LIDL: Local Intrinsic Dimension estimation using approximate Likelihood

Piotr Tempczyk[1,2,3], Rafał Michaluk[1,2], Łukasz Garncarek[2,4], Przemysław Spurek[5], Jacek Tabor[5], Adam Goliński[6]

06 November 2021

1. Institute of Informatics, University of Warsaw
2. Polish National Institute for Machine Learning Foundation (*opium.sh*)
3. *deeptale.ai*
4. Applica
5. Jagiellonian University
6. University of Oxford

Local Intrinsic Dimension

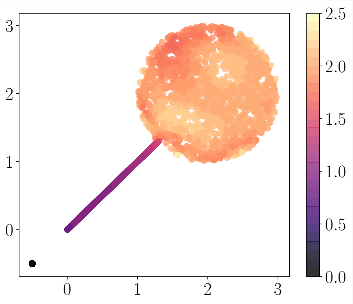


Figure 1: Lollipop dataset with composed of three manifolds of different dimensionality.

In our setting X is a subset of some d-dimensional data manifold M equipped with probability measure μ, and there exist an embedding $j\colon M \to \mathbb{R}^D$ into an Euclidean space, through which we can view X.

Why it is important?

- The estimation of the intrinsic dimension (ID) of a data set is a classical problem of pattern recognition and machine learning.
- LID estimation is a powerful analytical tool to study the process of training and representation learning in deep neural networks [1] [2].
- Intrinsic dimension is connected with model performance [3]. Our experiments show, that it can affect point-wise autoencoder performance for image datasets, and the accuracy of the classifier.

[1] Li et al. "Measuring the intrinsic dimension of objective landscapes. ICLR 2018."

[2] Ansuini et al. "Intrinsic dimension of data representations in deep neural networks." NeurIPS 2019.

[3] Pope et al. "The intrinsic dimension of images and its impact on learning", ICLR 2020

Intuition

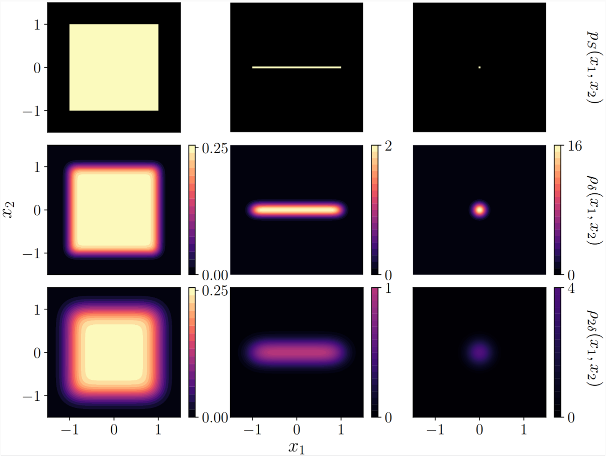


Figure 2: How density changes when perturbed with Gaussian noise.

LIDL estimate

For small values of δ and smooth on-manifold densities we have

$$\log \rho_\delta(x) \approx (d - D) \log \delta + const. \tag{1}$$

where:

$\rho_\delta(x)$ – the probability density of the perturbed data at point x with noise $\mathcal{N}(0, \delta^2 I_D)$,

δ – the noise amplitude,

d – manifold dimensionality at x,

D – ambient space dimensionality.

LIDL Algorithm

Require: $X \subset \mathbb{R}^D$; $x_1, \ldots, x_m \in \mathbb{R}^D$; $\delta_1, \ldots, \delta_n \in \mathbb{R}^+$;

  for $j = 1$ to n **do**

    $X_j \leftarrow X$ perturbed with $\mathcal{N}(0, \delta_j^2 I_D)$

    Fit the density model $\hat{\rho}_j$ to X_j

  end for

  for $i = 1$ **to** m **do**

    for $j = 1$ **to** n **do**

      $\xi_j \leftarrow \log \delta_j$

      $\eta_j \leftarrow \log \hat{\rho}_j(x_i)$

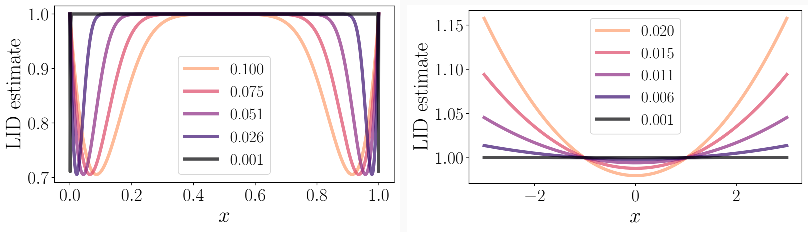
    end for

    $\beta \leftarrow$ regression coefficient for a set of n points (ξ_j, η_j)

    $\hat{d}_i \leftarrow D + \beta$

  end for

  return $(\hat{d}_1, \ldots, \hat{d}_m)$



(a) $\mathcal{U}(0, 1)$ (b) $\mathcal{N}(0, 1)$

Figure 3: LIDL estimates for points from distributions that break some of the assumptions for different values of δ (marked with different colors).

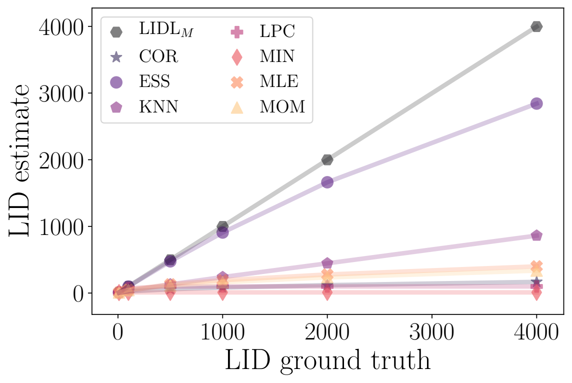
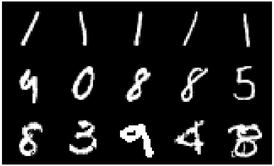


Figure 4: LID estimates for d dimensional uniform distribution on a hypercube. The dimensionality d of the distribution is plotted on the horizontal axis and the estimates for different algorithms on the vertical axis.

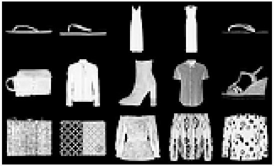
Image datasets

MNIST



Celeb-A

Fashion-MNIST



Contact

piotr.tempczk@opium.sh

https://ptempczyk.github.io/