Grouping Countries and Regions to Improve Covid-19 Dynamics Predictions

UFRN Covid-19 course Fall 2020

presented by **Joris Guérin**

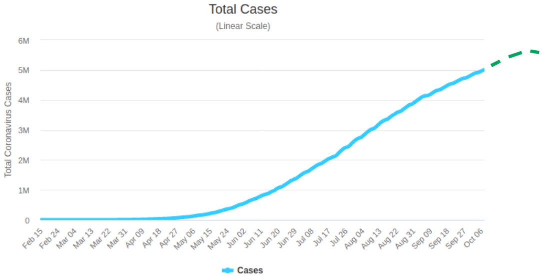
on October 13 2020

Table of content

Outline

Real Objective

Predicting future contaminations for a certain region

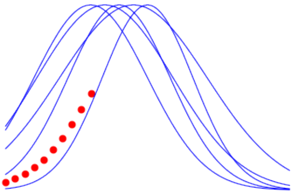


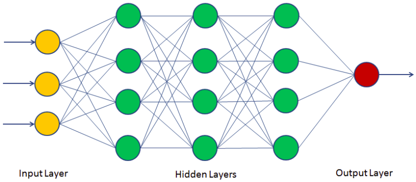
Why?

- Help local health authorities to manage hospital beds.
- Help local politics to take appropriate actions to protect population.

Data driven approaches

Train a Machine Learning model on previous data



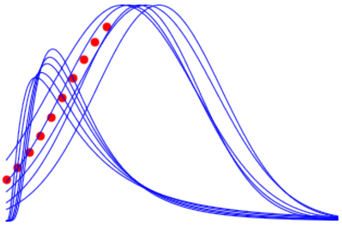


Advantages?

- No need to estimate contamination parameters (Difficult).
- Get better as we get more data.

Difficulty with Learning to Predict Contamination

Few data and different response form

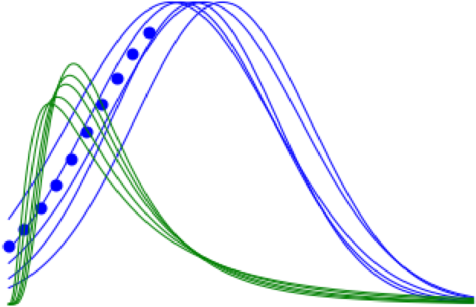


Sources of differences between countries/regions

- Policy to prevent spread of the disease (border closing, lockdown)
- Testing policy (Percentage of population, already sick people only)
- reporting of results (seasonality)

Solution?

Identify countries with similar responses in advance

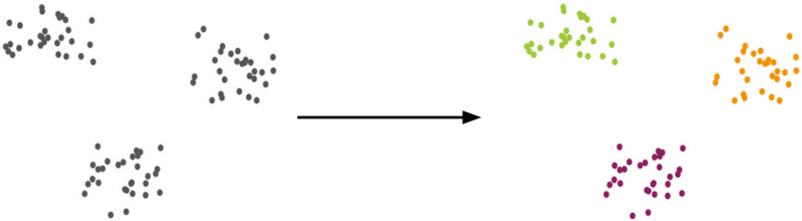


⇒ Clustering countries before learning

Outline

Problem definition

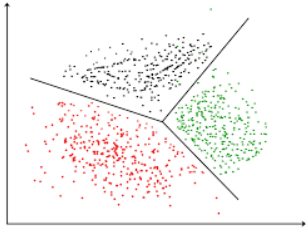
Grouping point in an unsupervised manner



- High intra-cluster similarity
- Low between-cluster similarity

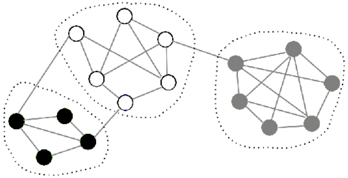
Taxonomy of Clustering Algorithms

Partitioning-based



- K-means, Fuzzy C-means, DEC, ...
- Easy integration of new points

Connectivity-based



- Hierarchical, Affinity Propagation, JULE, ...
- Only requires distances between data points

K-means clustering

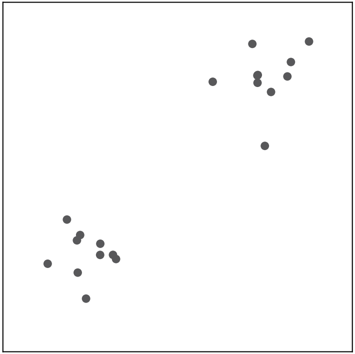
Formulation for M data points and K desired clusters.

$$
\begin{aligned}
\underset{A,\ c}{\text{Minimize}} \quad & \sum_{i=1}^{M}\sum_{k=1}^{K} a_{ik} \times d(x_i, c_k), \\
\text{subject to} \quad & \sum_{k=1}^{K} a_{ik} = 1,\ \forall i \in \{1, ..., M\}, \\
& a_{ik} \in \{0,\ 1\},\ \forall i,\ \forall k.
\end{aligned}
$$

With

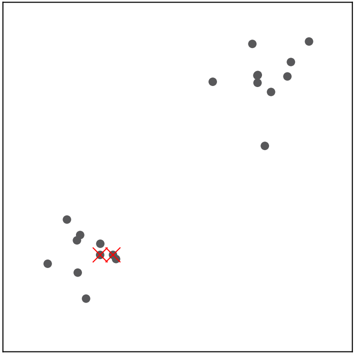
- c_k cluster centers,
- a_{ik} membership binary variables,
- $d(.,.)$ distance metric used.

K-means resolution using Alternating Optimization



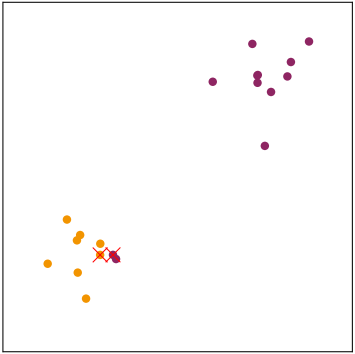
Initial data

K-means resolution using Alternating Optimization



Centroids initialization

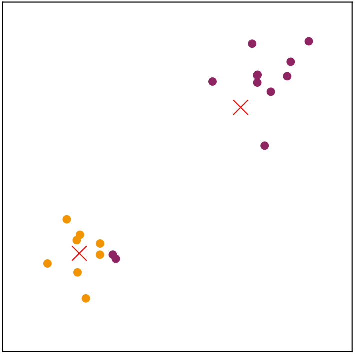
K-means resolution using Alternating Optimization



Classes actualization

$$x_i \in C_l \iff d(x_i, c_l) \leq d(x_i, c_k),\ \forall k \in \{1, ..., K\}.$$

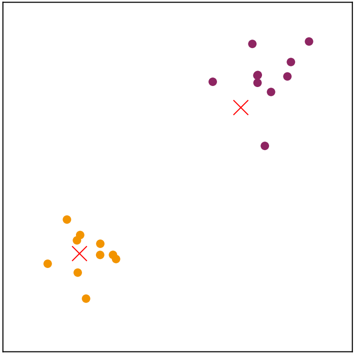
K-means resolution using Alternating Optimization



Centroids update

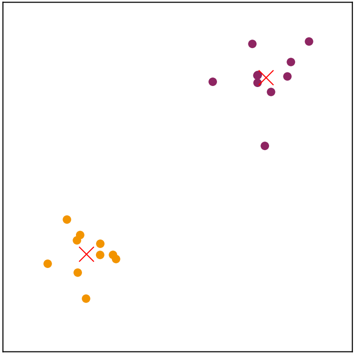
$$c_k = \frac{1}{\sum_{i=1}^{M} a_{ik}} \sum_{i=1}^{M} a_{ik} \times x_i$$

K-means resolution using Alternating Optimization



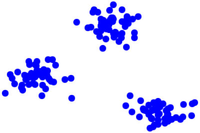
Classes actualization

K-means resolution using Alternating Optimization



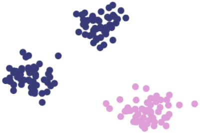
Centroids update

Number of clusters?

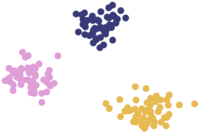


KMeans

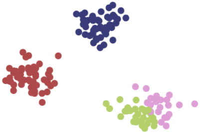
K = 2



K = 3

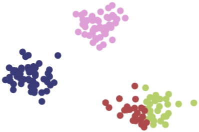


K = 4

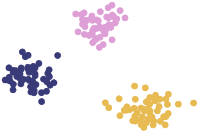


Affinity Propagation

Damping = 0.5



Damping = 0.8



Affinity Propagation

No need to know number of clusters

Responsibility

	A	B	C	D
A	-2.8	-0.7	-2.8	2.8
B	-0.9	-3.0	-2.3	3.0
C	-2.0	-1.3	-2.0	2.0
D	-4.9	-4.7	-5.7	0.0

Availability

	A	B	C	D
A	0.0	-3.0	-2.0	0.0
B	-2.8	0.0	-2.0	0.0
C	-2.0	-3.0	0.0	0.0
D	-2.8	-3.0	-2.0	7.7

Example of Affinity Propagation

Participant	Tax Rate	Fee	Interest Rate	Quantity Limit	Price Limit
Alice	3	4	3	2	1
Bob	4	3	5	1	1
Cary	3	5	3	3	3
Doug	2	1	3	3	2
Edna	1	1	3	2	3

↓

Participant	Alice	Bob	Cary	Doug	Edna
Alice	-22	-7	-6	-12	-17
Bob	-7	-22	-17	-17	-22
Cary	-6	-17	-22	-18	-21
Doug	-12	-17	-18	-22	-3
Edna	-17	-22	-21	-3	-22

Compute Responsibility

$$r(i,k) \leftarrow s(i,k) - \max_{k' \text{ such that } k' \neq k} \{a(i,k') + s(i,k')\}$$

Participant	Alice	Bob	Cary	Doug	Edna
Alice	-22	-7	-6	-12	-17
Bob	-7	-22	-17	-17	-22
Cary	-6	-17	-22	-18	-21
Doug	-12	-17	-18	-22	-3
Edna	-17	-22	-21	-3	-22

$\downarrow$

Participant	Alice	Bob	Cary	Doug	Edna
Alice	-16	-1	1	-6	-11
Bob	10	-15	-10	-10	-15
Cary	11	-11	-16	-12	-15
Doug	-9	-14	-15	-19	9
Edna	-14	-19	-18	14	-19

Compute Availability

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' \text{ such that } i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}$$

Participant	Alice	Bob	Cary	Doug	Edna
Alice	-16	-1	1	-6	-11
Bob	10	-15	-10	-10	-15
Cary	11	-11	-16	-12	-15
Doug	-9	-14	-15	-19	9
Edna	-14	-19	-18	14	-19

$\downarrow$

Participant	Alice	Bob	Cary	Doug	Edna
Alice	21	-15	-16	-5	-10
Bob	-5	0	-15	-5	-10
Cary	-6	-15	1	-5	-10
Doug	0	-15	-15	14	-19
Edna	0	-15	-15	-19	9

Compute Availability diagonal

$$a(k,k) \leftarrow \sum_{i' \text{ such that } i' \neq k} \max\{0, r(i',k)\},$$

Participant	Alice	Bob	Cary	Doug	Edna
Alice	-16	-1	1	-6	-11
Bob	10	-15	-10	-10	-15
Cary	11	-11	-16	-12	-15
Doug	-9	-14	-15	-19	9
Edna	-14	-19	-18	14	-19

$\downarrow$

Participant	Alice	Bob	Cary	Doug	Edna
Alice	21	-15	-16	-5	-10
Bob	-5	0	-15	-5	-10
Cary	-6	-15	1	-5	-10
Doug	0	-15	-15	14	-19
Edna	0	-15	-15	-19	9

Cluster assignment

Criterion Matrix: $\mathbf{C} = \mathbf{R} + \mathbf{A}$

Participant	Alice	Bob	Cary	Doug	Edna
Alice	**5**	-16	-15	-11	-21
Bob	**5**	-15	-25	-15	-25
Cary	**5**	-26	-15	-17	-25
Doug	-9	-29	-30	**-5**	-10
Edna	-14	-34	-33	**-5**	-10

- Cluster 1: Alice, Bob, Cary
- Cluster 2: Doug, Edna

Clustering Evaluation Metrics

Intrinsic measures

- The ground truth labels are not known
- Example: Silhouette Coefficient

$$s = \frac{b-a}{\max(a,b)}$$

a mean distance between sample & other points in the class.
b mean distance between sample & nearest cluster.

silhouette mean of **s** across all points.

Clustering Evaluation Metrics

Extrinsic measures

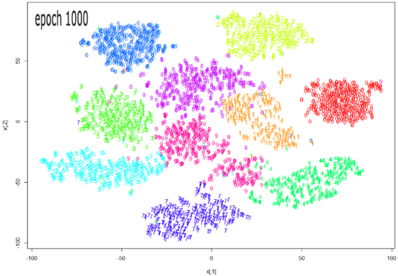
- The ground truth labels are available
- No label correspondence [0 0 0 1 1 1 2 2 2] vs [1 1 1 0 0 0 2 2 2]
- Example: Clustering Accuracy

$$ACC(Y, C) = \max_{perm \in P} \frac{1}{N} \sum_{i=0}^{n-1} 1\left(perm\left(C_i\right) = Y_i\right)$$

Outline

What is dimensionality reduction?

Original data in high dimension $N \rightarrow$ find an embedding of smaller dimension M which still represents the initial data.

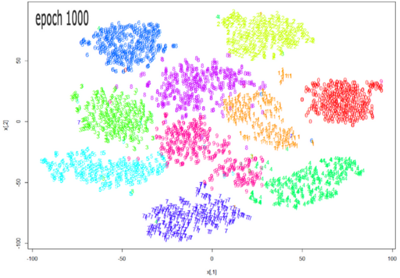


Usage?

- Data visualization
- Data exploration
- Extract underlying concepts
- Speed up learning algorithms
- Scale data to algorithm
- Data compression

How to achieve DR?

$N = 28 \times 28 = 784$



$M = 2$

How to go from 784 feature dimensions to 2 while keeping important information?

- Feature elimination
- Feature selection
- **Feature extraction**

Popular algorithms

Linear algebra methods

Matrix factorization methods drawn from the field of linear algebra can be used for dimensionality reduction.

- **Principal Components Analysis**
- Singular Value Decomposition
- Non-Negative Matrix Factorization
- Independent Components Analysis

Manifold Learning Methods

Manifold learning methods seek a lower-dimensional projection that captures some properties of the input.

- Isomap Embedding
- Locally Linear Embedding
- Spectral Embedding
- **t-distributed Stochastic Neighbor Embedding**
- Uniform Manifold Approximation and Projection

Principal Components Analysis(PCA)

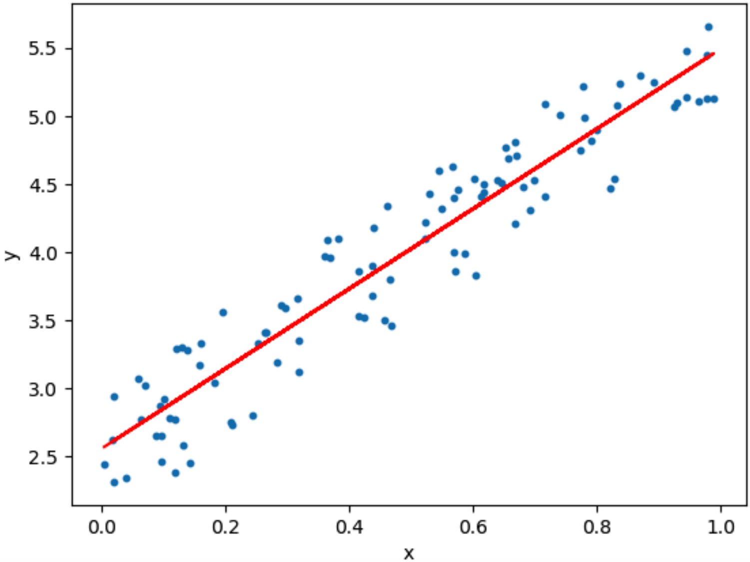
Goal: Find r-dim projection that best preserves variance

1. Compute mean vector μ and covariance matrix Σ of original points
2. Compute eigenvectors and eigenvalues of Σ
3. Select top r eigenvectors
4. Project points onto subspace spanned by them:

$$y = A(x - \mu)$$

where y is the new point, x is the old one, and the rows of A are the eigenvectors

Principal Components Analysis(PCA)



t-Distributed Stochastic Neighbor Embedding (t-SNE)

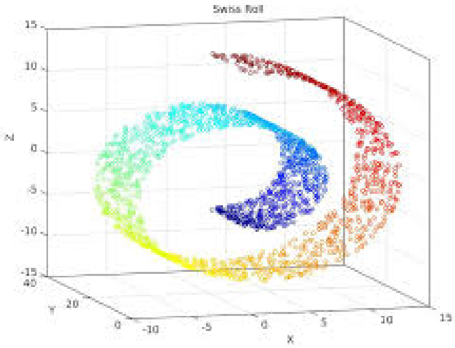
High dimensional space

$$p_{ij} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_k \sum_{l \neq k} \exp -||x_k - x_l||^2/2\sigma_i^2}$$

Low dimensional space

$$q_{ij} = \frac{(1+||y_i - y_j||^2)^{-1}}{\sum_k \sum_{l \neq k} (1+||y_k - y_l||^2)^{-1}}$$

$$\underset{\text{points}}{\text{Minimize}} \quad KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

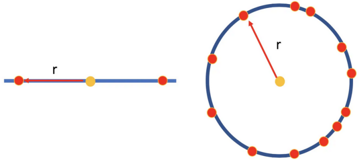


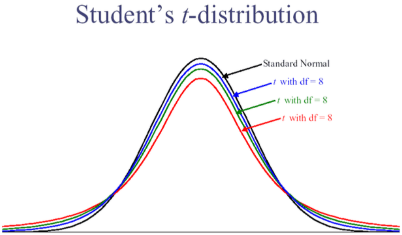
- Random initialization
- Gradient descent
- Preserves local structures
- Little dependant on tunable parameters

t-SNE explanations

Why different distributions in different spaces?

Crowding problem
(Curse of dimensionality)



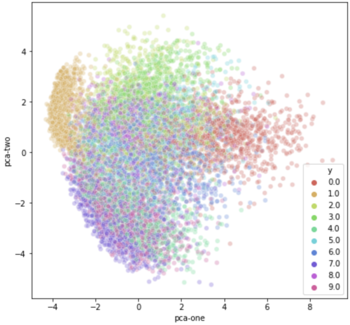


Hyperparameters?

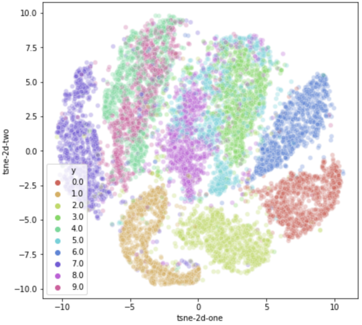
Perplexity - the number of neighbors for any point used to compute σ_i

- High perplexity: Takes more global structures into account
- Low perplexity: Takes more local structures into account

PCA vs t-SNE



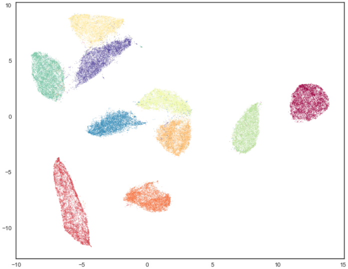
pca-two
pca-one
4
2
0
−2
−4
−4
−2
0
2
4
6
8
y
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0



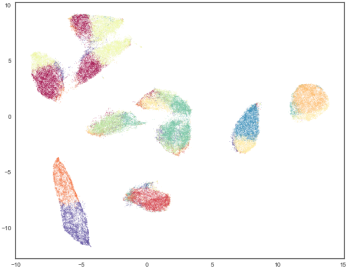
tsne-2d-two
tsne-2d-one
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10
−5
0
5
10
y
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0

Dimensionality Reduction before Clustering - MNIST example

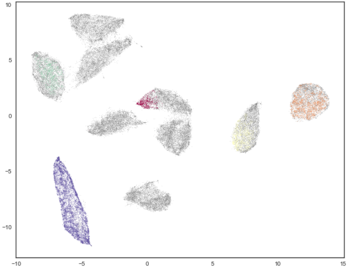
Ground-truth labels



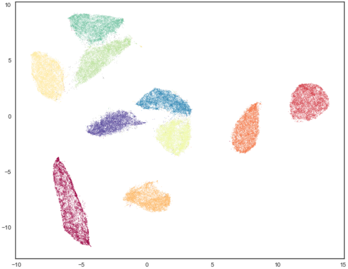
K-Means



HDBSCAN



UMAP + HDBSCAN



Using Dimensionality Reduction before Clustering

Points of concerns

- Does not completely preserve density.
- Can create false clusters.

$\longrightarrow$ Do some exploration and evaluation of the clusters that come out to try to validate them if possible.

Outline

1. Motivations

2. Overview of Clustering Techniques

3. Dimensionality Reduction

4. **Clustering countries based on early Covid-19 spreading data**

5. Improving further the results from the perspective of clustering

Features to represent a region

- All countries from the JHU dataset
- US states
- Canadian provinces
- Chinese provinces
- Australian states
- Brazilian states
- Italian regions

Objective

Cluster countries/regions together, and use non-Brazil data from the same group to predict Braziliam states epidemic propagation.

Features used for clustering

Early Mortality weekly number of deaths 14 days after the outbreak, divided by the number of confirmed cases, in the week of the outbreak. A two weeks period was used because it is the time required to know the outcome of a contamination.

Days until 10x the number of days it takes to multiply the confirmed cases by 10, from the day of the outbreak.

Early Acceleration if we denote $\Delta W0W1$ as the percentage increase of confirmed cases from the week of the outbreak to the week after, and $\Delta W1W2$ as he percentage increase from the 1st to the 2nd week after the outbreak, then the early acceleration is defined by:

$$earlyAccel = \frac{\Delta W1W2}{\Delta W0W1}$$

Approach

Dimensionality reduction

Uniform Manifold Approximation (UMAP) embedding

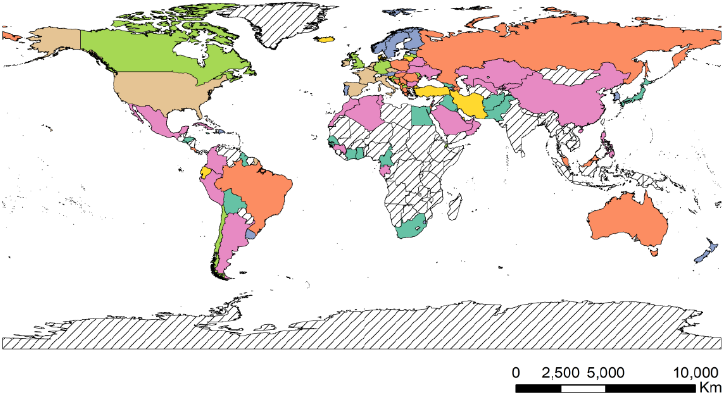
- Handles well the balance local vs global for keeping distances in the low dimensional embedding
- Based on studying the topology of the spaces studied
- Hyperparameters: min_dist = 0, n_neighbors=15

Clustering

Affinity propagation

- Hyperparameters: damping=0.8

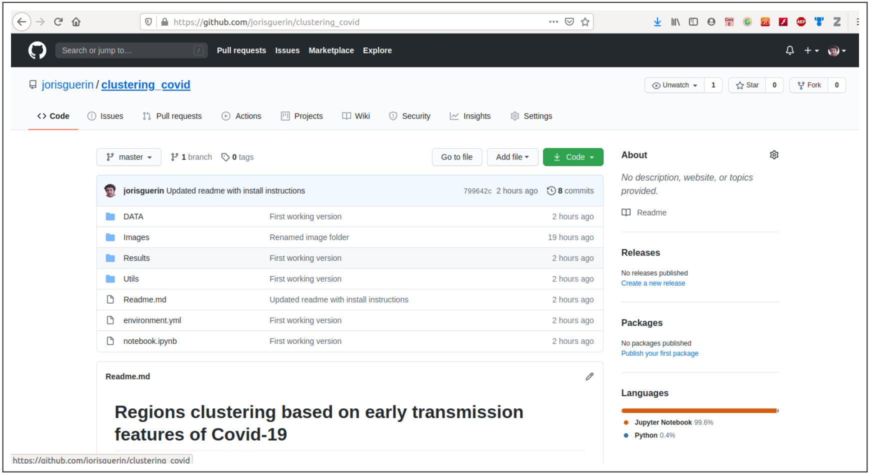
Validation



- There is no truth!
- Look if your intuitions are respected to see if clusters make sense.
- Look if the NN predictions improved after using the clusters for training.

Code

`https://github.com/jorisguerin/clustering_covid`



Outline

1. Motivations

2. Overview of Clustering Techniques

3. Dimensionality Reduction

4. Clustering countries based on early Covid-19 spreading data

5. Improving further the results from the perspective of clustering

Strengthen validation approach & Improve current clustering

Validation approach

Conduct a real numerical study on how much improvements are reach using countries from a cluster vs using all countries.

Going further

Use this validation to have a numerical feedback on different "dimasionality reduction + clustering" combination and optimize the grouping pipeline.

Adapt features to refine training regularly

Modify features to be "dynamic":

- Early Mortality $\rightarrow$ Current Mortality
- Days until 10x $\rightarrow$ Current Speed
- Early Acceleration $\rightarrow$ Current Acceleration

And retrain network with updated clusters every day/week/month.

Grouping Countries and Regions to Improve Covid-19 Dynamics Predictions

UFRN Covid-19 course Fall 2020

presented by **Joris Guérin**

on October 13 2020