Data structure in machine learning: estimators and models

Vittorio Erba$^{(1,2)}$
Marco Cosentino Lagomarsino$^{(1)}$, Marco Gherardi$^{(1,2)}$, Mauro Pastore$^{(1,2)}$, Pietro Rotondo$^{(2)}$

EPFL, 15 December 2020

1. Università degli Studi di Milano
2. INFN, Sezione di Milano & FELLINI fellowship
Real data is geometrically structured

2d t-SNE projection of MNIST dataset

Understanding data structure requires many tools

1) UNDERSTAND STRUCTURE IN REAL DATASETS (Manifold learning)
   Clustering, Intrinsic dimension estimation, Dimensionality reduction, ...

2) HOW TO EMBED REAL DATASETS IN MATHEMATICALLY TRACTABLE SPACES
   Word embeddings, One-hot encodings, … +
   Euclidean VS Non-Euclidean metrics (Wasserstein, p-norms, ...)

3) THEORETICAL MODELS
   Perceptual manifolds, Teacher-student, Hidden manifold model

TODAY
PART 1: INTRINSIC DIMENSION ESTIMATION

PART 2: LINEAR CLASSIFICATION OF GEOMETRICALLY STRUCTURE DATA
Part 1: Intrinsic Dimension Estimation (IDE)

1) Define the problem
2) Overview of algorithms and issues
3) Glimpse of our novel estimator
IDE: retrieve the dimension $d$ of manifold from a discrete, random sample of $N$ points

Low $d$-dim manifold + Random sampling

Embedding

High $D$-dim representation

IDE algorithm

Intrinsic dimension $d$
IDE: retrieve the dimension $d$ of manifold from a discrete, random sample of $N$ points

**MANIFOLD HYPOTHESY**
Real datasets are random samples of smooth manifolds
MAY NOT BE TRUE + EMBEDDING DEPENDENT

**INTERNSIC DIMENSION**
Minimal number of degrees of freedom that encode all information of the dataset
There are two classes of IDE algorithms

\[
\rho(r) = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \theta(r - ||\vec{x}_i - \vec{x}_j||)
\]

**ASSUME LOCAL LINEARITY + MEASURE LOCAL DENSITY AT SMALL SCALE** \( r \)

**FIT AGAINST** \( r^d \)

---

There are two classes of IDE algorithms

**GEOMETRIC ESTIMATORS (Corr Dim)**

\[
\rho(r) = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \theta(r - ||\vec{x}_i - \vec{x}_j||)
\]

**PROJECTIVE ESTIMATORS (PCA)**

**ASSUME LOCAL LINEARITY**

MEASURE LOCAL DENSITY AT SMALL SCALE \( r \)

\( = \)

FIT AGAINST \( r^d \)

**ASSUME GLOBAL LINEARITY**

USE LINEAR ALGEBRA TO DISTINGUISH BETWEEN INTRINSIC AND SPURIOUS DIMENSIONS


High ID + Curvature are the main enemies of IDE algorithms

EXPONENTIAL UNDERSAMPLING IN HIGH DIMENSION \((d > 6) = \text{LOCAL LINEARITY DIFFICULT TO PROBE}\)

High ID + Curvature are the main enemies of IDE algorithms

EXPONENTIAL UNDERSAMPLING IN HIGH DIMENSION (d > 6) = LOCAL LINEARITY DIFFICULT TO PROBE

Curvature => Proliferation of possible geometries + No global linearity

A new estimator: Full Correlation Integral (FCI)
FCI leverages not-so-small $r$ regime to avoid undersampling

\[ \rho(r; d) = \frac{1}{2} + \frac{\Omega_{d-1}}{\Omega_d} (r^2 - 2) {}_{2}F_{1} \left( \frac{1}{2}, \frac{1 - \frac{d}{2}}{\frac{3}{2}} \left| (r^2 - 2)^2 \right. \right) \]

Well sampled!

Undersampled
FCI leverages not-so-small \( r \) regime to avoid undersampling

Linear manifolds
Isotropic sampling measure
Linear embeddings

\[
\rho(r; d) = \frac{1}{2} + \frac{\Omega_{d-1}}{\Omega_d} (r^2 - 2)^2 F_1 \left( \frac{1}{2}, 1 - \frac{d}{2}, \frac{3}{2} \left| (r^2 - 2)^2 \right. \right)
\]

Manifold dependent features do not alter the overall shape of the correlation integral

Able to estimate in the extreme undersampled regime \( N < d \)

(Geometric: \( \exp d \) | Projective: \( d \log d \))

Vertices of cube \( d = 15 \)

Gaussian \( d = 15 \)

Cube \( d = 15 \)
FCI can be easily multiscaled to extend to curved manifolds

Use the "most persistent" + lowest minimum as the estimator of the Intrinsic Dimension
FCI can be easily multiscaled to extend to curved manifolds

Use the "most persistent" + lowest minimum as the estimator of the Intrinsic Dimension

5 degrees of freedom per blob:
translation x, translation y, eccentricity, scale, tilt
Part 1: Intrinsic Dimension Estimation (IDE)

1) Accurate IDE is tricky
2) FCI + multiscalability => first step towards more robust IDE
3) A lot of work to do! Rationalize the multiscale approach into a statistically robust estimator


https://github.com/vittorioerba/pyFCI
Part 2: Linear classification of geometrically structured data

1) Define the polytope model
2) Review of expressivity of linear classifiers for unstructured data
3) Expressivity of linear classifiers for structured data
4) A data-driven phase transition
Geometric correlations => Classification of manifolds

Before: Linear classification of UNSTRUCTURED data
Geometric correlations => Classification of manifolds

Before:
Linear classification of UNSTRUCTURED data


Geometric correlations => Classification of manifolds

Before:
Linear classification of UNSTRUCTURED data

After:
Linear classification of STRUCTURED data

1) Expand points into manifolds
2) Restrict the learning architectures to those that classify coherently points in the same manifolds


Understand how linear classification properties are changed by data structure
The expressivity of linear classifiers can be computed using two complementary frameworks.

\[ P = \text{# of training samples} \]
\[ N = \text{# of dimensions} \]
\[ C(N, P) = \text{# of dichotomies realizable by a linear classifiers} \]
The expressivity of linear classifiers can be computed using two complementary frameworks.

**Combinatorial (Cover)**

- Exact results for full $C(N,P)$ curves

$$C_{n,p} = 2 \sum_{k=0}^{p-1} \binom{p-1}{k}$$

$$\alpha_c = 2$$

- Additional insights:
  - $C(N,P)$ independent on the position of the points
  - Finite size corrections

**Statistical Physics (Gardner)**

- Exact results for storage capacity

$$\int D^n W \prod_{\mu=1}^{p} \theta \left( \sigma^\mu \sum_{i=1}^{n} W_i s_i^\mu \right)$$

- Position of the points + labels = quenched disorder

---


Both frameworks extend to geometrically structured data

\[ P = \# \text{ of training samples} \]
\[ N = \# \text{ of dimensions} \]
\[ K = \# \text{ of vertices of polytopes} \]
\[ C(N, P, K) = \# \text{ of dichotomies of polytopes realizable by a linear classifiers} \]
Both frameworks extend to geometrically structured data

\[ P = \# \text{ of training samples} \]
\[ N = \# \text{ of dimensions} \]
\[ K = \# \text{ of vertices of polytopes} \]
\[ C(N,P,K) = \# \text{ of dichotomies of polytopes realizable by a linear classifiers} \]
Both frameworks extend to geometrically structured data

Combinatorial

\[ C_{n,p+1}^{(2)} = \Psi_2(\rho) C_{n,p}^{(2)} + C_{n-1,p}^{(2)} + \left[ 1 - \Psi_2(\rho) \right] C_{n-2,p}^{(2)} \]


Statistical Physics

\[ \int D^n W \prod_{\mu=1}^n \theta \left[ \sigma^\mu \sum_{i=1}^n W_i \left( \zeta^\mu_{\alpha} \right)_i \right] \]

\[ \alpha_c = \frac{2}{3 - 2 \Psi_2(\rho)} \]


P = # of training samples
N = # of dimensions
K = # of vertices of polytopes
\( C(N,P,K) = \) # of dichotomies of polytopes realizable by a linear classifiers
One more thing: a novel phase transition
One more thing: a novel phase transition

\[ (\alpha_* + 1) \log(\alpha_* + 1) - \alpha_* \log \alpha_* + \\
+ (\alpha_* - 1) \log \theta_k(0) + \log \theta_k(1) = 0 \]


One more thing: a novel phase transition

\[
\begin{align*}
(\alpha_* + 1) \log(\alpha_* + 1) - \alpha_* \log \alpha_* + \\
+ (\alpha_* - 1) \log \theta_k(0) + \log \theta_k(1) &= 0
\end{align*}
\]


Is a set of manifolds classifiable at all?

\[
\int D\sigma \int D^nW \prod_{\mu,\alpha=1}^{p, k} \theta \left[ \sigma^\mu \sum_{i=1}^{n} W_i(\xi^\mu_{\alpha})_i \right]
\]

Transition common to all models of extended geometry
Part 2: Linear classification of structured data (simplex model)

1) Polytope model is analytically tractable
2) Classification of extended datasets induces new phase transition
3) Check other kinds of data structure models?

https://journals.aps.org/prresearch/abstract/10.1103/PhysRevResearch.2.023169
https://journals.aps.org/pre/abstract/10.1103/PhysRevE.102.032119
Thank you!