Statistical Learning

Week 1 - Multidimensional Data (I)

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

Academic Year 2017/2018

Master in Big Data Analytics

uc3m | Universidad Carlos III de Madrid

1 Introduction

2 Multidimensional data sets

3 Data quality problems

1 Introduction

2 Multidimensional data sets

3 Data quality problems

Introduction

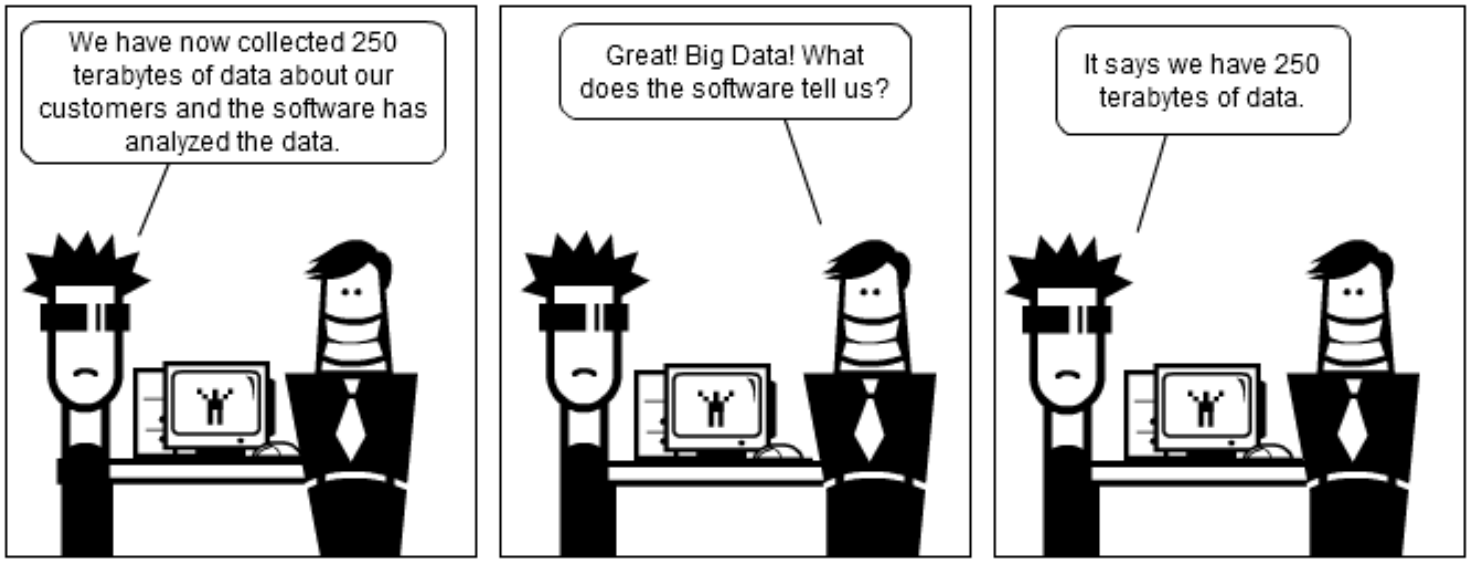
- Statistics: Wide-ranging discipline with certain doses of Mathematics, Empirical Science, Computer Science and Philosophy, among many others.
- Definition: The art and science of learning from data.
- Usual questions treated by Statistics:
  - Does smoking really cause cancer?
  - Does unemployment cause return migration?
  - What will be the maximum temperature of tomorrow?
- Statistical approach: Collect, process and analyze data.
- How to analyze data?: Develop appropriate statistical models and/or methods, perform inference, and predict (or forecast) future outcomes.
- Computer Age Statistics: Almost all topics in 21th-century Statistics are now computer dependent.

Introduction

- Nowadays, the world is awash with data:
  - Read your personal email: Find recommendations to buy products from certain stores based on your previous purchases.
  - Go to the bank: Everyone is subject of a scoring model that predicts whether he/she will default on his/her loan in the near future.
  - Use your smartphone: Your provider analyzes your calling behavior to predict whether you are going to churn in the near future.
  - Read an internet newspaper: You get advertisements based on pages you use to visit.
  - Buy products in the supermarket: I use my credit card, then my provider needs to know if this is a legitimate transaction or not.
  - ...

Introduction

- Nevertheless: "We are drowning in information and starving for knowledge" (Rutherford D. Rogers):



- Crucial need: Take advantage of the large amount of information available today to achieve a thorough understanding of it, by means of data analysis.
- The signal and the noise: The key problem is how to separate the signal from the noise, because only a small percentage of data available is useful.

Introduction

- How to do it?: Use computer-age statistical techniques properly.
- Otherwise: We will kill the goose that lays the golden eggs:
  - https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf
  - http://elpais.com/elpais/2016/11/07/talento_digital/1478535225_341110.html
- Alternative approaches: In machine learning.
- Master: See the courses "Big Data Intelligence: Methods and Technologies" and "Machine Learning" of the Master's program.
- Examples: The next slides show practical problems that are handled with statistical techniques seen in this course.

Introduction

- Business analytics: credit risk modeling
  - Reform measures such as Basel III: Banks need to develop systems in an attempt to model the credit risk arising from important aspects of their business lines.
  - How?: Take different snapshots of information: application and credit bureau information at loan origination, default status information,...
  - Problem: Classify a consumer as good or bad.
  - Technique: Use supervised classification methods taking into account that a labeled data set with good and bad clients (thousands, up to millions) is available.

Introduction

- Business analytics: fraud detection
  - Typical examples: Credit card fraud, insurance claim fraud, money laundering, tax evasion, product warranty fraud, and click fraud, among others.
  - Problem: Label certain operations as fraudulent or not.
  - Techniques:
    - ⋆ Use supervised classification methods taking into account that a labeled data set with fraud objects is available.
    - ⋆ Use unsupervised classification and outlier detection procedures to detect clusters of abnormal operations.

Introduction

- Business analytics: net lift modeling
  - Net lift modeling: Deepen customer relationships by means of targeted or win-back campaigns: mail catalog, email, coupon,...
  - Purpose: Identify customers most likely to respond based on demographic and relationship variables, social network information, and RFM (recency, frequency and monetary) variables.
  - Problem: Classify consumer attitudes (not buying, buying, buying if motivated correctly,...)
  - Technique: Use unsupervised classification to build consumers attitudes and supervised classification to perform future classifications.

Introduction

- Genomics:
  - Microarrays: There are more than 500000 microarrays that are publicly available with each array containing tends of thousands of expression values of molecules.
  - Goal: The large amount of genome sequencing data now make possible to uncover genetic markers of rare disorders and find associations between diseases and rare sequence variants.
  - Characteristics: In such microarrays, the number of variables scales in thousands and is usually larger than the number of individuals involved in the experiments.
  - Problem: Dimension reduction and unsupervised classification methods are frequently used to associate genes or diseases.

Introduction

- Image (and video) analysis:
  - Images and videos: Ones of the most frequent sources of information that are nowadays available for analysis.
  - Examples: Medical images, astrophysical images, surveillance images,...
  - Pixels: Each image can be represented by millions of pixels.
  - Techniques:
    - ⋆ Dimension reduction techniques are frequently used to reduce the size of image and video files.
    - ⋆ Supervised and unsupervised classification methods are used to find similar images (or videos).

Introduction

- Social networks:
  - Social network data: Massive amount of data are being produced by Twitter, Facebook, Youtube,...
  - Uses: These data have been exploited to predict influenza epidemic, stock market trends, and box-office revenues for movies, among many others.
  - Techniques: Use supervised classification, regression and time series models for prediction and forecasting.

Introduction

- Structured data sets: In this course, we will be concerned with techniques for handling well structured data sets.
- Well structured data set: After a data processing exercise, the idea is to analyze a certain collection of characteristics in a certain set of objects.
- Data base processing: Usually, a large amount of work should be done in order to get a well structured data set.

Introduction

- Non-structured data sets: Some non-structured data sets can be converted to well structured data sets.
- For instance: Digital media, like books or newspapers, can be converted into certain structured data sets through text mining techniques (see the courses "Big Data Intelligence: Methods and Technologies" and "Machine Learning" of the Master's program.).
- Examples of well structured data sets: See, next, a few examples of well structured data sets.

Introduction

- The Spam data set (see Week 1.R script):
  - The Spam data set from Hewlett-Packard Labs: Contains the values of 57 variables indicating the frequency of certain words and symbols in 4601 e-mails.
  - The e-mails are labeled: Spam (1813 out of 4601) or non-spam (2788 out of 4601).
  - Example of: Multidimensional data set.
  - Problem: Is it possible to predict if a new email is spam or not using the information provided by this data set?
  - Topics: 1, 2 and 4.

Introduction

- The Default data set (see Week 1.R script):
  - The Default data set: Contains the values of 3 variables on 10000 customers about credit card expenses.
  - The customers are labeled: No (9667 out of 10000) or Yes (333 out of 10000) indicating whether the customer defaulted on their debt.
  - Example of: Multidimensional data.
  - Problem: Is it possible to predict if a new customer will default on their credit card debt?
  - Topics: 1 and 4.

Introduction

- The Credit data set (see Week 1.R script):
  - The Credit data set: Contains the values of 11 variables on 400 customers.
  - Example of: Multidimensional data.
  - Problem: Is it possible to group customers in homogeneous groups based on this information?
  - Topics: 1 and 3.

Introduction

- The NCI60 data set (see Week 1.R script):
  - The data set: Contains expression levels on 6830 genes from 64 cancer cell lines.
  - Example of: Multidimensional data.
  - Problem: Are there groups among the cell lines based on their gene expression levels?
  - Topics: 1, 2 and 3.

Introduction

- The CanadianWeather data set (see Week 1.R script):
  - The data set: Contains daily temperature and precipitation at 35 different locations in Canada in the period from 1960 to 1994.
  - Example of: Functional data.
  - Problems: Several problems can arise including predicting precipitation in terms of temperature or group cities with similar weather behavior.
  - Difference with previous data sets: Here, we observe two processes over time.
  - Topic: 5.

Introduction

- Topic 2: Dimension reduction techniques
  - Noise: The essential structure of multidimensional data sets is obscured by noise.
  - Dimension reduction: It becomes vital to reduce the original data set in such a way that the interesting structure in the data is preserved while irrelevant features are removed.
  - Principal Component Analysis: Simple, elegant, and surprisingly powerful dimension reduction tool.
  - Complex methods: As time moves on, more complex methods are being developed, although PCA has not lost its appeal.

Introduction

- Topic 3: Unsupervised classification methods
  - Unsupervised classification: The problem is to group objects based on one or more variables observed on these objects under certain criteria.
  - Information about group membership: Assumed to be unknown.
  - Number of groups: Assumed to be unknown.
  - Challenging problem: There is not an easy way to check whether or not the groups constructed are appropriate or not.
  - In Statistics: Usually called clustering.

Introduction

- Topic 4: Supervised classification methods
  - Supervised classification: The problem is to classify objects into certain known groups based on one or more variables observed on these objects.
  - Assumption: We have a set of well classified observations, i.e., a set of observations with known associated group.
  - New objects: Use this information to classify new objects with unknown group membership into one of the groups.
  - Number of groups: Known.
  - Checking: See how well our model classify observations with known group membership.

Introduction

- Topic 5: Functional data analysis
  - Functional data: Consists on random functions or curves observed discretely at a finite interval.
  - In a conceptual sense: Functional data are intrinsically infinite dimensional and thus, methods designed for multidimensional data sets are no longer applicable.
  - Functional techniques: Solve similar problems to those of multidimensional data but taking into account the functional nature of the data.

Introduction

- The rest of this topic is devoted to:
  - Introduce the general structure of multidimensional data sets.
  - Present some data quality problems (data pre-processing).
  - Show some useful plots of the information given in the data set.
  - Introduce several interesting descriptive measures of the variables in the data set.
  - Summarize some useful concepts of multidimensional distributions and inference.
  - Illustrate some problems related with sample correlations and some ideas about how to deal with them.

1 Introduction

2 Multidimensional data sets

3 Data quality problems

Multidimensional data sets

- Multidimensional data sets: Multiple measurements or observations obtained on a collection of selected variables.
- Data matrix: Large rectangular arrays where rows represent measurements or observations and columns represent variables.
- Data matrix of massive sizes: Even if they are stored and manipulated in special database systems, from the mathematical point of view, we still think in terms of data matrices.
- Indicator vector: In supervised classification problems, additionally to the data matrix, there is a vector to indicate the group of the associated object.

Multidimensional data sets

- Data matrix: The most important object in multidimensional analysis.
- Usually: The data matrix, denoted by X, contains n multidimensional observations taken on p variables:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Size of the data matrix: $n \times p$.
- Sample size: n, the number of observed objects.
- Dimension: p, the number of observed variables.

Multidimensional data sets

- Generic element of X: x_{ij}, represents the value of the $j-$th univariate variable over the i-th object.
- Values of the $j-$th univariate variable: $x_{1j}, \ldots, x_{nj}$, for $j = 1, \ldots, p$, summarized in the column vector given by $x_{\cdot j} = (x_{1j}, \ldots, x_{nj})'$.
- Values of the $i-$th object: $x_{i1}, \ldots, x_{ip}$, for $i = 1, \ldots, n$, summarized in the column vector given by $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})'$.

Multidimensional data sets

- Indicator vector: Useful in supervised classification problems.
- Usually: The indicator vector contains n values that indicates the group of the associated object:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Vector size: $n \times 1$.
- Sample size: n.
- Generic element of Y: y_i, represents the value of the indicator variable over the i-th object.

Multidimensional data sets

- Week 1.R script:
  - Data matrix and indicator vector: Spam and Default data sets.
  - Data matrix: Credit and NCI60 data sets.

Multidimensional data sets

- Two main data types:
  - Quantitative variables: Their value is a measurable quantity, such as the variables in the data matrices of the spam and NCI60 data sets.
  - Qualitative variables: Their value can be attributed to a category, such as the variable student in the data matrix of the Default data set, or the variable gender in the data matrix of the Credit data set.

Multidimensional data sets

- Quantitative variables:
  - Continuous: Their values can be any number in a finite or infinite interval, such as the expression levels in the data matrix of the NCI60 data set.
  - Discrete: Their values are distinct and separate, such as the variables capitalLong and capitalTotal in the data matrix of the spam data set.
- Qualitative variables:
  - Binaries: There are only two possible values, such as the variable student in the data matrix Default data set.
  - General: There are more than two possible values, such as the variable Ethnicity in the data matrix of the Credit data set.

Multidimensional data sets

- Usually: Qualitative variables are coded numerically.
- For example: The indicator variable spam in the spam data set can be coded with 1, if spam, and 2, if non-spam, or vice-versa.
- Codification: Depends on the problem at hand and the technique used.
- Thus: It is better not to give now a general rule.
- Consequently: More information will be given in subsequent topics.

1 Introduction

2 Multidimensional data sets

3 Data quality problems

Data quality problems

- Problems in multidimensional data sets: Problems of all kinds exists.
- Data cleaning: Problems easy to detect will most likely be found at the data cleaning stage.
- Data analysis: Quite resistant problems might only be discovered during data analysis.

Data quality problems

- Examples:
  - Inconsistencies: Matching data coming from different sources can create inconsistencies.
  - Uninteresting variables: Variables with a very large number of repeated values.
  - Missing data: Incomplete or totally missed observations.
  - Outliers: Observations that do not appear to fit the pattern of the other data values.

Data quality problems

- Inconsistencies:
  - Example 1: Consider different registers from the same person because his name is recorded in different ways: John, Johnny or Jack.
  - Example 2: Assign a wrong code to a product that already exists.
  - Example 3: Negative values of a positive variable, such as age.
- Unfortunately: Some of these mistakes are very difficult to find.
- Solution: Check the behavior of the variables carefully and try to correct everything wrong.
- Also: Sometimes, these problems are found when data analysis is performed.

Data quality problems

- Uninteresting variables:
  - Example: Consider a binary variable with 99.99% of the observations with one value and the 0.01 of the observations with another value.
  - Example: Place of birth.
- Solution: These variables are usually skipped from the analysis unless they provide a very important information.

Data quality problems

- Missing data:
  - Missing values: Very frequent in databases.
  - Marks of missing data: In R, missing values are flag as NA, in SQL, as null,...
  - Important: Sometimes, the label can be very confusing, such as -999 in a variable measuring the age of a person.
  - Thus: Check inconsistencies in the data matrix first.

Data quality problems

- Week 1.R script:
  - Example of missing data: The births2006 data set.

Data quality problems

- Missing data:
  - Popular solution: Delete those observations with missing values in variables of interest (complete case analysis).
  - However: Only acceptable if the number of observations with missing values is small relative to the size of the data set and if the missing data mechanism is independent of the variables.
  - Example of dependency: A survey where participants older than a certain age refuse to answer a particular survey question and age is measured in the study.

Data quality problems

- Missing data:
  - Imputation: Fill an estimated value for each missing observation.
  - Type of variable: Imputation will depend on the characteristics of the variables.
  - Some popular methods:
    - ⋆ Mean or median imputation: Substituting the sample mean or median of all the completely recorded values for that variable (quantitative variables).
    - ⋆ Hot-deck imputation: A missing value is imputed by substituting a value from a similar but complete record in the data set (both type of variables).
    - ⋆ Imputation based on regression: Impute missing values by predicting their values using regression models (both type of variables depending on the regression method used).
  - Consequently: More information will be given at the end of this topic.

Data quality problems

- Outliers:
  - Gross errors: Outliers can occur for many different reasons but should not be confused with gross errors that are cases where something went wrong, such as human or mechanical errors.
  - Outliers in single variables: Are usually easy to detect as they are values that are very large or very small compared with others in the sample.
  - Multidimensional outliers: Much more difficult to detect.
  - Indeed: Multidimensional outliers are not necessary outliers in the single variables.

Data quality problems

- Outliers:
  - Methods: There are several multidimensional outlier detection procedures available, although most of them are based on Gaussian assumptions on the variables.
  - Idea: Obtain low-dimensional visual displays of the data and try to detect the most obvious outliers.
  - For that: We need plots for multidimensional data sets.
  - Consequently: More information will be given in this topic.

Data quality problems

- Week 1.R script:
  - Example of potential outliers: The NCI60 data set.

1. Introduction

2. Multidimensional data sets

3. Data quality problems