Distance learning using Euclidean percolation: Following Fermat's principle

Matthieu Jonckheere

with P. Groisman (UBA) and F. Sapienza (Berkeley)

UBA and IMAS - CONICET

Invited Professor Centrale-Supelec and DataIA

Motivation

Our original motivation: A problem at Aristas SRL

Problem

- Clustering of high dimensional chemical formulas

Data size

- 10^6 formulas
- Dimension $d \sim 4000$

Clustering in high-dimensional spaces is usually very difficult

Our original motivation: A problem at Aristas SRL

Problem

- Clustering of high dimensional chemical formulas

Data size

- 10^6 formulas
- Dimension $d \sim 4000$

Clustering in high-dimensional spaces is usually very difficult and Euclidian or ad-hoc distances might be misleading...

A curse of dimensionality

Bad news

Let $\omega_D(r) = \omega_D(1) r^D$ be the volume of the ball of radius r in $\mathbb{R}^D$.

$$\frac{\omega_D(1) - \omega_D(1-\varepsilon)}{\omega_D(1)} = 1 - (1-\varepsilon)^D \xrightarrow{D \to \infty} 1$$

A curse of dimensionality

Bad news

Let $\omega_D(r) = \omega_D(1) r^D$ be the volume of the ball of radius r in $\mathbb{R}^D$.

$$\frac{\omega_D(1) - \omega_D(1-\varepsilon)}{\omega_D(1)} = 1 - (1-\varepsilon)^D \xrightarrow{D \to \infty} 1$$

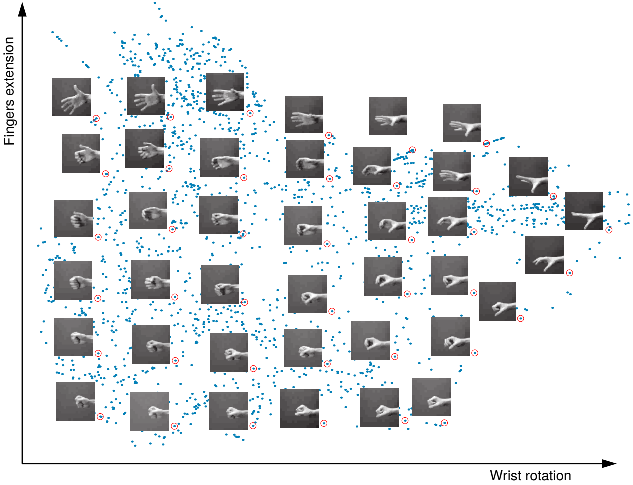
In high dimensional Euclidean spaces every two points of a typical large set are at similar distance.

Manifold hope

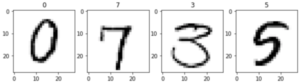
Good news: many structured data live in a manifold of dimension much lower than ambient space ($d \ll D$).

Manifold hope

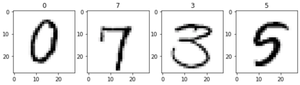
Good news: many structured data live in a manifold of dimension much lower than ambient space ($d \ll D$).

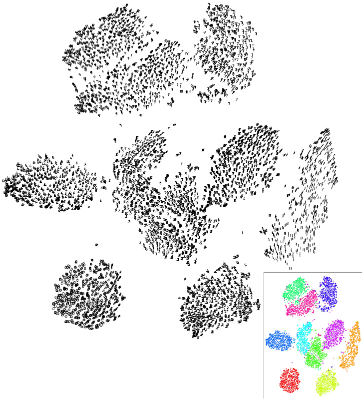


Motivation: MNIST Dataset



Motivation: MNIST Dataset





van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research. 9: 2579–2605.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.
- The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.
- The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.
- Since the data lies in an (unknown) lower dimensional surface, this distance has to be inferred from the data itself.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.
- The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.
- Since the data lies in an (unknown) lower dimensional surface, this distance has to be inferred from the data itself.
- Delicate game between dimensionality reduction, choice of the distance and clustering...

Dimension reduction and distance learning techniques

Dimensionality reduction and distance learning

There are many techniques to address dimensionality reduction and possibly finding distances in lower dimensional spaces:

- Principal components analysis (PCA),
- Multidimensional scaling (MDS),
- t-Stochastic neighbor embbeding (t-SNE),
- Isomap and variants.

Dimensionality reduction and distance learning

Dimensionality reduction

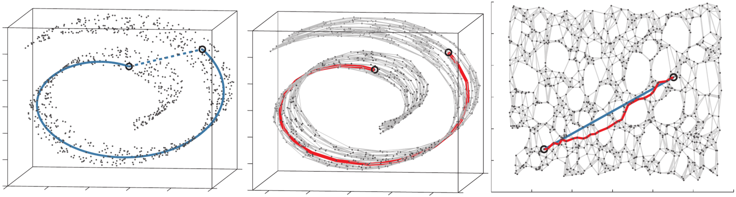
- Principal components analysis (PCA),
- Multidimensional scaling (MDS),
- t-Stochastic neighbor embbeding (t-SNE).

Distance learning

- Isomap and variants.

Dimensionality Reduction/distance learning: Isomap

Constructs the k-nn graph and finds the optimal path. The weight of an edge is given $|q_i - q_j|$.



Isomap

Theorem

Given $\varepsilon > 0$ *and* $\delta > 0$, *for* n *large enough*

$$\mathbb{P}\left(1 - \varepsilon \leq \frac{d_{\textsf{geodesic}}(x, y)}{d_{\textsf{graph}}(x, y)} \leq 1 + \varepsilon\right) > 1 - \delta.$$

[Bernstein, de Silva, Langford, Tenenbaum (2000)].

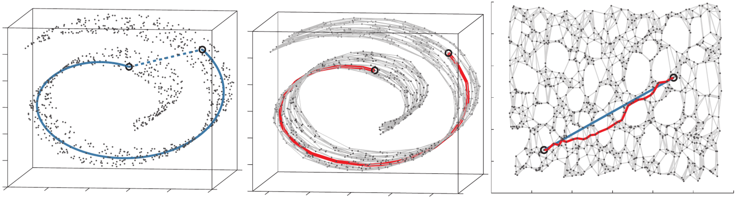
Isomap

Theorem

Given $\varepsilon > 0$ and $\delta > 0$, for n large enough

$$\mathbb{P}\left(1-\varepsilon \leq \frac{d_{\textit{geodesic}}(x,y)}{d_{\textit{graph}}(x,y)} \leq 1+\varepsilon\right) > 1-\delta.$$

[Bernstein, de Silva, Langford, Tenenbaum (2000)].



Fermat's distance

The Problem

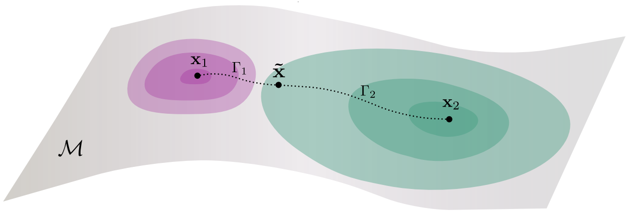
- Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).

The Problem

- Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).
- Consider n independent points on $\mathscr{M}$ with common density $f : \mathscr{M} \mapsto \mathbb{R}_{\geq 0}$.

The Problem

- Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).
- Consider n independent points on $\mathscr{M}$ with common density $f : \mathscr{M} \mapsto \mathbb{R}_{\geq 0}$.



Can we learn a better notion of distance between points (for say clustering)?

Objectives

We look for a distance that takes into account the underlying manifold $\mathscr{M}$ and the underlying density f.

Sample Fermat's distance

- $\alpha \geq 1$ a parameter, $\mathbb{X} =$ a discrete set of points q, $x, y \in \mathbb{X}$.

Sample Fermat's distance

- $\alpha \geq 1$ a parameter, $\mathbb{X} =$ a discrete set of points q, $x, y \in \mathbb{X}$.

$$\mathscr{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q}) = \inf\{ \sum_{j=1}^{K-1} |\mathbf{y}_{i+1} - \mathbf{y}_i|^{\alpha} \colon K \geq 2,$$

$$\text{y } (\mathbf{y}_1, \ldots, \mathbf{y}_K) \text{ is a } \mathbb{X}\text{-path from } \mathbf{p} \text{ to } \mathbf{q}\}.$$

Visualisation

▸ http://www.aristas.com.ar/fermat/index.html

Sample to Macroscopic Fermat's distance

Theorem (Groisman, Jonckheere, Sapienza, 2018+)
Under mild assumptions on f, there exists $\mu > 0$, such that for $x, y \in \mathscr{M}$ and $\mathbb{X}_n$ i.i.d $\sim f$ we have

$$\lim_{n \to \infty} n^{\beta} D_{\mathbb{X}_n}(x, y) = \mu \mathscr{D}(x, y),$$

almost surely, with $\beta = (\alpha - 1)/d$.

$$\mathscr{D}(x, y) = \inf_{\Gamma} \int_{\Gamma} \frac{1}{f^{\beta}}.$$

Fermat's principle

In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}, \quad \mathrm{n} = \text{refractive index}$$

Fermat's principle

In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}, \quad \mathrm{n} = \text{refractive index}$$

$$\mathscr{D}(x,y) = \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \qquad f^{-\beta} \sim \mathbf{n}$$

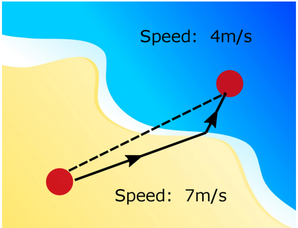
Fermat's principle

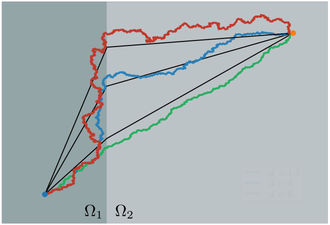
In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}, \quad \mathrm{n} = \text{refractive index}$$

$$\mathscr{D}(x,y) = \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \qquad f^{-\beta} \sim \mathbf{n}$$

Snell's law, the lifeguard and Fermat's distance





Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff$$

Heuristics:

$r = (q_1, \dots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha - 1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$n c_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

$$n^{(\alpha-1)/d} |q_{i+1} - q_i|^{\alpha-1} \asymp c \frac{1}{f(q_i)^{(\alpha-1)/d}}$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

$$n^{(\alpha-1)/d} |q_{i+1} - q_i|^{\alpha-1} \asymp c \frac{1}{f(q_i)^{(\alpha-1)/d}}$$

$$\inf_r n^{(\alpha-1)/d} \sum |q_{i+1} - q_i|^\alpha \asymp \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \, d\ell.$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

$$n^{(\alpha-1)/d} |q_{i+1} - q_i|^{\alpha-1} \asymp c \frac{1}{f(q_i)^{(\alpha-1)/d}}$$

$$\inf_r n^{(\alpha-1)/d} \sum |q_{i+1} - q_i|^\alpha \asymp \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \, d\ell. \qquad \square$$

Some mathematical insights

Homogeneous Poisson Point Process : Shape theorem

We based our analysis on:

Theorem (Howard and Newman (1997))

Let $\mathbb{X}$ *a PPP with intensity* $\lambda = 1$. *Then there exists* $0 < \mu < \infty$ *such that*

$$\lim_{|\mathbf{q}| \to \infty} \frac{\mathscr{D}_{\mathbb{X}}(\mathbf{0}, \mathbf{q})}{|\mathbf{q}|} = \mu, \qquad \textit{almost surely.}$$

Homogeneous Poisson Point Process : Shape theorem

We based our analysis on:

Theorem (Howard and Newman (1997))

Let $\mathbb{X}$ *a PPP with intensity* $\lambda = 1$. *Then there exists* $0 < \mu < \infty$ *such that*

$$\lim_{|\mathbf{q}| \to \infty} \frac{\mathscr{D}_{\mathbb{X}}(\mathbf{0}, \mathbf{q})}{|\mathbf{q}|} = \mu, \qquad \textit{almost surely.}$$

Also give bounds on the fluctuations!

Other previous mathematical results

Sung Jin Hwang, Steven B. Damelin, Alfred O. Hero III,

Shortest Path through Random Points,

The Annals of Applied Probability, 2016, Vol. 26, No. 5, pp 2791-2823.

Algorithmic considerations and generalizations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}}^{(\alpha,k)}(x,y) = \inf_{\substack{r=(q_1,\ldots,q_K) \\ q_{i+1} \in \mathcal{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1} - q_i|^{\alpha}.$$

Generalization of Isomap and Fermat's distance.

Algorithmic considerations and generalizations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}}^{(\alpha,k)}(x,y) = \inf_{\substack{r=(q_1,\ldots,q_K)\\ q_{i+1}\in\mathscr{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1}-q_i|^{\alpha}.$$

Generalization of Isomap and Fermat's distance.

Proposition [Groisman, Jonckheere, Sapienza, 2018+]: *Given $\varepsilon > 0$, we can choose $k = \mathscr{O}(\log(n/\varepsilon))$ such that*

$$\mathbb{P}\left(D_{\mathbb{X}_n}^{(k)}(x,y) = D_{\mathbb{X}_n}(x,y)\right) > 1-\varepsilon.$$

Algorithmic considerations and generalizations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}}^{(\alpha,k)}(x,y) = \inf_{\substack{r=(q_1,\ldots,q_K) \\ q_{i+1}\in\mathcal{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1}-q_i|^{\alpha}.$$

Generalization of Isomap and Fermat's distance.

Proposition [Groisman, Jonckheere, Sapienza, 2018+]: *Given $\varepsilon > 0$, we can choose $k = \mathcal{O}(\log(n/\varepsilon))$ such that*

$$\mathbb{P}\left(D_{\mathbb{X}_n}^{(k)}(x,y) = D_{\mathbb{X}_n}(x,y)\right) > 1-\varepsilon.$$

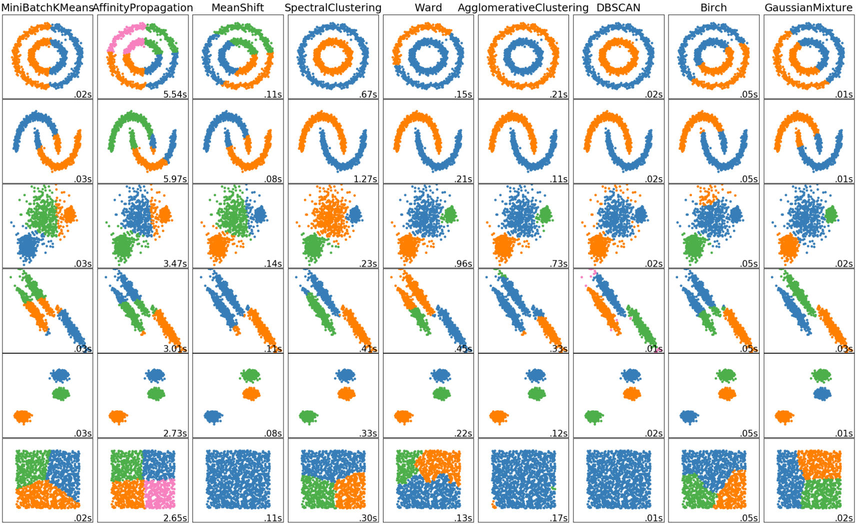
$\rightarrow$ We can reduce the running time from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2(\log n)^2)$.

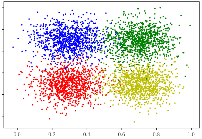
Open theoretical questions

- General proof of convergence for k fixed?
- How to choose α, k ??

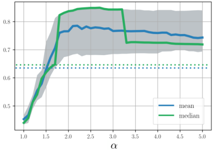
Clustering

Clustering

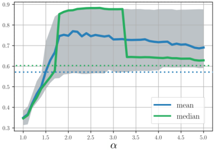

MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.02s
5.54s
.11s
.67s
.15s
.21s
.02s
.05s
.01s
.03s
5.97s
.08s
1.27s
.21s
.11s
.02s
.05s
.01s
.03s
3.47s
.14s
.23s
.96s
.73s
.02s
.05s
.02s
.03s
3.01s
.11s
.41s
.45s
.33s
.01s
.05s
.03s
.03s
2.73s
.08s
.33s
.22s
.12s
.02s
.05s
.01s
.02s
2.65s
.11s
.30s
.13s
.17s
.01s
.05s
.02s



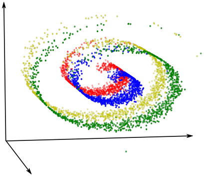
(a) 2D data



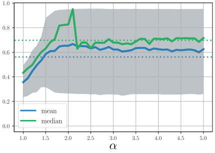
(c) Adjusted mutual information



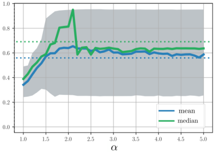
(e) Adjusted Rand index



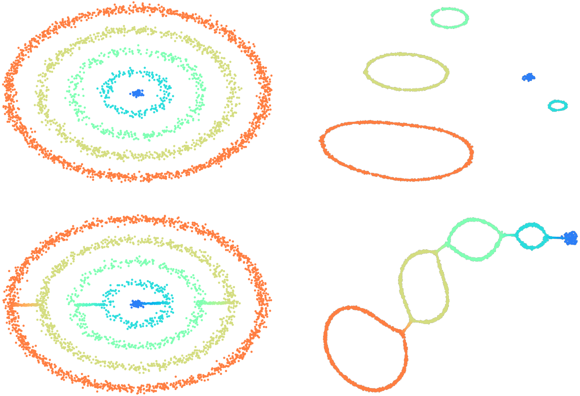
(b) 3D data



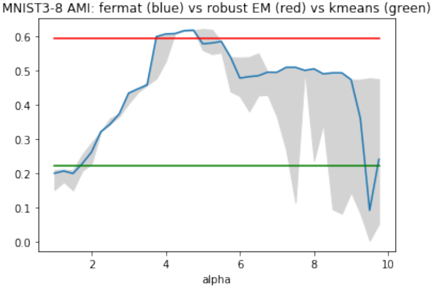
(d) Accuracy



(f) F1 score



MNIST



Performance of Fermat + k-medoids compared to state of the art robust clustering

Simulations Violeta Roizman and Alfredo Umfurer.

Application in genetics

Fingerprints of cancer by persistent homology,

A. Carpio, L. L. Bonilla, J. C. Mathews, A. R. Tannenbaum,

2019.

- They compute Fermat's distance between genes'expressions (dimension 77) (They choose $\alpha \sim 3$.)
- They study clusters based on the Fermat distance.
- "These clusters make noticeable the relations between gene expressions in healthy samples and those in cancerous samples."

Conclusions

- We have introduced Fermat's distance and way to estimate it with a sample.

Conclusions

- We have introduced Fermat's distance and way to estimate it with a sample.
- It defines a notion of distance between sample points that takes into account the geometry of the clouds of point, including possible non-homogeneities.

Conclusions

- We have introduced Fermat's distance and way to estimate it with a sample.
- It defines a notion of distance between sample points that takes into account the geometry of the clouds of point, including possible non-homogeneities.
- We have proved that this estimator in fact approximates Fermat's distance, which is a good way to measure distance in this (general) setting.

Applications

- Clustering

Applications

- Clustering
- Dimensionality reduction

Applications

- Clustering
- Dimensionality reduction
- Density estimation

Applications

- Clustering
- Dimensionality reduction
- Density estimation
- Regression

Applications

- Clustering
- Dimensionality reduction
- Density estimation
- Regression
- Any learning task that requires a notion of distance (not necessarily in Euclidian space) as an input.

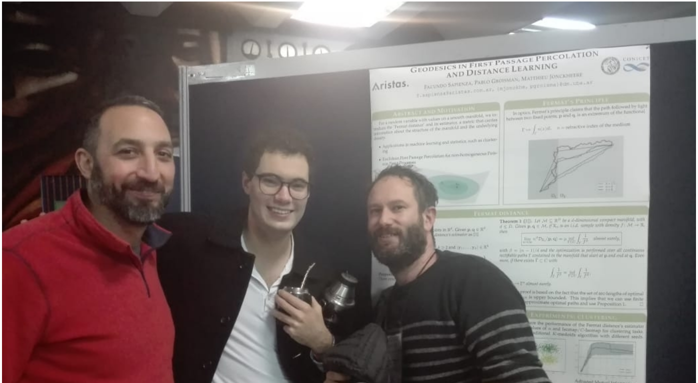
Download

A prototype implementation is available at

▸ http://www.aristas.com.ar/fermat/index.html

- *Weighted Geodesic Distance Following Fermat's Principle* (2018); F. Sapienza, P. Groisman, M. Jonckheere; 6th International Conference on Learning Representations (ICRL 2018).

- *Geodesics in First Passage Percolation and Distance Learning* (2019); P. Groisman, M. Jonckheere, F. Sapienza; submitted



Thanks!