

Data Analysis and Unsupervised Learning

Introduction

MAP 573, 2020 – Julien Chiquet

École Polytechnique, Autumn semester, 2020

<https://jchiquet.github.io/MAP573>



Exploratory analysis of (modern) data sets

Assume a table with n individuals described by p features/variables

Questions

Look for **patterns** or **structures** to summarize the data by

- Finding **groups** of "similar" individuals
- Finding variables **important** for these data
- Performing **visualization**

Challenges

Size data may be **large** ("big data ": large n large p)

Dimension data may be **high dimensional** (more variables than individual or $n \ll p$)

Redundancy many variables may carry the **same information**

Unsupervised we **don't necessary know** what we are looking after

An example in genetics: 'snp'

Genetics variant in European population

Description: *medium/large data, high-dimensional*

500, 000 Genetics variants (SNP – Single Nucleotide Polymorphism) for 3000 individuals (1 meter \times 166 meter (height \times width))

- SNP : 90 % of human genetic variations
- coded as 0, 1 or 2 (10, 1 or 2 allele different against the population reference)

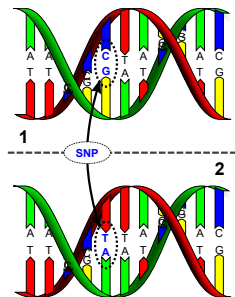


Figure: SNP (wikipedia)

Summarize 500,000 variables with 2 features

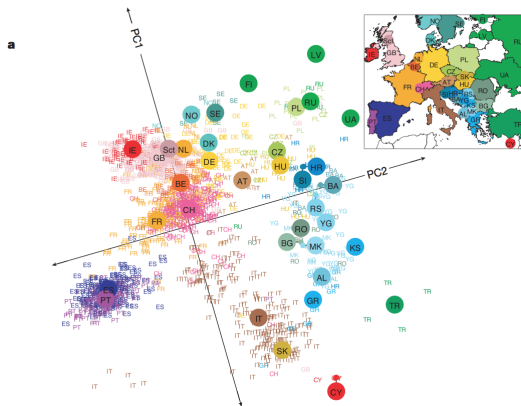


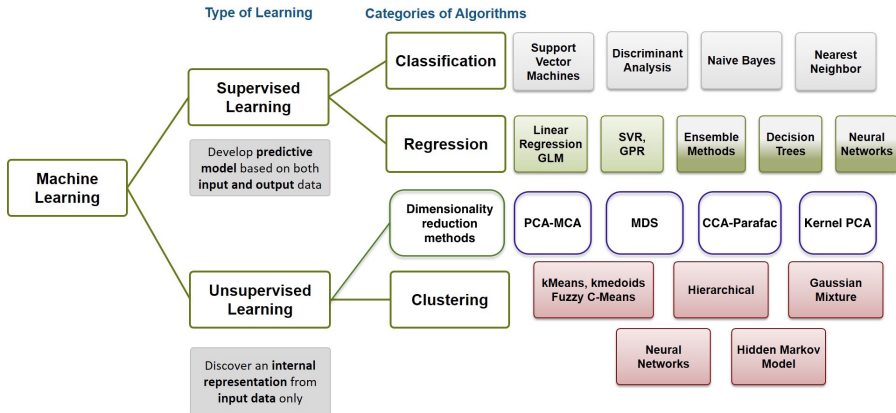
Figure: Dimension reduction + labels source: Nature "Gene Mirror Geography Within Europe", 2008

In the original messy $3,000 \times 500,000$ table, we may find

- an extremely strong structure between individuals ("**clustering**")
- a very simple subspace where it is obvious ("**dimension reduction**")

Overview of Statistics & Machine Learning

Where is MAP573 in this big picture?



Supervised vs Unsupervised Learning

Supervised Learning

- Training data $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, $X_i \sim^{\text{i.i.d}} \mathbb{P}$
- Construct a predictor $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$ using \mathcal{D}_n
- Loss $\ell(y, f(x))$ measures how well $f(x)$ predicts y
- Aim: minimize the generalization error
- Task: Regression, Classification

↪ The goal is clear: predict y based on x (regression, classification)

Unsupervised Learning

- Training data $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$
- Loss? , Aim?
- Task: Dimension reduction, Clustering

↪ The goal is less well defined, and *validation* is questionable

Supervised vs Unsupervised Learning

Supervised Learning

- Training data $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, $X_i \sim^{\text{i.i.d}} \mathbb{P}$
- Construct a predictor $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$ using \mathcal{D}_n
- Loss $\ell(y, f(x))$ measures how well $f(x)$ predicts y
- Aim: minimize the generalization error
- Task: Regression, Classification

↪ The goal is clear: predict y based on x (regression, classification)

Unsupervised Learning

- Training data $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$
- Loss? , Aim?
- Task: Dimension reduction, Clustering

↪ The goal is less well defined, and *validation* is questionable

Dimension reduction: general goals

Main objective: find a **low-dimensional representation** that captures the "essence" of (high-dimensional) data

Application in Machine Learning

Preprocessing, Regularization

- Compression, denoising, anomaly detection
- Reduce overfitting in supervised learning

Application in Statistics/Data analysis

Better understanding of the data

- descriptive/exploratory methods
- visualization (difficult to plot and interpret $> 3d!$)

Clustering: general goals

Main objective: construct a map f from \mathcal{D} to $\{1, \dots, K\}$ where K is a fixed number of clusters.

Motivations and Applications

- describe large masses of data in a simplified way,
- reveal structures, hidden causes,
- structure a set of knowledge,
- use of the groups in further processing, etc.

Careful! classification \neq clustering

- Classification presupposes the existence of classes
- Clustering labels only elements of the dataset
 - \rightsquigarrow no ground truth (no given labels)
 - \rightsquigarrow discovers a structure "natural" to the data
 - \rightsquigarrow not necessarily related to a known classification

Goals of MAP573

Comprehensive introduction to unsupervised learning

- **Dimension Reduction** (PCA and variants, nonlinear methods, a bit of AE)
- **Clustering** (partitioned based approaches, model-based approaches, spectral variants)
- + handling missing data

Practical skills for data/exploratory analysis

- by applying classical unsupervised approaches and some recent developments
- by performing complete analyses via projects
- for developing critical evaluation

Outline

Part 1: 2 sessions to get started with R

↪ 2 lectures, a deck of tutorials, 2 homework assignments

Part 2: 4 sessions on dimension reduction and clustering

↪ 4 lectures, 2 tutorials, 2 labs, 4 homework assignments

Part 3: Projects follow-up

↪ Apply/develop methods and skills seen in parts 1 and 2 on "real world" data sets

Practical information

Time – Tuesday, 1.30pm - 6pm (zoom essentially...)

Team – Florian Bourgey / Julien Chiquet / Élise Dumas

Grades

- 50% homework (6)
Rmd reports must be submitted during the first 6 weeks
- 50% projects
Report + talk (groups of 4 or 5 people)

Resources

- Website <https://jchiquet.github.io/MAP573>
- Moodle of MAP573 (material, forum, assignments, references)