Diffusion Maps

Ángela Fernández Pascual

Universidad Autónoma de Madrid

March 28, 2011

Contenido

1. Introduction

2. Classical Methods

3. Spectral Embedding

4. Diffusion Maps

5. Anisotropic Diffusion

6. Nyström Formula

7. Examples

8. Conclusions

Introduction

Dimensionality reduction

- **Problem:**
  The sample data sets lie in a differential manifold of low dimension embedded in a space of a higher dimension.
- **Interest:**
  - Working in low dimension is always easier and has many computational advantages.
  - We try to obtain and use neighborhood information of each point that describe the geometry of our data.
- We achieve both goals defining Spectral Embeddings or Diffusion Maps.

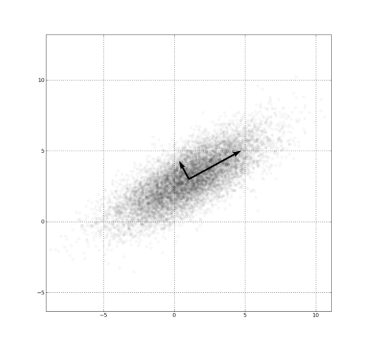
Clustering

- Clustering is one of the most widely used techniques for data analysis.
- We can also apply Spectral Embeddings or Diffusion Maps with clustering objectives.

Classical Methods

PCA

- **Objective:** Simplify the data structure transforming the original features into others, named principal components, applying linear combinations of those features retaining variance.
- This method only rotate the coordenates, changing the data axis in the maximum variance direction.
- The principal components are decorrelated.



Classical Methods II

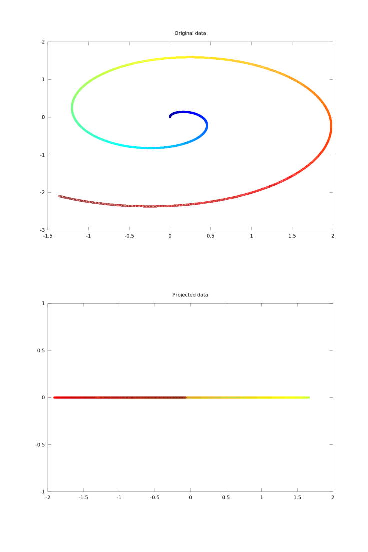
PCA: Advantages

- PCA studies the features relation, finding feature groups that are highly correlated.
- It is an useful technique for feature selection, for outlier detection or for clustering.
- It works properly when our features present a high correlation because few factors would explain a high part of the total variability.

PCA: Disadvantages

- PCA does not take into account the vector's classes, so it does not look at the classes' separability.
- PCA makes a linear data transformation.
- PCA is not able, for example, to reduce properly a spiral because it does not take into account the geometry of the data.

A spiral in dimension 1 with PCA



Original data

Projected data

Spectral Embedding Algorithm

1. Constructing the adjacency graph and choosing its weights.

$$G = (S = \{\mathbf{x}_i\}, E), \text{ where } (i,j) \in E \text{ if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are near},$$

$$W_{ij} = w(x_i, x_j) = e^{\frac{-||x_i - x_j||^2}{\epsilon}}, \text{ the most comun option for weights.}$$

2. Computing eigenmaps:
   - Laplacian Graph L.
   - Compute eigenvalues and eigenfunctions of $I - L$.
   - Obtain the first k eigenvectors $v_1, \dots, v_k$ of L, corresponding to the major eigenvalues $\lambda_1 > \cdots > \lambda_k$.
     We reject the trivial solution $f_0 = \mathbf{1}$, associated to $\lambda_0 = 1$.
   - Let $V \in \mathbb{R}^{N \times k}$ be the matrix with the selected eigenvectors in columns.
     To obtain the embedding:

```
for $i = 1$ to $N$
    Compute $y_i \in \mathbb{R}^k$ as the $i$-th row of $V$.
```

$$\begin{array}{cc}
\begin{array}{ccc} v_1 & \qquad & v_k \\ \downarrow & & \downarrow \end{array} & \\
\left(\begin{array}{ccc} & & \\ & & \\ & & \end{array}\right) & \begin{array}{l} \leftarrow y_1 \\ \\ \leftarrow y_n \end{array}
\end{array}$$

Laplacian Graphs

Laplacian Graphs

- Laplacian Graphs are the main tool for Spectral Embedding.
- The Laplacian is an operator that describes the connections in our graph, so it describes the geometry of our data.
- We assume that our graph G is an undirected weighted graph whose weighted matrix W has positive entries ($w_{ij} = w_{ji} \geqslant 0$).

Classification

- Unnormalized Laplacian Graph

$$L_{un} = D - W$$

  where D is the diagonal degree matrix with $d_{ii} = \sum_{j=1}^{N} w_{ij}$.
- Normalized Laplacian Graph
  - Symmetric Laplacian

$$L_{sym} = D^{-\frac{1}{2}} L_{un} D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}.$$

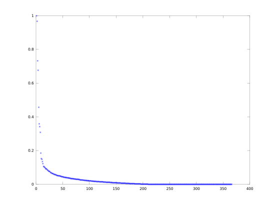
  - Random Walk Laplacian

$$L_{rw} = D^{-1} L_{un} = I - D^{-1} W.$$

Embedding Justification

Why the first eigenvalues?

In mathematics, the highest eigenvalues are closely related with the geometry of the manifold.



And why this embedding?

In our proyection, we search for the construction of an embedding function that preserves the local information.
We assure this fact searching for the optimal embedding, the one that minimizes distances between closed nodes: $J(Y) = \frac{1}{2}\sum_{i,j}(y_i - y_j)^2 w_{ij} \geqslant 0$.

Diffusion Maps

Motivation

The Spectral Embeddings don't have a clear justification about why they work.

Objective

We want to define a new representation of the data preserving some quantities of interest such as local mutual distances.

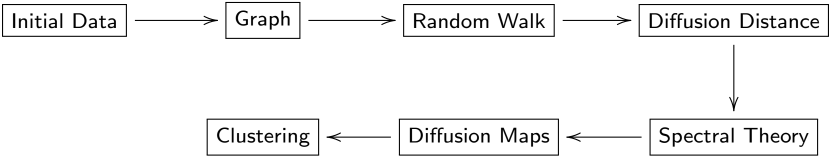


Figure: Steps for Diffusion Maps.

1. Initial Data: construction of a graph

Initial data

- $\Omega = \{x_1, \dots, x_n\}$.

Constructing a graph

- Nodes: x_i.
- Weights: $W_{ij} = w(x_i, x_j)$.

  Usually, we fix them following a Gaussian Kernel: $w(x_i, x_j) = e^{\frac{-||x_i - x_j||^2}{\epsilon}}$.
  W must always be:
  - Symmetric.
  - Pointwise positive.
- $G = (\Omega, W)$ represents the local geometry of the graph.

A Markov Random Walk

- Thanks to the properties of W we can define a Markov random walk on the graph.
- In this probabilistic formulation, the transition probability comes defined by:

$$P_{ij} = p(x_i, x_j) = \frac{w(x_i, x_j)}{d(x_i)},$$

where $d(x_i) = \sum_{k=1}^{n} w(x_i, x_k)$ is the degree of the graph.

- We can consider larger neighborhoods with the powers of P, so $P_{ij}^t = p_t(x_i, x_j) \equiv$ the probability transition from x_i to x_j in t time steps.

3. Diffusion Distance

Main Goal

Define a distance metric on the original set that reflects the connectivity of the data.

Diffusion distance

- Two points are close if they are connected by many short paths in the graph.
- The square of the diffusion distance in t time steps is defined by

$$\begin{aligned} D_t^2(x,z) &= ||p_t(x,\cdot) - p_t(z,\cdot)||^2_{L^2(\frac{1}{\phi_0})} \\ &= \sum_{y\in\Omega} \frac{(p_t(x,y) - p_t(z,y))^2}{\phi_0(y)}, \end{aligned}$$

where ϕ_0 is the stationary distribution of the Markov process, defined as

$$\phi_0(x) = \frac{d(x)}{\sum_{z\in\Omega} d(z)}$$

and $\frac{1}{\phi_0}$ penalizes discrepancies on domains of low density more than on those of high density.

4. Spectral Theory

Dimensionality Reduction

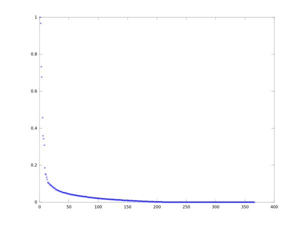
- We use the spectral theory to simplify the diffusion distance calculation.

$$D_t^2(x,z) = \sum_{j=1}^{n-1} \lambda_j^{2t} (\psi_j(x) - \psi_j(z))^2,$$

  where λ_j are the eigenvalues and ψ_j are the eigenvectors of P.

- $\psi_0 = 1$ is the trivial solution.

Eigenvalue decay



Approximation to diffusion distance:

$$D_t^2(x,z) \sim \sum_{j=1}^{d(t)} \lambda_j^{2t} (\psi_j(x) - \psi_j(z))^2.$$

where $d(t) = \max\{l : |\lambda_l^t| > \delta |\lambda_1^t|\}$
with δ, the precision.

Definition

- The approximation of D_t^2 can be interpreted as the Euclidean Distance in $\mathbb{R}^{d(t)}$ if we define an embedding as:

$$\Psi_t = \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \vdots \\ \lambda_{d(t)}^t \psi_{d(t)}(x) \end{pmatrix}.$$

- The diffusion distance between two points of the original set coincide with the euclidean distance between the embedded points.

$$D_t^2(x,z) \sim \sum_{j=1}^{d(t)} \lambda_j^{2t} (\psi_j(x) - \psi_j(z))^2 = ||\Psi_t(x) - \Psi_t(z)||^2.$$

6. Clustering

Clustering in euclidean spaces

- We have the data embedded in an Eucliden space of lower dimension.
- We could apply k-means algorithm over this new set, obtaining k clusters $C_1, \cdots, C_k$.
- The clusters in the original space Ω are $A_1, \cdots, A_k$ such that

$$A_i = \{x_j | \Psi_t(x_j) \in C_i\}.$$

Summary: Diffusion Maps Algorithm

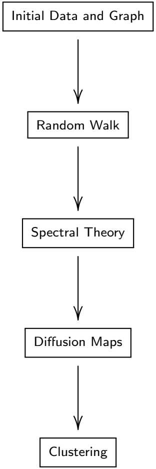
1. $\Omega = \{x_1, \cdots, x_n\}$ our data set.
2. Construct $G = (\Omega, W)$ where $W_{ij} = K(x_i, x_j)$ a symetric, positive kernel.
3. Define the transition probability $P_{ij} = p(x_i, x_j) = \frac{K(x_i,x_j)}{d(x_i)}$.
4. Obtain eigenvalues $\{\lambda_l\}_{l\geqslant 0}$ and eigenfunctions $\{\psi_l\}_{l\geqslant 0}$ of P such that

$$\left\{ \begin{array}{ccl} 1 & = & \lambda_0 > |\lambda_1| \geqslant \cdots \\ P\psi_l & = & \lambda_l \psi_l. \end{array} \right.$$

5. The threshold $d(t) = \max\{l : |\lambda_l^t| > \delta|\lambda_1^t|\}$.
6. Diffusion Map:

$$\Psi_t = \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \vdots \\ \lambda_{d(t)}^t \psi_{d(t)}(x) \end{pmatrix}.$$

7. Clustering over the embedding (if desired).



Anisotropic Diffusion: the Relevance of the Density

Relevance of the Density

- Context: Points are sampled from a probability density on a submanifold of $\mathbb{R}^n$.
- Problem: The sample distribution is often not related to the geometry of the manifold.
- Goal: To recover the manifold structure regardless of the distribution of the data.

Anisotropic Diffusion

- To define $\alpha \in \mathbb{R}$ that specifies the influence of the density in the transition of the diffusion process: $\alpha = 0$ means maximal influence, $\alpha = 1$ means there isn't influence.
- To construct the diffusion family: we normalize twice the kernel matrix.

1. Fix α and $K_\sigma(x, y) = e^{-\frac{||x-y||^2}{\sigma}}$.
2. $q_\sigma(x_i) = \sum_{j=1}^{n} K_\sigma(x_i, x_j)$ is the density function.
3. First normalization: $K_\sigma^{(\alpha)}(x, y) = \frac{K_\sigma(x,y)}{q_\sigma(x)^\alpha q_\sigma(y)^\alpha}$.
4. The new degree is $d_\sigma^{(\alpha)}(x_i) = \sum_{j=1}^{n} \frac{K_\sigma(x_i,x_j)}{q_\sigma(x_i)^\alpha q_\sigma(x_j)^\alpha}$.
5. The final probability matrix is given by $p_{\sigma,\alpha}(x, y) = \frac{K_\sigma^{(\alpha)}(x,y)}{d_\sigma^{(\alpha)}(x)}$.

Anisotropic Diffusion II

- The Markov chain in continuous time is usually defined by an infinitesimal generator instead of by the probability matrix that we have just seen.

Theorem

Let

$$L_{\sigma,\alpha} = \frac{I - P}{\sigma}$$

be the infinitesimal generaton of the Markov chain. Then, for a fixed K we have:

$$\lim_{\sigma \to 0} L_{\sigma,\alpha} f = \frac{\Delta(fq^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}} f,$$

where Δ is the Laplace Beltrami operator.

- When $\alpha = 1$,

$$\lim_{\sigma \to 0} L_{\sigma,\alpha} f = \Delta f.$$

- When $\alpha = 0$ and the density of q is uniform in the submanifold, also occur that

$$\lim_{\sigma \to 0} L_{\sigma,\alpha} f = \Delta f.$$

Out-of-sample method: Nyström Formula

Motivation

The previous algorithm only works for train samples. If we have some new data points to cluster, we must repeat the whole algorithm to classify them.

Idea

To extend the algorithm in a way such that they admit out-of-sample points without repeat the algorithm with the full dataset.

The Nyström Formula

It is a general method of kernel eigenfunctions learning that is based on the prediction of the eigenvector value and the eigenvalue convergence.

$$\phi_i^n(\mathbf{x}) = \frac{\sqrt{n}}{\ell_i} \sum_{j=1}^{n} v_{ji}^n K^n(\mathbf{x}, \mathbf{x}_j) \quad i = 1, \cdots, k$$

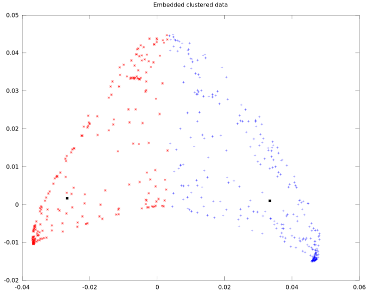
where

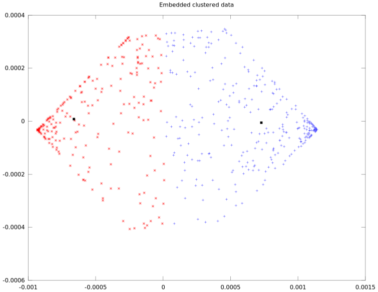
- ϕ_i^n eigenfunctions desired.
- $K^n(\mathbf{x}_i, \mathbf{x}_j)$, sample kernel function (determines the weights of the graph).
- ℓ_i y v_{ji}^n autovalores y autofunciones de $K^n(\mathbf{x}, \mathbf{x}_j)$.

Some embedding examples in a real problem

Framework

- Original data: *dimension* $= 4$. They represent the cloudiness of the main 4 hours in a day of solar radiation.
- Embedding: *dimension* $= 2$.
- Depending on the parameters (α, σ of the Gaussian Kernel, ...) we can obtain a different data representation on the embedding.





Conclusions

Diffusion Maps *versus* Spectral Clustering

Diffusion Maps present some advantages opposite to Spectral Embeddings:

1. They are a more general technique.
2. They have a solid theory behind, as they are based on Markov processes and a new metric, the Diffusion Distance.
3. They are more versatile as they allow us to take into account more steps in the diffusion processes and also to evaluate the relevance of the density.