CCMI : Classifier based Conditional Mutual Information Estimation

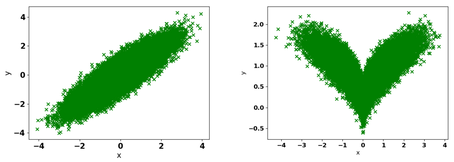
Sudipto Mukherjee, Himanshu Asnani, Sreeram Kannan.
Department of Electrical and Computer Engineering, University of Washington.

Introduction

Conditional Mutual Information (CMI) is a measure of conditional dependence between random variables X and Y, given another random variable Z. *Can we estimate CMI accurately using lower bounds on mutual information ? Can this lead to better conditional independence (CI) testing ?*

Motivation

- Pearson Correlation ($\rho(X,Y)$) cannot capture non-linear dependencies. But Mutual Information (MI) has no such limitation.
  Moreover, $I(X;Y) = 0 \iff X \perp Y$.



- CMI extends properties of MI to conditional settings. $I(X;Y|Z) = 0 \iff X \perp Y|Z$.
  Is Salary $\perp$ Gender | Years of Education ?

- MI and CMI are special cases of KL-divergence. CMI can be expressed as :

$$I(X;Y|Z) = D_{KL}(p_{X,Y,Z}||p_{X,Z}p_{Y|Z})$$

So, *how can we estimate KL-divergence ?*

- MINE [1] used lower bounds of MI and neural network function approximation to estimate MI. But, the estimator has high variance [2] and is challenging to tune.

▪Donsker-Varadhan (DV) representation :

$$D_{KL}(p\|q) = \sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim p(x)}[f(x)] - \log(\mathbb{E}_{x\sim q(x)}[\, e^{f(x)}])$$

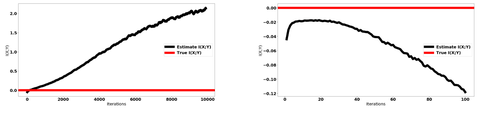
$$\nabla_\theta D_{KL}(p\|q) = \mathbb{E}_{x\sim B_p(x)}[\nabla_\theta f_\theta(x)] - \frac{\log(\mathbb{E}_{x\sim B_q(x)}[\,\nabla_\theta e^{f_\theta(x)}])}{\mathbb{E}_{x\sim B_q(x)}[\,\nabla_\theta e^{f_\theta(x)}]}$$

Biased Gradients.

▪ f-divergence bound :

$$D_{KL}(p\|q) = \sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim p(x)}[f(x)] - \mathbb{E}_{x\sim q(x)}[\, e^{f(x)-1}]$$

Even with low learning rate of $1e{-}4$, the estimator can diverge. The training is sensitive to network size, learning rate, batch size.

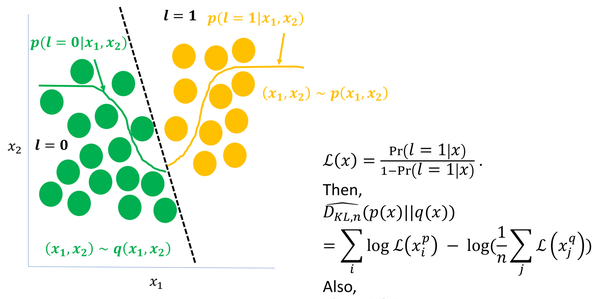


Likelihood ratio Estimation : Classifiers

$$Let\ f^*(x) = log\frac{p(x)}{q(x)}, \text{Then}$$

$$\mathbb{E}_{x\sim p(x)}[f^*(x)] - \log(\mathbb{E}_{x\sim q(x)}[\, e^{f^*(x)}])$$
$$= \mathbb{E}_{x\sim p(x)}\left[\log\frac{p(x)}{q(x)}\right] - \log\left(\mathbb{E}_{x\sim q(x)}\left[\frac{p(x)}{q(x)}\right]\right)$$
$$= \int p(x)\log\frac{p(x)}{q(x)}dx = D_{KL}(p||q)$$

$$Now, \frac{p(x)}{q(x)} = \frac{u(x\mid l=1)}{u(x\mid l=0)} = \left(\frac{Pr(l=1|x)\,u(x)}{Pr(l=1)}\right)\left(\frac{\Pr(l=0)}{Pr(l=0|x)\,u(x)}\right)$$
$$= \frac{Pr(l=1|x)}{Pr(l=0|x)} = \frac{\Pr(l=1|x)}{1-\Pr(l=1|x)}$$



$$\mathcal{L}(x) = \frac{\Pr(l=1|x)}{1-\Pr(l=1|x)}.$$

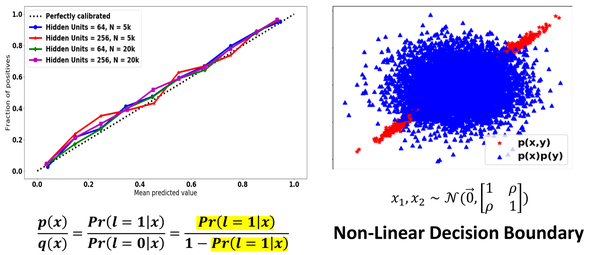
Then,

$$\widehat{D_{KL,n}}(p(x)||q(x)) = \sum_i \log\mathcal{L}(x_i^p) - \log(\frac{1}{n}\sum_j \mathcal{L}(x_j^q))$$

Also,

$$\hat{I}_n = \widehat{D_{KL,n}}(p(x,y)||p(x)p(y))$$

Figure 1: (a) Optimal function in DV-Representation (b) Plugging in Likelihood-ratio from Classifiers



$$\frac{p(x)}{q(x)} = \frac{Pr(l=1|x)}{Pr(l=0|x)} = \frac{Pr(l=1|x)}{1-Pr(l=1|x)}$$

$$x_1, x_2 \sim \mathcal{N}(\vec{0}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix})$$

Non-Linear Decision Boundary

▪ Modular Estimation

$(x,y) \sim p(x,y)$: $(x_1,y_1), (x_2,y_2), (x_3,y_3), (x_4,y_4), (x_5,y_5), (x_6,y_6)$

$(x,y) \sim p(x)p(y)$: $(x_1,y_5), (x_2,y_6), (x_3,y_1), (x_4,y_4), (x_5,y_2), (x_6,y_3)$

But what about samples from $p(x,z)p(y|z)$?

$(y_i,z_i)_{i=1}^n \Rightarrow$ CGAN/CVAE/ 1-NN $\Rightarrow (y_i',z_i)_{i=1}^n \sim p(y|z)$

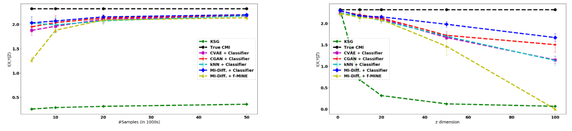
$(x_i,y_i,z_i)_{i=1}^n, (x_i,y_i',z_i)_{i=1}^n \Rightarrow$ Classifier-MI $\Rightarrow D_{KL}(p(x,y,z)\|p(x,z)p(y|z))$

▪ΣI Approach:
$I(X;Y|Z) = I(X;Y,Z) - I(X;Z)$

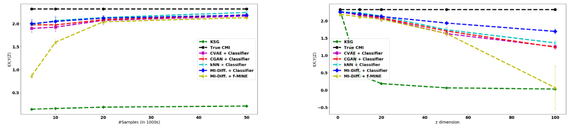
Figure 2: (a) Well-calibrated non-linear classifiers (b) From MI Estimation to CMI Estimation

CMI Estimation

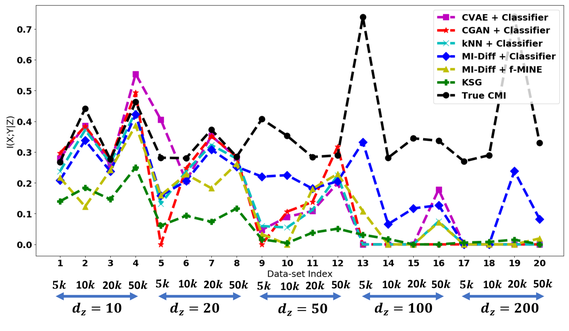
- **Linear and Non-Linear Models :**
  (a) Model I: $X \sim \mathcal{N}(0,1), Z \sim \mathcal{U}(-0.5,0.5)^{d_z}, \epsilon \sim \mathcal{N}(Z_1,\sigma_\epsilon^2), Y \sim X+\epsilon$.
  (b) Model II: $X \sim \mathcal{N}(0,1), Z \sim \mathcal{N}(0,1)^{d_z}$, $U = w^T Z$, $\|w\|_1 = 1, \epsilon \sim \mathcal{N}(U,\sigma_\epsilon^2), Y \sim X+\epsilon$.
  (c) Non-linear : $Z \sim \mathcal{N}(\mathbf{1}, I_{d_z}), X = f_1(\eta_1), Y = f_2(A_{zy}Z + A_{xy}X + \eta_2)$, f_1 and f_2 being non-linear.



(a) Model I : Variation with n, $d_z = 20$ (b) Model I : Variation with d_z, $n = 20,000$



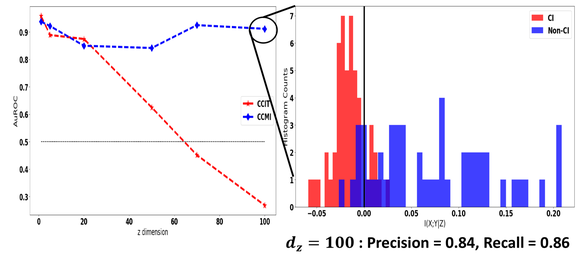
(c) Model II : Variation with n, $d_z = 20$ (d) Model II : Variation with d_z, $n = 20,000$



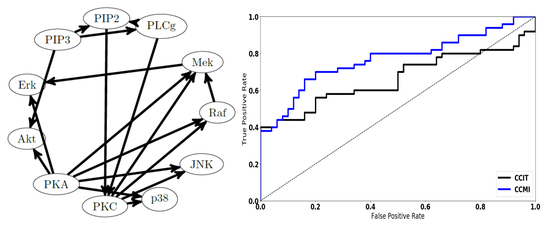
- **Theoretical Properties of CCMI :** Under some assumptions on the data densities and richness of classifier class $\mathcal{C}$, the following theorems hold : (a) Classifier-MI is (weakly) consistent. (b) The finite sample estimate from Classifier-MI is a lower bound on the true MI value with high probability.

CI Testing

We threshold CMI estimates $I(X;Y|Z)$ at 0 to decide if $X \perp Y \mid Z$. CCMI outperforms state-of-the-art tester, CCIT, on synthetic and real data-sets.

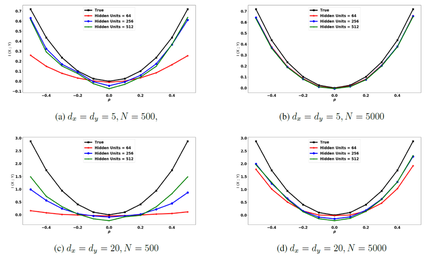


$d_z = 100$: Precision = 0.84, Recall = 0.86



Hyper-parameter choice

The estimates lie below true $I^*(X;Y)$ value. So, we can tune hyper-parameters to maximize $\hat{I}(X;Y)$.



(a) $d_x = d_y = 5, N = 500$, (b) $d_x = d_y = 5, N = 5000$

(c) $d_x = d_y = 20, N = 500$ (d) $d_x = d_y = 20, N = 5000$

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, 2018.

[2] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.