Intrinsic Dimension of Image Data

Ruixin Guo

Department of Computer Science
Kent State University

February 7, 2023

Contents

Contents

A Theoretical View of Supervised Learning

Let $x_i \in \mathbb{R}^m$ be sampled from a distribution P, $y_i \in \mathbb{R}$ be the label of x_i, and $S = (x_i, y_i = f^*(x_i)), 1 \leq i \leq n$ be the dataset. Here f^* is a function mapping each x_i to its label y_i.

Let f be a machine learning model, the goal of supervised learning is to make the error between f and f^* as small as possible.

The error between f and f^* on a single point x_i is defined as the loss function $L_{x_i}(f)$. Usually we use square loss, i.e., $L_{x_i}(f) = (f(x_i) - f^*(x_i))^2$.

The risk function $\mathcal{R}(f) = \mathbb{E}_{x \sim P}(f(x) - f^*(x))^2$ is the expectation of loss of x from P. We use it to measure the error between f and f^*. If P is known, $\mathcal{R}(f)$ can be solved by integral, i.e., $\mathcal{R}(f) = \int (f(x) - f^*(x))^2 dF(x)$, where $F(x)$ is the CDF of P.

However, in many real world applications, P is unknown. What we only know is the samples x_is from P. So we instead use the empirical risk function $\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2$. Note that the empirical risk is the expectation of loss of all samples, not the entire distribution.

The training process of supervised learning is to minimize $\hat{\mathcal{R}}(f)$. It is known as an Empirical Risk Minimization (ERM) problem.

Curse of Dimensionality

Curse of Dimensionality (CoD): The number of samples needed in order to successfully learn a target function (sample complexity) grows exponentially with the dimension of the samples.

Formally, let n be the number of samples and d is the dimension of the samples; $\mathcal{R}$ be the risk function, $\hat{\mathcal{R}}$ be the empirical risk function of n samples. Let $\mathcal{H}$ be the unit ball in the Lipschitz space, $\hat{f} = \text{argmin}_{f \in \mathcal{H}} \hat{\mathcal{R}}(f)$. Then[1]

$$|\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})| \leq \sup_{f \in \mathcal{H}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)| \sim \frac{1}{n^{1/d}}$$

Let $\epsilon = |\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})|$ be the error between $\mathcal{R}$ and $\hat{\mathcal{R}}$, then $(\frac{1}{\epsilon})^d \sim n$. This shows for a fixed ϵ (suppose ϵ is very small and $\frac{1}{\epsilon} \gg 1$), n grows exponentially with d.

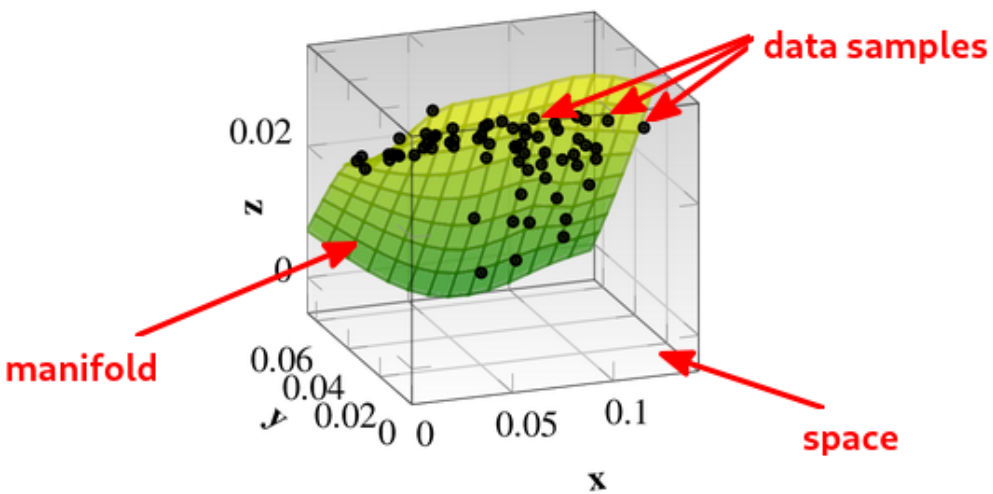
[1] Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu. Towards a Mathematical Understanding of Neural Network-Based Machine Learning: what we know and what we don't. arXiv preprint arXiv:2009.10713 (2020).

Why Deep Learning does not suffer from CoD

By the CoD, the number of samples should grow exponentially with the dimension of the samples. But in real world applications, neural networks can learn from small amounts of data (e.g., thousands of images). What is the reason behind it?

Manifold Hypothesis: The high-dimensional data are not truly high-dimensional. They can be on a low-dimensional manifold embedded in a high-dimensional space.

Thus the intrinsic dimension (the dimension of the manifold) of the data is low.



In the left figure, although the data samples live in a high-dimensional space (3D), they are from a low-dimensional manifold (2D) in the space. The shape of the manifold may be irregular.

Intrinsic Dimension of Image Data

The image data are supposed to be in low dimensional manifold because randomly sampled pixels cannot form meaningful images.

- If the image has the same intrinsic dimension as the space, then the manifold will have the same dimension as the space, which will make the pixels random.

For example, in ImageNet dataset, each image has $224 \times 224 \times 3 = 150528$ pixels, but the intrinsic dimension of it is estimated between only 26 and 43 [2].

[2] Phillip Pope, Chen Zhu, Ahmed Abdelkader et al, The Intrinsic Dimension of Images and Its Impact on Learning, ICLR 2021

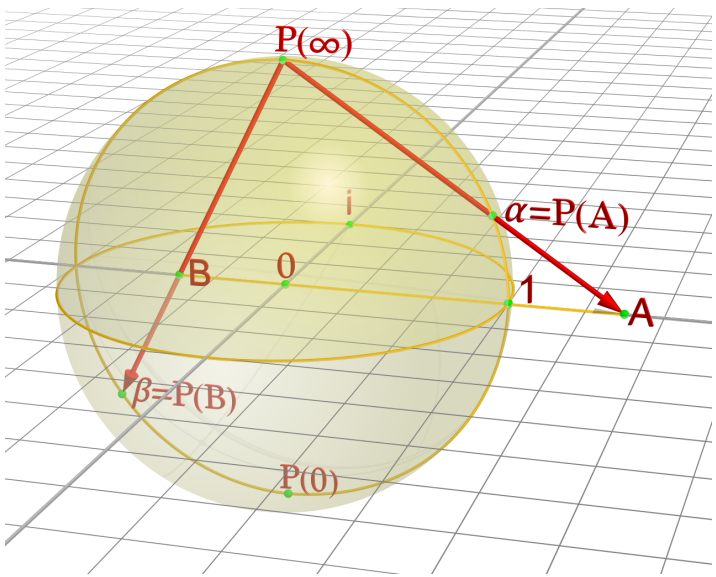
Contents

Manifold

Definition: A topological space X is called locally Euclidean if there is an non-negative integer n such that every point in X has a neighborhood which is homeomorphic to real n-space $\mathbb{R}^n$.

Definition: A topological manifold is a locally Euclidean Hausdorff space.

For example, we can "wrap" the 2D Euclidean space into a sphere. The sphere is a 2D manifold, although it is in the 3D space.



The left picture shows the sphere is homeomorphic to the 2D space. We can construct a one-to-one projection from the North pole of a sphere to its equatorial plane. [a]

[a] https://en.wikipedia.org/wiki/Stereographic_projection

Measure the Dimension of Manifold

The Manifold Hypothesis suggests that the data are on a low-dimensional manifold embedded in a high-dimensional space.

- The dimension of the manifold may be greater than 3, which makes it hard to visualize.
- The shape of the manifold may be irregular, not as smooth as a plane or a sphere.

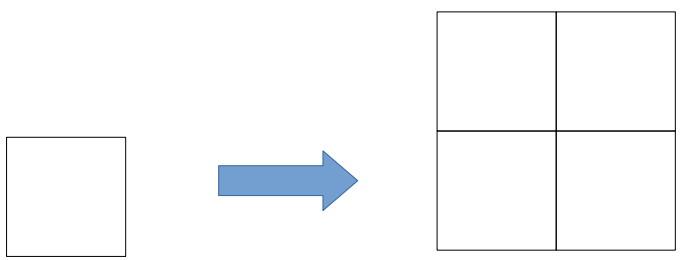
Question: How do we measure the dimension of the manifold formed by data?

- Before answering the problem, let's introduce a general definition of "the dimension of an object" – Fractal Dimension.
- Then we will introduce methods to measure fractal dimension including box-counting, correlation dimension and maximum likelihood estimation.

Fractal Dimension

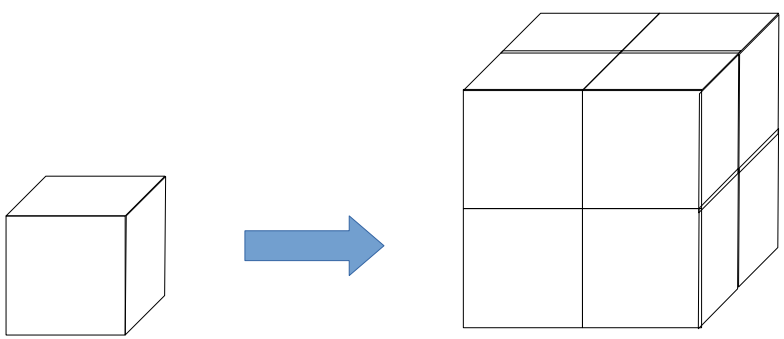
Dimension Definition by Coordinates: The dimension of an object is the minimum number of coordinates needed to specify any point within it. (The dimension is always an integer.)

Dimension Definition by the Rule of Scale: If the length of an object scales by ϵ, the mass of the object will scale by N, then the object has dimension $D = \log_\epsilon N$.



Scale Factor:
Length: 2
Mass (area): 4

Dimension:
$D = \log_2 4 = 2$

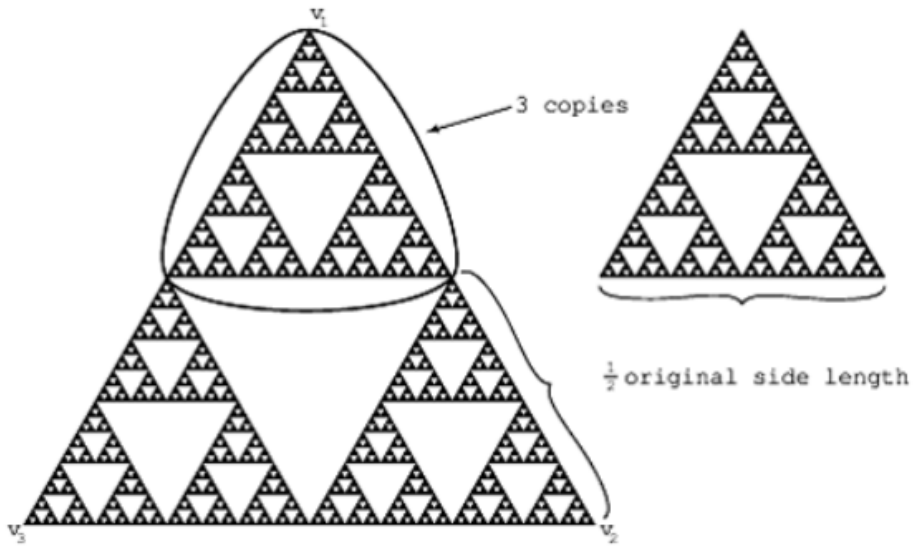


Scale Factor:
Length: 2
Mass (volume): 8

Dimension:
$D = \log_2 8 = 3$

Fractal Dimension

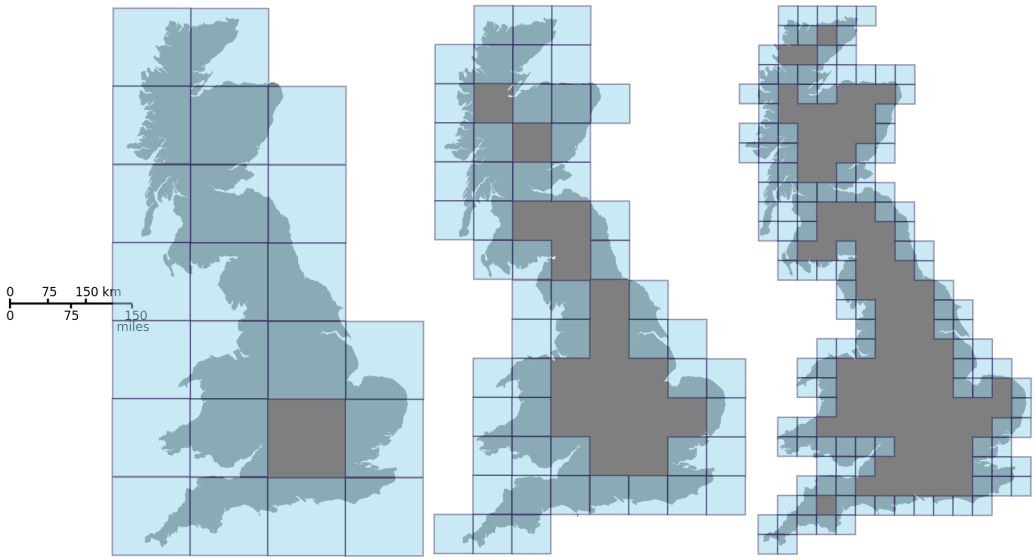
By the Definition of the Rule of Scale, some object will have non-integer dimension. For example, Sierpinski Triangle:



When the length of the Sierpinski Triangle scales by $1/2$, the mass of it scales by $1/3$, thus it has fractal dimension $D = \frac{\log 1/3}{\log 1/2} \approx 1.585$.

Compute Fractal Dimension by Box-Counting

Estimating the fractal dimension of the coast of Great Britain[3]:



Let ϵ be the scaling factor, $N(\epsilon)$ be the number of boxes to cover the coast. Then the dimension of the coast D has the equation that $c\,\epsilon^D \approx N(\epsilon)$, where c is an constant. We compute D by letting $\epsilon \to 0$. It is estimated that $D \approx 1.21$.

[3] https://en.wikipedia.org/wiki/Minkowski–Bouligand_dimension

Hausdorff Dimension

The fractal dimension induced by box-counting is called Minkowski–Bouligand dimension. Hausdorff dimension is a successor of Minkowski–Bouligand dimension, which is usually equivalent to the former (not always) but its definition is more complicated.

Definition (Hausdorff Measure): Let (X, ρ) be a metric space. For any subset $U \subset X$, let diam U denote its diameter:

$$\text{diam } U := \sup\{\rho(x, y) : x, y \in U\}$$

Let S be any subset of X, and $\delta > 0$ a real number. Define

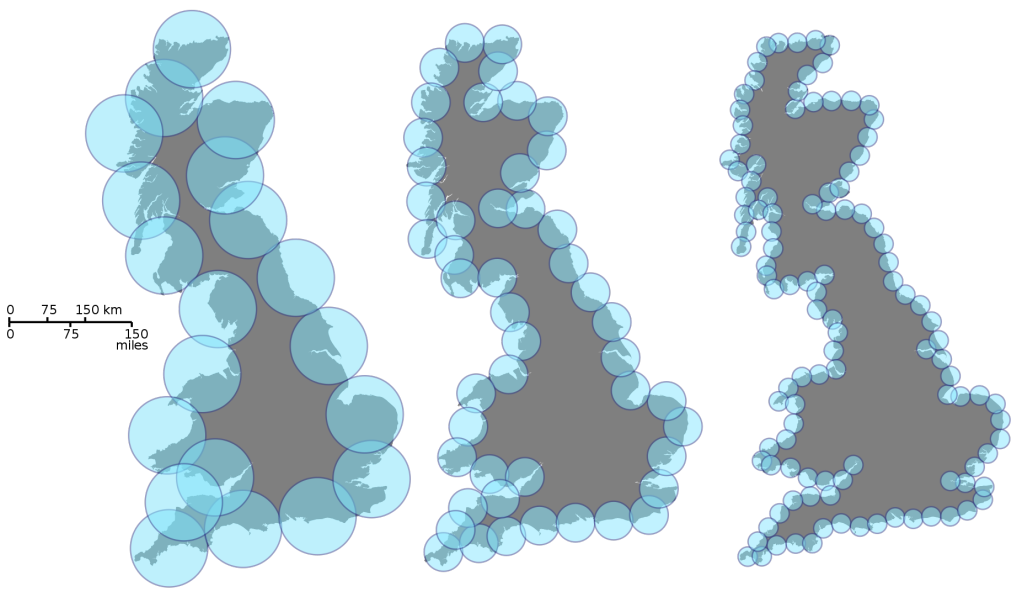
$$H^d_\delta(S) = \inf\{\sum_{i=1}^{\infty}(\text{diam } U_i)^d : \bigcup_{i=1}^{\infty} U_i \supseteq S, \text{diam } U_i < \delta\}$$

Let $\delta \to 0$, then $H^d(S) = \lim_{\delta \to 0} H^d_\delta(S)$ is called the d-dimensional Hausdorff measure of S.

Hausdorff Dimension

An example of Hausdorff Measure of the coast of Great Britain:

- Consider the part of the coast in each open ball is a subset U_i, the diameter of the open ball is diam U_i.
- Increasing the number of subsets but let each diam $U_i \to 0$, compute the limit $H^d(\text{coast})$ for any given d.



Hausdorff Dimension

$H^d(S)$ has an interesting property:

Theorem: For any S, there exists a $d_0 \in [0, \infty)$ to make $H^{d_0}(S) \in [0, \infty)$, such that: (1) for any $t < d_0$, $H^t(S) = \infty$; (2) for any $t > d_0$, $H^t(S) = 0$.

Proof: For (2), suppose $d_0 \in [0, \infty)$, by definition

$$H_\delta^{d_0}(S) = \inf\{\sum_{i=1}^{\infty}(\text{diam } U_i)^{d_0} : \bigcup_{i=1}^{\infty} U_i \supseteq S, \text{diam } U_i < \delta\}$$

For any $t > d_0$,
$H_\delta^t(S) = \inf\{\sum_{i=1}^{\infty}(\text{diam } U_i)^t\} = \inf\{\sum_{i=1}^{\infty}(\text{diam } U_i)^{d_0}(\text{diam } U_i)^{t-d_0}\} <$
$\inf\{\sum_{i=1}^{\infty}(\text{diam } U_i)^{d_0}\delta^{t-d_0}\} = \delta^{t-d_0}\inf\{\sum_{i=1}^{\infty}(\text{diam } U_i)^{d_0}\}$. Since $t - d_0 > 0$, when $\delta \to 0$, $\delta^{t-d_0} \to 0$, thus $H^t(S) = 0$. For (1), $t - d_0 < 0$, when $\delta \to 0$, $\delta^{t-d_0} \to \infty$, thus $H^t(S) = \infty$.

We call d_0 the Hausdorff Dimension of S, denoted as $\dim_{\mathsf{H}}(S)$. Note that d_0 can be either an integer or a fraction. Formally, $\dim_{\mathsf{H}}(S) = \inf\{d : H^d(S) = 0\}$.

Hausdorff Dimension

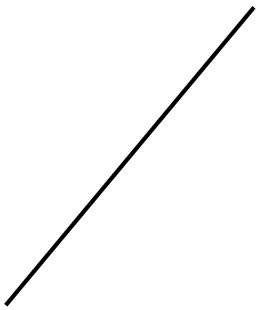
Theorem: Any object in an n-dimensional space cannot have Hausdorff dimension greater than n.

Proof: We first prove that if $X \subseteq S$, then $H^d(X) \leq H^d(S)$. This is because since X is a subset of S, the subset U_is that can cover X may not cover S. Thus, $\inf\{d : H^d(X) = 0\} \leq \inf\{d : H^d(S) = 0\}$, and $\dim_{\mathsf{H}}(X) \leq \dim_{\mathsf{H}}(S)$.

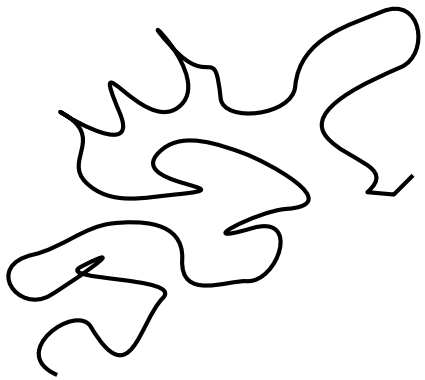
Let S be the n-dimensional space itself, which has Hausdorff dimension n. Any object in the n-dimensional space itself is a subset of S, thus have Hausdorff dimension no greater than n.

Hausdorff Dimension

Hausdorff Dimension measures the roughness of an object.



Smooth 1D manifold,
Hausdorff Dim = 1



Rough 1D manifold,
Hausdorff Dim between
1 and 2

Smooth 2D manifold
Hausdorff Dim = 2

Objects from nature are mostly fractal. If an object is not fractal, it is most likely man-made.

Contents

Methods of Estimating Intrinsic Dimension

The methods of estimating intrinsic dimension can roughly be classified into two categories [4]:

(1) Based on eigenvalue or projection: Project all data into a lower dimension.

- For example, PCA.

(2) Based on geometry: Using the neighborhood information to avoid projecting all data into a lower dimension.

- For example, Isomap, correlation dimension, maximum likelihood estimation.

[4] Francesco Camastra, Alessandro Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. IEEE Transactions on pattern analysis and machine intelligence. 2002

Correlation Dimension

In Box-Counting method, counting the boxes is impractical for high-dimensional objects.

Definition[5]: Given a set $S = \{x_1, x_2, ..., x_N\}$, and $C(l)$ be the correlation integral, where

$$C(l) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I\{\|x_i - x_j\| < l\}$$

when l is small, $C(l)$ grows like a power $C(l) \sim l^v$, where v is the correlation dimension of S.

In the above equation, $I\{\|x_i - x_j\| < l\}$ is the identity function where $I = 1$ if $\|x_i - x_j\| < l$ or 0 otherwise. $\sum_{i=1}^{N} \sum_{j=i+1}^{N} I\{\|x_i - x_j\| < l\}$ is the number of pairs of x_i and x_j that have distance less than l. There are $\binom{N}{2} = \frac{N(N-1)}{2}$ pairs in total, and the ratio of pairs that have distance less than l is $C'(l) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I\{\|x_i - x_j\| < l\}$. When $N \to \infty$, we have $C'(l) = 2C(l)$, thus $C'(l) \sim l^v$.

[5] Peter Grassberger, Procaccia Itamar. Measuring the strangeness of strange attractors. Physica. 1983

Correlation Dimension

We can prove that correlation dimension is a lower bound of Hausdorff dimension.

Theorem: Let v be the correlation dimension, D be the Hausdorff dimension of a dataset, then $v \leq D$.

Proof: For Hausdorff dimension, let $M(l)$ be the number of cells, where l is the diameter of each cell, then $M(l) \sim (\frac{1}{l})^D \Rightarrow \frac{1}{M(l)} \sim l^D$ when $l \to 0$.

For $C(l)$, suppose the points are uniformly distributed in each cell, and the pairs in the same cell have distance $< l$. Let N be the number of points, then each cell will have $\frac{N}{M(l)}$ points and $\frac{N}{M(l)}(\frac{N}{M(l)} - 1) \leq (\frac{N}{M(l)})^2$ pairs. Since there are $M(l)$ cells, the total number of pairs is bounded by $M(l)(\frac{N}{M(l)})^2 = \frac{N^2}{M(l)}$. Note that we do not consider the pairs across two cells having distance $< l$ because the cells can be joint by definition of Hausdorff dimension. Therefore,

$$l^v \sim C(l) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I\{\|x_i - x_j\| < l\} \leq \lim_{N\to\infty} \frac{1}{N^2} \frac{N^2}{M(l)} = \frac{1}{M(l)} \sim l^D$$

Thus $v \leq D$.

Maximum Likelihood Estimation

Correlation dimension implicitely uses Nearest Neighbor (NN) distances. And NN distance based methods tend to underestimate the intrinsic dimension for high-dimensional datasets.

Definition[6]: Let $x \in \mathbb{R}^n$ be a fixed point, $T_k(x)$ be the distance of the kth nearest neighbor of x to x. Then the intrinsic dimension with respect to x can be estimated as

$$\hat{m}_k(x) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1}$$

The global estimate can be obtained by taking the average of the local estimate at each point, i.e., $\bar{m}_k = \frac{1}{N} \sum_{i=1}^{N} \hat{m}_k(x_i)$.

[6] Elizaveta Levina, Peter J. Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. NIPS. 2004

Maximum Likelihood Estimation

We will explain how $\hat{m}_k(x)$ comes.

Suppose $\boldsymbol{x} = \{x_1, ..., x_N\}$ are samples from a distribution of PDF $f(\boldsymbol{x}|\boldsymbol{\theta})$, where the parameter $\boldsymbol{\theta}$ is unknown. We want to find $\boldsymbol{\theta}$ using $\boldsymbol{x}$ to determine the distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$ that the samples most likely come from.

To do this, we first compute the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$, then solve $\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta}|\boldsymbol{x})$ by computing $\frac{\partial L}{\partial \boldsymbol{\theta}} = 0$.

To facilitate computation, sometimes we compute the log-likelihood function $\log L(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$ instead, and solve $\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\boldsymbol{x})$ by computing $\frac{\partial \log L}{\partial \boldsymbol{\theta}} = 0$. Note that the $\boldsymbol{\theta}$ solved from $\log L(\boldsymbol{\theta}|\boldsymbol{x})$ is the same as the one from $L(\boldsymbol{\theta}|\boldsymbol{x})$.

Maximum Likelihood Estimation

Poisson Process: A Poisson Process is a model for a series of discrete event where the average time between events is known, but the exact timing of events is random.

- Events are independent of each other.
- The average rate (events per time period) is constant.
- Two events cannot occur at the same time.

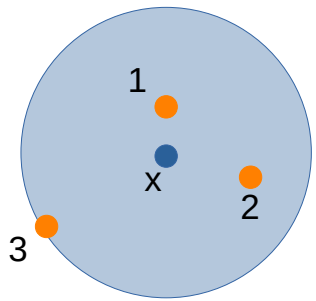
Let $N(t)$ be the number of events happened in t time periods, λ be the average number of events happened in each time period, then $N(t)$ satisfies Poisson distribution: $P(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$. When t is fixed, the PDF can be simplified as $P(N = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.

The likelihood function for Poisson distribution is: $L(\lambda|\boldsymbol{x}) = (\prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}) e^{-n\lambda}$. This is for synchronous data (all events have the same λ).

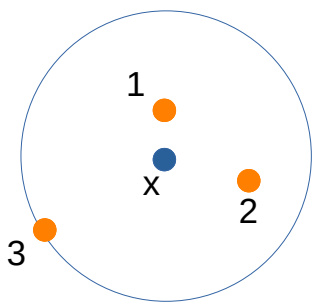
For asynchronous data (the events come one at a time on the timeline, each event have different λ, depending on the time t it happens), the likelihood function becomes $L(\lambda|\boldsymbol{x}) = \prod_{i=1}^{N(T)} \lambda(t_i) \exp\{-\int_0^T \lambda(t)dt\}$. $N(T)$ is the number of events happened in time T, t_i is the time that event i happens, $\lambda(t)$ is the λ at time t.

Maximum Likelihood Estimation

Now we consider the number of random points that occurs with the proximity to x (Here the proximity is analogous to the timeline).



The probability of the first k points is proportional to the volume of the ball of radius $T_k(x)$



The probability of the k-th points is proportional to the volume of the sphere of the ball of radius $T_k(x)$

Let $x \in \mathbb{R}^m$ be a fixed point, the n neighbors of x: $X_1, X_2, ..., X_n \in \mathbb{R}^m$ be iid from a density distribution $f(x)$, $T_k(x)$ be the distance of x to its kth nearest neighbor. Then we have the following relationship[a]:

$$\frac{k}{n} \approx f(x)V(m)[T_k(x)]^m$$

Here $V(m) = \frac{\pi^{m/2}}{\Gamma(m/2+1)}$ is the volume of an unit m-ball. $V(m)[T_k(x)]^m$ is the volume of a m-ball with radius $T_k(x)$.

Let t be the distance to x. Then we can set λ with respect to $t(0 \leq t \leq R)$ as follows:

$$\lambda(t) = f(x)V(m)mt^{m-1}$$

where $V(m)mt^{m-1} = \frac{d}{dt}V(m)t^m$ is the volume of the sphere of the m-ball.

---

[a] Richard O. Duda, Peter E. Hart and David G. Stork. Pattern classification and scene analysis. Wiley, 2000. (Section 4.2: Density Estimation)

Maximum Likelihood Estimation

Let $N(R)$ be the number of points we observed in the n-ball centered at x with radius R. t_i be the distance of x to its ith nearest neighbor. m be the intrinsic dimension of these points with respect to x and $m \leq n$. We set $\theta = \log f(x)$ be a constant, i.e., $\lambda(t) = e^{\theta} V(m) m t^{m-1}$.

Likelihood function of Poisson distribution for asynchronous data[7]:

$$L(\theta, m) = \prod_{i=1}^{N(R)} \lambda(t_i) \exp\{-\int_0^R \lambda(t) dt\}$$

Take the log-likelihood:

$$\begin{aligned} \log L(\theta, m) &= \sum_{i=1}^{N(R)} \log \lambda(t_i) - \int_0^R \lambda(t) dt \\ &= \sum_{i=1}^{N(R)} (\theta + \log(V(m)m) + (m-1) \log t_i) - e^{\theta} V(m) R^m \end{aligned}$$

[7] https://stats.stackexchange.com/questions/360814/mle-for-a-homogeneous-poisson-process

Maximum Likelihood Estimation

Solve the Maximum likelihood:

$$\frac{\partial \log L}{\partial \theta} = N(R) - e^{\theta} V(m) R^m = 0 \tag{1}$$

$$\frac{\partial \log L}{\partial m} = \sum_{i=1}^{N(R)} (\frac{V'(m)}{V(m)} + \frac{1}{m}) + \sum_{i=1}^{N(R)} \log t_i - e^{\theta} V(m) R^m (\frac{V'(m)}{V(m)} + \log R) = 0 \tag{2}$$

Substituting (1) into (2), we get

$$m = [\frac{1}{N(R)} \sum_{i=1}^{N(R)} \log \frac{R}{t_i}]^{-1}$$

Let $N(R)$ be the kth nearest neighbor, we have

$$m = [\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{t_k}{t_i}]^{-1}$$

We use $k-1$ here because of omitting the last zero term.

Contents

Basic Ideas of the Experiments

The problem is we do not know the intrinsic dimension of real world dataset since its manifold is quite complicated and hard to measure.

The steps of experiments in this paper[8] are as follows:
(1) Generate synthetic dataset with upper-bounded intrinsic dimensionality.

- Using GAN to generate synthetic images. Let the size of the image be fixed. Set most pixels to be zero but leave $\bar{d}$ pixels to be chosen at random. Then the synthetic dataset will have intrinsic dimension at most $\bar{d}$.

(2) Validate that MLE can estimate the intrinsic dimension of synthetic data accurately.

- The MLE used in this paper computes the global intrinsic dimension by inverse average $\bar{m}_k = [\frac{1}{N}\sum_{i=1}^{N} \hat{m}_k(x_i)^{-1}]^{-1}$, suggested by[9].

(3) Apply MLE to real world datasets.

[8] Phillip Pope, Chen Zhu, Ahmed Abdelkader et al, The Intrinsic Dimension of Images and Its Impact on Learning, ICLR 2021

[9] David J.C. MacKay and Zoubin Ghahramani. Comments on 'maximum likelihood estimation of intrinsic dimension' by e. levina and p. bickel (2004), 2005. http://www.inference.org.uk/mackay/dimension/

Intrinsic Dimension of Image Datasets

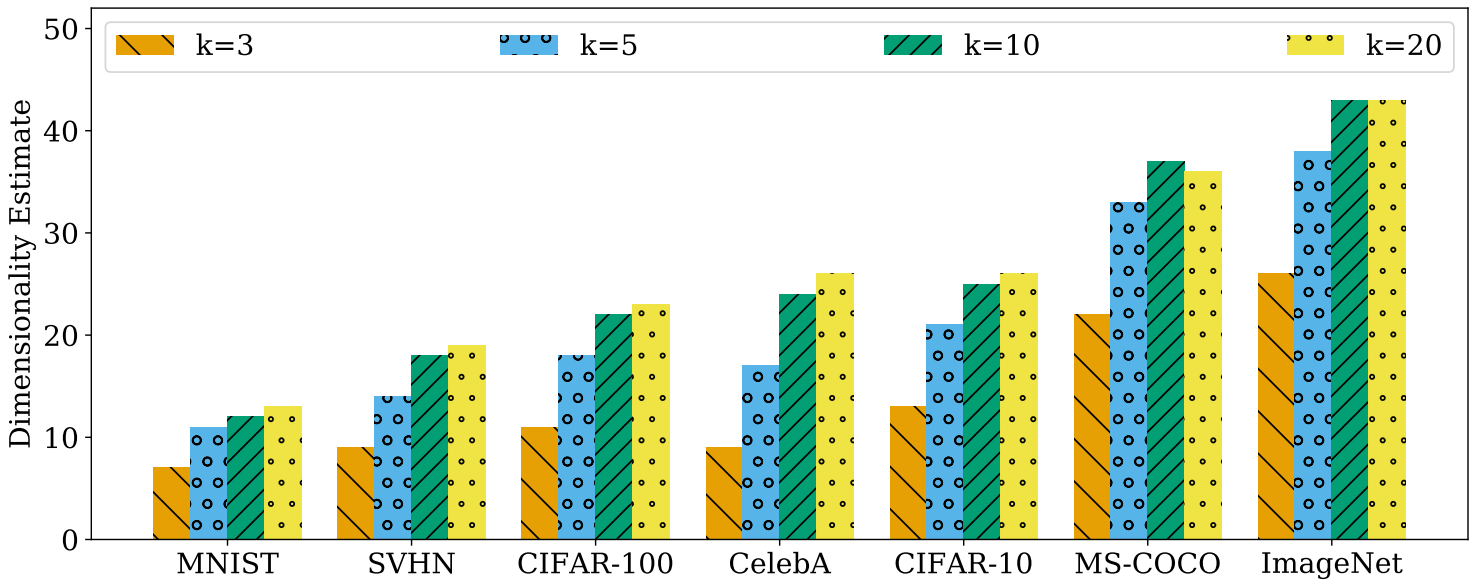


Figure 1: Estimates of the intrinsic dimension of commonly used datasets obtained using the MLE method with $k = 3, 5, 10, 20$ nearest neighbors (left to right). The trends are consistent using different k's.

Intrinsic Dimension and Testing Error

The following experiment results show that the intrinsic dimension is negatively correlated with test error.

Dataset	MNIST	SVHN	CIFAR-100	CelebA	CIFAR-10	MS-COCO	ImageNet
MLE (k=3)	7	9	11	9	13	22	26
MLE (k=5)	11	14	18	17	21	33	38
MLE (k=10)	12	18	22	24	25	37	43
MLE (k=20)	13	19	23	26	26	36	43
SOTA Accuracy	99.84	99.01	93.51	-	99.37	-	88.55

Table 1: The MLE estimates for practical image datasets, and the state-of-the-art test-set image classification accuracy (for classification problems only) for these datasets.

SOTA means State-Of-The-Art. k is the number of nearest neighbors used to estimate the intrinsic dimension.

From left to right, the intrinsic dimension of the dataset grows, but the SOTA Accuracy drops (except CIFAR-10).

Intrinsic Dimension and Sample Complexity

Verify two hypothesis:

- Data of lower intrinsic dimensionality has lower sample complexity than that of higher intrinsic dimensionality.
- Extrinsic dimensionality is irrelevant for sample complexity.

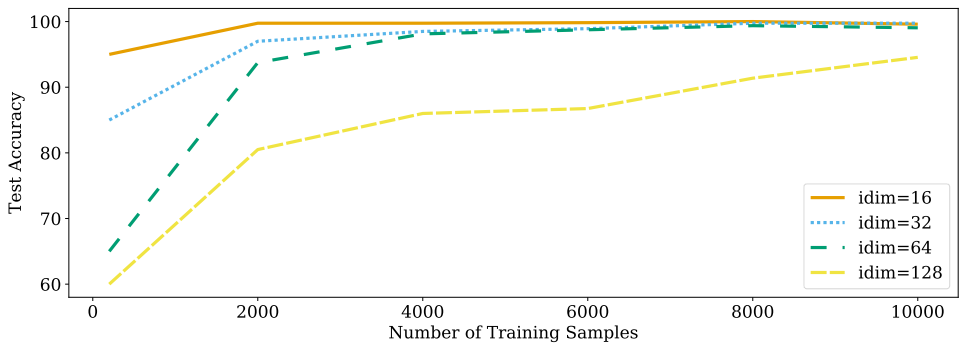


Figure 4: Sample complexity of synthetic datasets of varying intrinsic dimensionality.

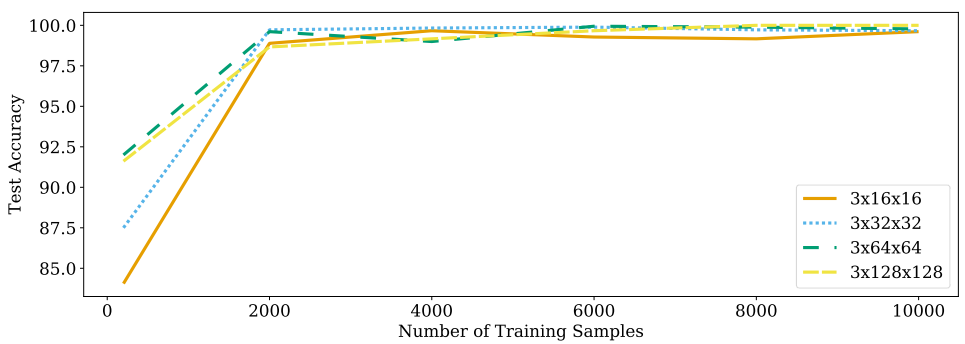


Figure 5: Sample complexity of synthetic datasets of varying extrinsic dimensionality.

The two figures shows the test accuracy (positively correlated to sample complexity) with the number of training samples. The first figure compares different intrinsic dimension while fixing extrinsic dimension. The second figure compares different extrinsic dimension while fixing intrinsic dimension.

The result shows intrinsic dimension affects more significantly on test accuracy because the gap is more obvious when intrinsic dimension increases.

Adding Noise to Images Increases Intrinsic Dimension

This experiment adds noise to CIFAR-10 dataset. The noise dim is the number of pixels randomly chosen to add noise. Each pixel is added a noise uniformly sampled in $[0, 1]$.

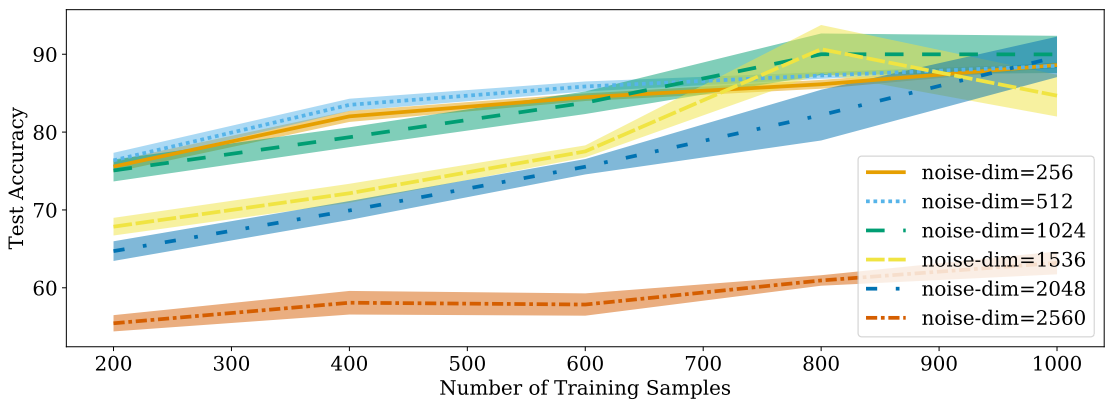


Figure 7: Sample complexity of noisy datasets. Standard errors are shown $N = 5$ random subsets of the data.

The result shows that with the noise dim increases, the test accuracy decreases, which means adding noise will increase the intrinsic dimension.

The results of different intrinsic dimension estimators

Intrinsic dimension estimation comparison of MLE with GeoMLE, TwoNN and kNN Graph Distance on real world datasets.

Dataset	MNIST	CIFAR-10	CIFAR-100	SVHN
MLE ($k = 5$)	11	21	18	14
GeoMLE ($k_1 = 20$,$k_2 = 55$)	25	96	93	21
TwoNN	15	11	9	7
kNN Graph Distance	7	7	8	6

Table 6: Additional ID estimators on popular datasets.