The Nearest Neighbor Information Estimator is Adaptively Near Minimax Rate-Optimal

Jiantao Jiao	Berkeley EECS
Weihao Gao	UIUC ECE
Yanjun Han	Stanford EE

NIPS 2018, Montréal, Canada

Differential Entropy Estimation

Differential entropy of a continuous density on $\mathbb{R}^d$:

$$h(f) = \int_{\mathbb{R}^d} f(x) \log \frac{1}{f(x)} dx$$

Differential Entropy Estimation

Differential entropy of a continuous density on $\mathbb{R}^d$:

$$h(f) = \int_{\mathbb{R}^d} f(x) \log \frac{1}{f(x)} dx$$

- machine learning tasks, e.g., classification, clustering, feature selection

Differential Entropy Estimation

Differential entropy of a continuous density on $\mathbb{R}^d$:

$$h(f) = \int_{\mathbb{R}^d} f(x) \log \frac{1}{f(x)} dx$$

- machine learning tasks, e.g., classification, clustering, feature selection
- causal inference

Differential Entropy Estimation

Differential entropy of a continuous density on $\mathbb{R}^d$:

$$h(f) = \int_{\mathbb{R}^d} f(x) \log \frac{1}{f(x)} dx$$

- machine learning tasks, e.g., classification, clustering, feature selection
- causal inference
- sociology

Differential Entropy Estimation

Differential entropy of a continuous density on $\mathbb{R}^d$:

$$h(f) = \int_{\mathbb{R}^d} f(x) \log \frac{1}{f(x)} dx$$

- machine learning tasks, e.g., classification, clustering, feature selection
- causal inference
- sociology
- computational biology

Differential Entropy Estimation

Differential entropy of a continuous density on $\mathbb{R}^d$:

$$h(f) = \int_{\mathbb{R}^d} f(x) \log \frac{1}{f(x)} dx$$

- machine learning tasks, e.g., classification, clustering, feature selection
- causal inference
- sociology
- computational biology
- $\dots$

Differential Entropy Estimation

Differential entropy of a continuous density on $\mathbb{R}^d$:

$$h(f) = \int_{\mathbb{R}^d} f(x) \log \frac{1}{f(x)} dx$$

- machine learning tasks, e.g., classification, clustering, feature selection
- causal inference
- sociology
- computational biology
- $\cdots$

Our Task

Given empirical samples $X_1, \cdots, X_n \sim f$, estimate $h(f)$.

Ideas of Nearest Neighbor

Ideas of Nearest Neighbor

Notations:

- n: number of samples
- d: dimensionality
- k: number of nearest neighbors
- $R_{i,k}$: ℓ_2 distance of i-th sample to its k-th nearest neighbor
- $\mathrm{vol}_d(r)$: volumn of the d-dimensional ball with radius r

Ideas of Nearest Neighbor

Notations:

- n: number of samples
- d: dimensionality
- k: number of nearest neighbors
- $R_{i,k}$: ℓ_2 distance of i-th sample to its k-th nearest neighbor
- $\mathrm{vol}_d(r)$: volumn of the d-dimensional ball with radius r

Idea

$$h(f) = \mathbb{E}[-\log f(X)] \approx -\frac{1}{n}\sum_{i=1}^{n} \log f(X_i)$$

$$f(X_i) \cdot \mathrm{vol}_d(R_{i,k}) \approx \frac{k}{n}$$

Ideas of Nearest Neighbor

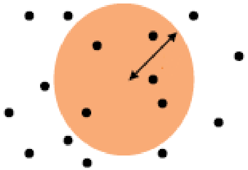
Notations:

- n: number of samples
- d: dimensionality
- k: number of nearest neighbors
- $R_{i,k}$: ℓ_2 distance of i-th sample to its k-th nearest neighbor
- $\mathrm{vol}_d(r)$: volumn of the d-dimensional ball with radius r

Idea

$$h(f) = \mathbb{E}[-\log f(X)] \approx -\frac{1}{n}\sum_{i=1}^{n} \log f(X_i)$$

$$f(X_i) \cdot \mathrm{vol}_d(R_{i,k}) \approx \frac{k}{n}$$



Kozachenko–Leonenko Estimator

Definition (Kozachenko–Leonenko Estimator)

$$\hat{h}_{n,k}^{\mathsf{KL}} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n}{k}\mathrm{vol}_d(R_{i,k})\right) + \underbrace{\log(k) - \psi(k)}_{\text{bias correction term}}$$

Kozachenko–Leonenko Estimator

Definition (Kozachenko–Leonenko Estimator)

$$\hat{h}_{n,k}^{\mathrm{KL}} = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{n}{k} \mathrm{vol}_d(R_{i,k})\right) + \underbrace{\log(k) - \psi(k)}_{\text{bias correction term}}$$

- Easy to implement: no numerical integration

Kozachenko–Leonenko Estimator

Definition (Kozachenko–Leonenko Estimator)

$$\hat{h}_{n,k}^{\mathrm{KL}} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n}{k}\mathrm{vol}_d(R_{i,k})\right) + \underbrace{\log(k) - \psi(k)}_{\text{bias correction term}}$$

- Easy to implement: no numerical integration
- Only tuning parameter: k

Kozachenko–Leonenko Estimator

Definition (Kozachenko–Leonenko Estimator)

$$\hat{h}_{n,k}^{\mathrm{KL}} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n}{k}\mathrm{vol}_d(R_{i,k})\right) + \underbrace{\log(k) - \psi(k)}_{\text{bias correction term}}$$

- Easy to implement: no numerical integration
- Only tuning parameter: k
- Good empirical performance without theoretical guarantee, especially when the density may be close to zero.

Main Result

Let $\mathcal{H}_d^s$ be the class of probability densities supported on $[0,1]^d$ which are Hölder smooth with parameter $s \geq 0$.

Main Result

Let $\mathcal{H}_d^s$ be the class of probability densities supported on $[0,1]^d$ which are Hölder smooth with parameter $s \geq 0$.

Theorem (Main Result)

For fixed k and $s \in (0,2]$,

$$\left(\sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f \left(\hat{h}_{n,k}^{KL} - h(f) \right)^2 \right)^{\frac{1}{2}} \lesssim n^{-\frac{s}{s+d}} \log n + n^{-\frac{1}{2}}.$$

Main Result

Let $\mathcal{H}_d^s$ be the class of probability densities supported on $[0,1]^d$ which are Hölder smooth with parameter $s \geq 0$.

Theorem (Main Result)

For fixed k and $s \in (0,2]$,

$$\left(\sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f \left(\hat{h}_{n,k}^{KL} - h(f) \right)^2 \right)^{\frac{1}{2}} \lesssim n^{-\frac{s}{s+d}} \log n + n^{-\frac{1}{2}}.$$

First theoretical guarantee of Kozachenko–Leonenko estimator without assuming density bounded away from zero.

Matching Lower Bound

Theorem (Han–Jiao–Weissman–Wu'17)

For any $s \geq 0$,

$$\left(\inf_{\hat{h}} \sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f \left(\hat{h} - h(f)\right)^2\right)^{\frac{1}{2}} \gtrsim n^{-\frac{s}{s+d}} (\log n)^{-\frac{s+2d}{s+d}} + n^{-\frac{1}{2}}.$$

Matching Lower Bound

Theorem (Han–Jiao–Weissman–Wu'17)

For any $s \geq 0$,

$$\left(\inf_{\hat{h}} \sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f \left(\hat{h} - h(f) \right)^2 \right)^{\frac{1}{2}} \gtrsim n^{-\frac{s}{s+d}} (\log n)^{-\frac{s+2d}{s+d}} + n^{-\frac{1}{2}}.$$

Take-home Message

- Nearest neighbor estimator is nearly minimax

Matching Lower Bound

Theorem (Han–Jiao–Weissman–Wu'17)

For any $s \geq 0$,

$$\left(\inf_{\hat{h}} \sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f \left(\hat{h} - h(f)\right)^2\right)^{\frac{1}{2}} \gtrsim n^{-\frac{s}{s+d}} (\log n)^{-\frac{s+2d}{s+d}} + n^{-\frac{1}{2}}.$$

Take-home Message

- Nearest neighbor estimator is nearly minimax
- Nearest neighbor estimator adapts to the unknown smoothness s

Matching Lower Bound

Theorem (Han–Jiao–Weissman–Wu'17)

For any $s \geq 0$,

$$\left(\inf_{\hat{h}} \sup_{f \in \mathcal{H}_d^s} \mathbb{E}_f \left(\hat{h} - h(f)\right)^2\right)^{\frac{1}{2}} \gtrsim n^{-\frac{s}{s+d}} (\log n)^{-\frac{s+2d}{s+d}} + n^{-\frac{1}{2}}.$$

Take-home Message

- Nearest neighbor estimator is nearly minimax
- Nearest neighbor estimator adapts to the unknown smoothness s
- Maximal inequality plays a central role in dealing with small densities.