Spectral Clustering

Machine Learning and Deep Learning - 2021

Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello

November 4th, 2021

University of Modena and Reggio Emilia

GRAFO DI SEGNALI, NxN è MATRICE CHE LO DESCRIVE

Spectral clustering: algorithm (1)

Clustering model based on the spectral graph theory.

- build a graph over examples, representing it with the adjacency matrix A

$$A_{i,j} = e^{-\frac{\sum_{k=1}^{d} ||x_i^k - x_j^k||^2}{\sigma^2}}$$

REGOLA DOME. AFFINITÀ
SE ALTO LE DISTANZE VENGONO EVIDENZIATE

- build the degree matrix D of the graph. It is a diagonal matrix holding for each element the sum of the incoming adges.
- compute the normalized laplacian L

$$L = D - A \qquad \text{(SOTTRAZIONE MATRICI STESSA DIM)} \tag{1}$$

Spectral clustering: algorithm (2)

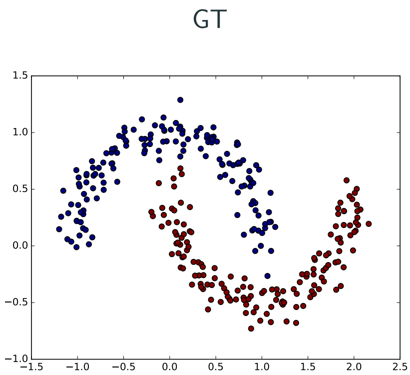
- Compute the eigenvectors and sort them for increasing eigenvalues
- Choose the eigenvectors from the second to the desired number of clusters
- Those eigenvectors provide a representation of data in a fancy embedding space!

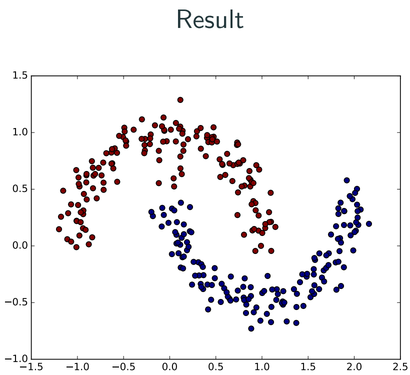
Spectral clustering: algorithm (3)

You may implement the solution according to one of the following:

- **Fiedler-Vector Solution:** in the event of a **binary clustering problem**, you may simply consider the eigenvector corresponding to the 2^{nd} smallest eigenvalue (i.e. the Fiedler Vector) and use the positivity of its components as labels;
- **K-Means solution:** in a more general case where you need n **distinct clusters**, you may consider the eigenvectors corresponding to the n smallest eigenvalues and use their components as the representation of old data in a new embedding space. You can apply K-Means to this new data representation.

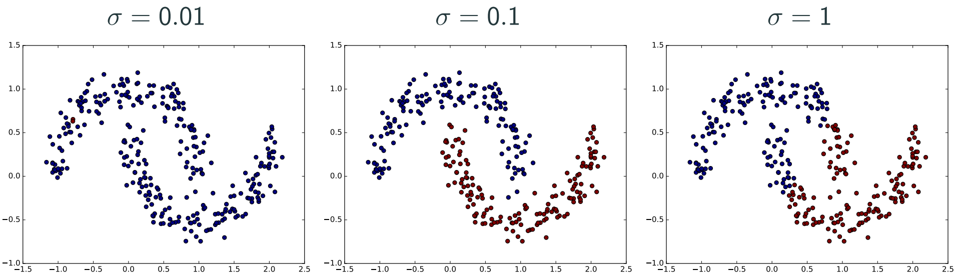
Spectral clustering: result





Beware: σ counts (in large amounts!)

b È UN IPERPARAMETRO



b è la TEMPERATURA e DEFINISCE L'AGGLOMERAZIONE DEI PUNTI NEL NUOVO SPAZIO

- $b \to 0 \Rightarrow$ PUNTI SI ALLONTANANO $\Rightarrow$
  - KMEANS AGGLOMERA I° PUNTO AL I° CLUSTER
  - I RESTANTI SONO DEL II CLUSTER
- $b \to \infty \Rightarrow$ PUNTI SI AVVICINANO $\Rightarrow$
  - I PRIMI N-1 PUNTI SONO DEL I CLUSTER
  - N-ESIMO PUNTO È DA SOLO NEL II CLUSTER

$\Rightarrow$ TRADE OFF b

VA FATTO IL TUNING DI b PER EVITARE DI GENERARE LE SITUAZIONI DESCRITTE.

Useful functions

- datasets.gaussian_dataset

```
data, cl = gaussians_dataset(3, [100,100,70], [[1, 1],[-4, 6],[8, 8]], [[1, 1],[3, 3],[1, 1]])
```

- datasets.two_moon_dataset

```
data, cl = two_moon_dataset(n_samples=300, noise=0.1)
```

- matplotlib.pyplot.plot
- matplotlib.pyplot.scatter
- scipy.linalg.fractional_matrix_power
- numpy.linalg.eig
- numpy.argsort