A Differential Entropy Estimator for Training Neural Networks

Georg Pichler, Pierre Colombo, Malik Boudiaf,
Günther Koliander, Pablo Piantanida

ICML 2022

Introduction

- Learning tasks require information in the form of training data. Thus, information measures (*e.g.* entropy, conditional entropy and mutual information) have been a source of inspiration for the design of learning objectives in modern machine learning.
- For the use in training deep neural networks, estimators of information-theoretic quantities need to
  - **(R1)** be differentiable w.r.t. the data distribution,
  - **(R2)** be computationally tractable, and
  - **(R3)** rapidly adapt to changes in the underlying distribution.
- We propose a **K**ernelized **N**eural **I**n**f**ormation **E**stimator (KNIFE), a simple, yet effective estimator of differential entropy that satisfies these requirements.

KNIFE is a plug-in estimator:

- Estimate differential entropy $\mathrm{h}(X) = -\int p(x) \log p(x) \, \mathrm{d}x$ of $X \in \mathbb{R}^d$ from N samples $\mathcal{D}_\mathrm{x} = (x_1, \ldots, x_N)$:

$$\hat{p}_{\mathrm{KNIFE}}(x; \boldsymbol{\theta}) = \sum_{m=1}^{M} u_m \kappa_{A_m}(x - b_m),$$

$$\widehat{\mathrm{h}}_{\mathrm{KNIFE}}(\mathcal{D}_\mathrm{x}) = -\frac{1}{N} \sum_{n=1}^{N} \log \hat{p}_{\mathrm{KNIFE}}(x_n)$$

using parameters $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{b}, \mathbf{u})$, where

- $\mathbf{u} = (u_1, u_2, \ldots, u_M)$; $0 \leq u_m \leq 1$ and $\mathbf{1} \cdot \mathbf{u} = 1$,
- $\mathbf{b} = (b_1, \ldots, b_M)$; $b_m \in \mathbb{R}^d$,
- $\mathbf{A} = (A_1, \ldots, A_M)$ are symmetric, positive definite covariance matrices, and
- $\kappa_A(x) = \mathcal{N}(x; 0, A) = \det(2\pi A)^{-\frac{1}{2}} \exp(-\frac{1}{2} x^\intercal A^{-1} x)$.

Training KNIFE

By the Law of Large Numbers (LLN),

$$\begin{aligned}\widehat{\mathrm{h}}_{\mathrm{KNIFE}}(\mathcal{D}_{\mathrm{X}}, \boldsymbol{\theta}) &\overset{\mathsf{LLN}}{\approx} -\mathbb{E}\big[\log \hat{p}_{\mathrm{KNIFE}}(X; \boldsymbol{\theta})\big] \\ &= \mathrm{h}(X) + \mathrm{D}_{\mathrm{KL}}(p \| \hat{p}_{\mathrm{KNIFE}}(\,\cdot\,; \boldsymbol{\theta})) \\ &\geq \mathrm{h}(X).\end{aligned}$$

Thus, we learn $\boldsymbol{\theta}$ by minimizing $\widehat{\mathrm{h}}_{\mathrm{KNIFE}}$.

Conditional Differential Entropy Estimation

- To estimate conditional differential entropy $\mathrm{h}(X|Y)$, consider $\boldsymbol{\theta}$ to be a parameterized function $\boldsymbol{\Theta}(y)$ of y.
- With $\mathcal{D} = (\mathcal{D}_\mathrm{x}, \mathcal{D}_\mathrm{y}) = (x_n, y_n)_{n=1}^N$, we have

$$\hat{p}_{\textsc{Knife}}(x|y;\boldsymbol{\Theta}) = \hat{p}_{\textsc{Knife}}(x;\boldsymbol{\Theta}(y)),$$

$$\widehat{\mathrm{h}}_{\textsc{Knife}}(\mathcal{D}_\mathrm{x}|\mathcal{D}_\mathrm{y};\boldsymbol{\Theta}) = \frac{1}{N}\sum_{n=1}^{N} \log \frac{1}{\hat{p}_{\textsc{Knife}}(x_n|y_n;\boldsymbol{\Theta})}. \qquad (1)$$

- Minimize (1) over the parameters of $\boldsymbol{\Theta}$.
- Use of an artificial neural network $\boldsymbol{\Theta}(y)$, taking y as its input.
- If $Y \in \mathcal{Y}$ is discrete, use one parameter $\boldsymbol{\theta}$ for each $y \in \mathcal{Y}$, i.e., $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_y\}_{y\in\mathcal{Y}}$ in (1).
- Mutual Information can be estimated as $\mathrm{I}(X;Y) = \mathrm{h}(X) - \mathrm{h}(X|Y)$.

Baselines and Related Work

Differential Entropy Estimation:

- (Schraudolph, 2004): Special case of KNIFE, where
  $u_m = \frac{1}{M}$ is uniform,
  A_m are diagonal matrices, and
  $b_m = x'_m$ are fixed with an independent training set $\mathcal{E} = (x'_m)_{m=1}^M$.
- DoE (McAllester and Stratos, 2020): Special case of KNIFE, where $M = 1$ and A_1 is a diagonal matrix

Mutual Information Estimation:

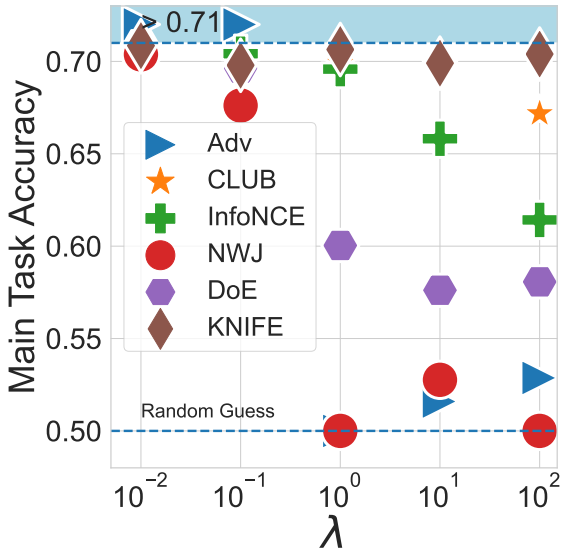
- MINE (Belghazi et al., 2018)
- NWJ (Nguyen, Wainwright, and Jordan, 2010)
- InfoNCE (Oord, Li, and Vinyals, 2018)
- CLUB (Cheng et al., 2020)

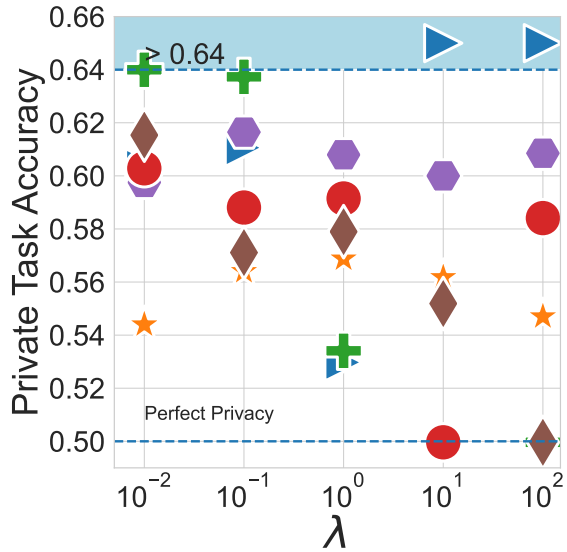
Fair Classification Task

- Goal: Train a model to take its decision without utilizing private information such as gender, age, or race.
- Datasets: DIAL corpus (Blodgett, Green, and O'Connor, 2016) (>50mio. tweets)
- Minimize the mutual information between output $Z = \Phi_{\boldsymbol{\psi}}(X)$ and a private label S (race).
- Loss function $\mathcal{L} = \mathrm{CE}(Y; \Phi_{\boldsymbol{\psi}}(X)) + \lambda \cdot \mathrm{I}(\Phi_{\boldsymbol{\psi}}(X); S)$
- Experiment performed for two target labels Y: "sentiment" and "mention"
- Experimental setting of (Elazar and Goldberg, 2018) (Adv) and (Barrett et al., 2019)

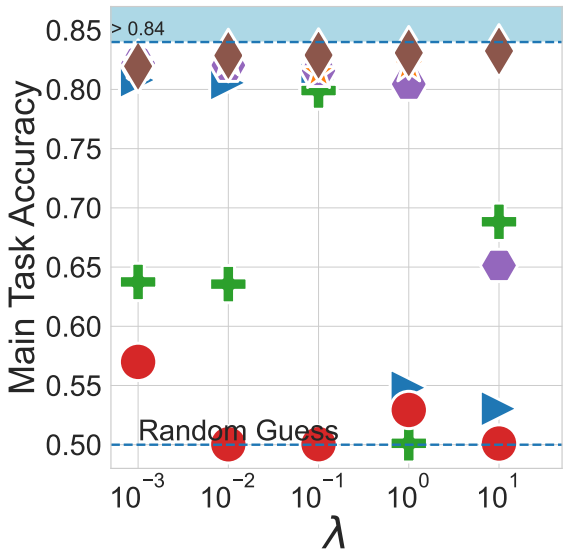
Fair Classification Task – Results

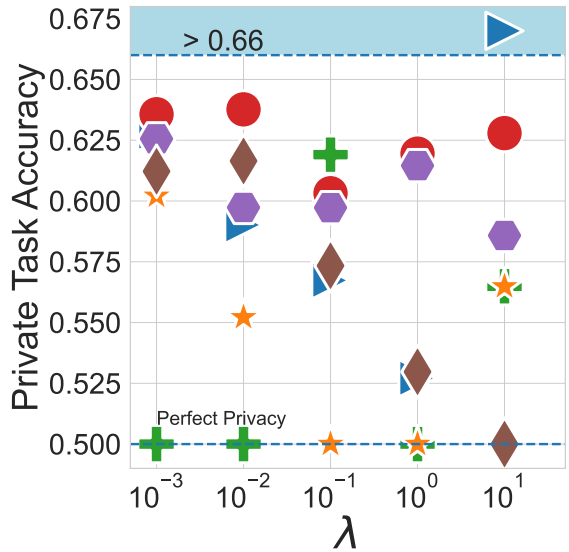
Sentiment:





Mention:





Domain Adaptation

- Goal: Transfer knowledge from the source domain (S) (with labeled examples) to a target domain (T) (with unlabeled examples).
- Several different datasets: M (MNIST), MM (MNIST M), U (USPS), SV (SVHN), C (CIFAR10) and S (STL10); averaged over 3 seeds;
- We follow the adversarial approach (Cheng et al., 2020) closely, using the method proposed by (Gholami et al., 2020).

	M → MM	S → C	U → M	M → U	C → S	SV → M	Mean
Source only	51.9 ±0.8	58.3 ±0.2	91.1 ±0.7	93.5 ±0.6	**72.3** ±0.5	54.7 ±2.8	70.3 ±0.9
CLUB	79.1 ±2.2	59.9 ±1.9	96.0 ±0.2	96.8 ±0.5	71.6 ±1.3	83.8 ±3.4	81.2 ±1.7
DoE	**82.2** ±2.6	58.9 ±0.8	97.2 ±0.3	94.2 ±0.9	68.8 ±1.4	86.4 ±5.4	81.3 ±1.9
InfoNCE	77.3 ±0.5	61.0 ±0.1	97.4 ±0.2	97.0 ±0.3	70.6 ±0.8	89.2 ±4.1	82.1 ±1.0
MINE	76.7 ±0.4	61.2 ±0.3	**97.7** ±0.1	97.3 ±0.1	70.8 ±1.0	91.8 ±0.8	82.6 ±0.4
NWJ	77.1 ±0.6	61.2 ±0.3	97.6 ±0.1	97.3 ±0.5	72.1 ±0.7	91.4 ±0.8	82.8 ±0.5
KNIFE	78.7 ±0.7	**61.8** ±0.5	**97.7** ±0.3	**97.4** ±0.4	71.2 ±1.8	**93.2** ±0.2	**83.4** ±0.6

The End.

Thank you for your attention.

References I

- Barrett, Maria et al. (2019). "Adversarial removal of demographic attributes revisited". In: *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6331–6336.
- Belghazi, Mohamed Ishmael et al. (July 2018). "Mutual Information Neural Estimation". In: *International Conference on Machine Learning (ICML)*. Vol. 80. PMLR, pp. 531–540.
- Blodgett, Su Lin, Lisa Green, and Brendan O'Connor (2016). "Demographic Dialectal Variation in Social Media: A Case Study of African-American English". In: *Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130.
- Cheng, Pengyu et al. (2020). "CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 1779–1788.

References II

- Elazar, Yanai and Yoav Goldberg (2018). "Adversarial removal of demographic attributes from text data". In: *Conference on Empirical Methods in Natural Language Processing*, pp. 11–21.
- Gholami, Behnam et al. (2020). "Unsupervised multi-target domain adaptation: An information theoretic approach". In: *IEEE Transactions on Image Processing* 29.
- McAllester, David and Karl Stratos (26–28 Aug 2020). "Formal Limitations on the Measurement of Mutual Information". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 108. PMLR, pp. 875–884.
- Nguyen, XuanLong, Martin J Wainwright, and Michael I Jordan (2010). "Estimating divergence functionals and the likelihood ratio by convex risk minimization". In: *IEEE Transactions on Information Theory*.

References III

- Oord, Aäron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding". In: *arXiv:1807.03748*.
- Schraudolph, N. N. (2004). "Gradient-based manipulation of nonparametric entropy estimates". In: *IEEE Transactions on Neural Networks* 15.4.