Gamma(\hat{P})} \mathbf{E}[\mathcal{L}(Y, \phi(X))] \Leftrightarrow \underset{P \in \Gamma}{\mathrm{argmax}} H(Y|X)$$

$$H(Y|X) := \inf_{\psi} \mathbb{E}[L(Y, \psi(X))]$$

Current state of the field

- Robust minimax classification [3], 2003
  - Minimazes $\mathcal{L}$ with continuous $\hat{P}$
  - Fixed first- and second-order moments
- Discrete Chebyshev Classifier [4], 2014
  - Minimizes Hinge loss
  - Fixed low-order marginals
- Discrete Renyi Classifier [5], 2015
  - Minimizes max correlation
  - Fixed pairwise marginals
- **Investigated method** [1], 2017
  - Minimizes the worst-case expected loss
  - Most generalistic approach

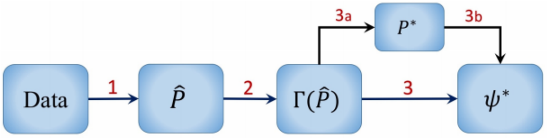

Figure: Minimax Approach

$$\phi^* = \arg\min_{\phi} \max_{P \in \Gamma(\hat{P})} \mathbf{E}[\mathcal{L}(Y, \phi(X))]$$

1 Compute the empirical distribution $\hat{P}$ from the data,

2 Form a distribution set $\Gamma(\hat{P})$ based on $\hat{P}$,

3 Learn a prediction rule ϕ that minimizes the worst-case expected loss over $\Gamma(\hat{P})$:

   3a Search for P^* the distribution maximizing the $H(\Gamma(\hat{P}))$

   3b Compute optimal ϕ^*.

Maximum Entropy Machine

$$L_{0-1}(y, \hat{y}) = \mathbf{1}(\hat{y} \neq y)$$

$$\min_{\boldsymbol{\alpha}} \frac{1}{n} \sum_{i=1}^{n} \max \left\{ 0, \frac{1 - y_i \boldsymbol{\alpha}^T \mathbf{x}_i}{2}, -y_i \boldsymbol{\alpha}^T \mathbf{x}_i \right\} + \epsilon \|\boldsymbol{\alpha}\|_*$$

Expectations and Delivery

- ☑ Implemented minimax SVM, DRC, TAN [7].
- ☑ Compare performance of minimax SVM to SVM, TAN, DRC [5] on datasets from UCI repository (used in paper and some new ones)[6]
- ☑ Compared performance of minimax SVM, SVM, DRC, TAN on high dimentional artificial datasets.

Results

Main experiments

Values are $1 - \mathit{accuracy}$

Dataset	**MEM**	**SVM**	**TAN**	**DRC**
adult	**23 (14)**	17	18	-
credit	11 (10)	16	**13**	13
kr-vs-kp	7	3	7	6
promoters	13	1	**18**	8
votes	6	5	**7**	8
hepatitis	20 (17)	**34**	12	36

Challenges

1. Need to implement not only minimax SVM, but also DRC
2. Experiments are computationally intensive
3. Unclear set up of the experiments in paper [1]

References

- Farnia, Tze. *A Minimax Approach to Supervised Learning*. 2017. [link]
- Chow, Lui. *Approximating Discrete Probability Distributions with Dependence Trees*. 1968. [link]
- Lanckriet et al. A robust minimax approach to classification. 2003. [link]
- Eban et al. Discrete chebyshev classifiers. 2014. [link]
- Razaviyayn, Farnia, Tse. Discrete renyi classifiers. 2015. [link]
- D. Dua and C. Graff. UCI Machine Learning Repository. [link]
- TAN implementation for Python [GitHub link]

Supplementary materials

$$\Gamma(Q) = \{P_{\mathbf{X},Y} : P_{\mathbf{X}} = Q_{\mathbf{X}} \quad \forall 1 \le i \le t : \|\mathbb{E}_P\left[\theta_i(Y)\mathbf{X}\right] - \mathbb{E}_Q\left[\theta_i(Y)\mathbf{X}\right]\| \le \epsilon_i\}$$

$$H(Y) := \inf_{a\in\mathcal{A}} \mathbb{E}[L(Y,a)]$$
$$H(Y|X) := \inf_{\psi} \mathbb{E}[L(Y,\psi(X))]$$

$$L_{0-1}(y,\hat{y}) = \mathbf{1}(\hat{y} \neq y): \; H_{0-1}(Y) = 1 - \max_{y\in\mathcal{Y}} P_Y(y),$$
$$\mathrm{H}_{0-1}(Y|X) = 1 - \sum_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} P_{X,Y}(x,y)$$