Statistical Learning

Functional data analysis

Pedro Galeano
Department of Statistics
UC3M-BS Institute on Financial Big Data
Universidad Carlos III de Madrid
pedro.galeano@uc3m.es

Academic year 2016/2017

Master in Big Data Analytics

uc3m | Universidad Carlos III de Madrid

1. Introduction

2. Functional data analysis

3. Functional principal component analysis

4. Functional unsupervised classification

5. Functional supervised classification

1. Introduction

2. Functional data analysis

3. Functional principal component analysis

4. Functional unsupervised classification

5. Functional supervised classification

Introduction

- The CanadianWeather data set:
  - The data set: Contains average daily temperature and precipitation at 35 different locations in Canada in the period from 1960 to 1994.
  - Therefore: For every of the 35 locations, we have 365 consecutive observations corresponding to the average daily temperature of the 365 days of the year, and similarly for the average daily precipitation.
  - Important point: We observe two processes over time.
  - Problems:
    - Predict precipitation in terms of temperature.
    - Understand the main sources of variation of temperatures or precipitations in terms of time.
    - Obtain clusters of similar locations in terms of temperature or precipitation.

Introduction

- Chapter 4.R script:
  - Functional data analysis: The CanadianWeather data set.

Introduction

- The Phoneme data set:
  - The data set: Contains 500 log-periodograms correponding to 5 phonemes, “sh”, “iy”, “dcl”, “aa” and “ao”, measured at 256 consecutive frequencies.
  - Problem: Use such data to classify a spoken phoneme.

Introduction

- Chapter 4.R script:
  - Functional data analysis: The Phoneme data set.

Introduction

- Functional data analysis: Analysis of data sets in which a process is observed over a continuum.
- Continuum: Usually, the time, as in the Canadian weather data set, but it is not the only case, as the frequency in the Phoneme data set.
- Discrete data: Even if we assume that there is a process observed in a continuum, the data that we have is discrete because the process is observed at a certain number of points.

Introduction

- The rest of this chapter is devoted to:
  - Introduce briefly functional data analysis.
  - Present some characteristic measures for functional data analysis.
  - Have a very brief look at:
    - Principal component analysis for functional data analysis.
    - Unsupervised classification for functional data analysis.
    - Supervised classification for functional data analysis.

1. Introduction

2. Functional data analysis

3. Functional principal component analysis

4. Functional unsupervised classification

5. Functional supervised classification

Functional data analysis

- Functional data sets: Composed of n observations of processes observed at p consecutive points.
- The process: $x(t)$, for every $t \in [a, b]$.
- Observation points: $t_1 < \cdots < t_p \in [a, b]$.
- Observed data: $x_i(t_j)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- Question: Is it possible to know the values of $x(t)$ at points t that are not those in the set $t_1, \ldots, t_p$?
- Answer: It is possible to approximate these values using smoothing methods (not entering into details here).

Functional data analysis

- Mean function:

  $$\mu(t) = E[x(t)]$$

  is the mean of the process at any point $t \in [a, b]$.

- Sample mean function:

  $$\widehat{\mu}(t_j) = \frac{1}{n}\sum_{i=1}^{n} x(t_j)$$

  for $j = 1, \ldots, p$.

- Consequently: The mean function $\mu(t)$ is estimated at points $t_1, \ldots, t_p$.

Functional data analysis

- Covariance function:

$$\Gamma(s,t) = E\left[(x(s)-\mu(s))(x(t)-\mu(t))\right]$$

  is the covariance between $x(s)$ and $x(t)$, for any pair $s, t \in [a,b]$.

- Sample covariance function:

$$\widehat{\Gamma}(t_j,t_k) = \frac{1}{n}\sum_{i=1}^{n}(x_i(t_j)-\widehat{\mu}(t_j))(x_i(t_k)-\widehat{\mu}(t_k))$$

  for $j = 1,\ldots,p$.

- Consequently: The covariance function $\Gamma(s,t)$ is estimated at pairs of points (t_j,t_k), where t_j and t_k are in the set $t_1,\ldots,t_p$.

1 Introduction

2 Functional data analysis

3 Functional principal component analysis

4 Functional unsupervised classification

5 Functional supervised classification

Functional principal component analysis

- Functional principal components of $x(t)$:

$$x(t) = \sum_{k=1}^{\infty} s_k v_k(t)$$

where:

- $v_1, v_2, \ldots$ are the eigenfunctions defined in $t \in [a, b]$.
- $s_1, s_2, \ldots$ are the scores given by:

$$s_k = \int_a^b (x(t) - \mu(t)) v_k(t) \, dt$$

for $k = 1, 2, \ldots$

- The scores $s_1, s_2, \ldots$ are real random variables with mean 0 and variances $\lambda_1, \lambda_2, \ldots$, respectively.
- $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of the covariance function Γ.

Functional principal component analysis

- In practice: We can approximate every observed function by means of:

$$x_i\left(t_j\right) \simeq \sum_{k=1}^{K} s_{i,k} \widehat{v}_k\left(t_j\right)$$

  where:

  - $\widehat{v}_1, \widehat{v}_2, \ldots$ are the estimated eigenfunctions defined in $t \in [a, b]$.
  - $s_{i,1}, s_{i,2}, \ldots$, are the scores of the observed function x_i, given by:

$$s_{i,k} = \int_a^b \left(x_i\left(t\right) - \widehat{\mu}\left(t\right)\right) \widehat{v}_k\left(t\right) dt$$

    for $k = 1, 2, \ldots$

  - The sample scores $s_{1,k}, \ldots, s_{n,k},$, for $k = 1, 2, \ldots$ have sample mean 0 and sample variance $\widehat{\lambda}_k$.
  - $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots$ are the estimated eigenvalues of the covariance function Γ.

Functional principal component analysis

- Estimation: The eigenfunctions and eigenvalues can be estimated through the sample mean and covariance functions (not entering into details here).
- Utility: Functional principal components are useful, among many other things:
  - To understand the main sources of variation of the process.
  - To obtain low dimensional representations of the infinite dimensional functions.
- Indeed: The FPCs provide a way of looking at the covariance structure of the process that can be much more informative than a direct examination of the covariance function.

Functional principal component analysis

- Chapter 4.R script:
  - Functional principal components: The CanadianWeather data set.

1. Introduction
2. Functional data analysis
3. Functional principal component analysis
4. Functional unsupervised classification
5. Functional supervised classification

Functional unsupervised classification

- Functional unsupervised classification: There are several procedures to perform unsupervised classification for functional data.

- Functional K-means: The K-means procedure is the most popular approach to perform unsupervised classification for functional data.

- Algorithm: The algorithm is similar to the one seen in Chapter 3 for multidimensional data sets, thus we do not repeat it here.

- Number of clusters, K: Most complicated to be determined than in the multidimensional case.

Functional unsupervised classification

- Distance: The only think that we have to take into account is the distance used between functional observations.
- L^2 distance: Equivalent to the Euclidean distance between multidimensional observations.
- Definition: Given two observed functions, x_i and $x_{i'}$, the L^2 distance between them is given by:

$$d_{L^2}\left(x_i, x_{i'}\right) = \left(\int_a^b \left(x_i\left(t\right) - x_{i'}\left(t\right)\right)^2 dt \right)^{1/2}$$

Functional unsupervised classification

- Chapter 4.R script:
  - Functional K-means: The CanadianWeather data set.

1. Introduction

2. Functional data analysis

3. Functional principal component analysis

4. Functional unsupervised classification

5. Functional supervised classification

Functional supervised classification

- Functional supervised classification: There are several procedures to perform supervised classification for functional data.
- Nevertheless: Most of the procedures are different to those in the multidimensional framework.
- Functional KNN: The KNN procedure is one of the few procedures that is similar in both frameworks (multidimensional and functional).
- Algorithm: The algorithm is similar to the one seen in Chapter 3 for multidimensional data sets, thus we do not repeat it here.
- Number of neighbors, K: Can be determined as in the multidimensional case, i.e., using training and tests samples and/or cross-validation.

Functional supervised classification

- Distance: As in K-means, the distance used in KNN is the L^2 distance.
- Nevertheless: Note that here, the concept of neighbor is more ambiguous than in the multidimensional case.
- Shape of the functions: Two functions with different shapes can be very close using the L^2 distance and two functions with similar shapes can be very far with the L^2 distance.
- For instance: Consider the case of two functions such that the second is the first plus a certain constant.

Functional supervised classification

- Chapter 4.R script:
  - Functional KNN: The phoneme data set.

1. Introduction

2. Functional data analysis

3. Functional principal component analysis

4. Functional unsupervised classification

5. Functional supervised classification