Data anaysis and Unsupervised Learning

Introduction

MAP 573, 2020 – Julien Chiquet

École Polytechnique, Autumn semester, 2020

https://jchiquet.github.io/MAP573

Exploratory analysis of (modern) data set

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

Questions

Look for pattern or structure, 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 (much more variables than individual or $n \ll p$)
Redundancy	Many variables can carry the same information
Unsupervised	We do not 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 allel different against the population reference)

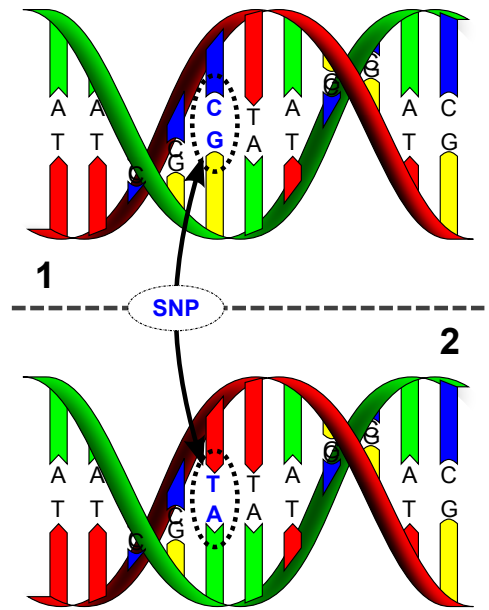


Figure: SNP (wikipedia)

Summarize 500,000 variables in 2

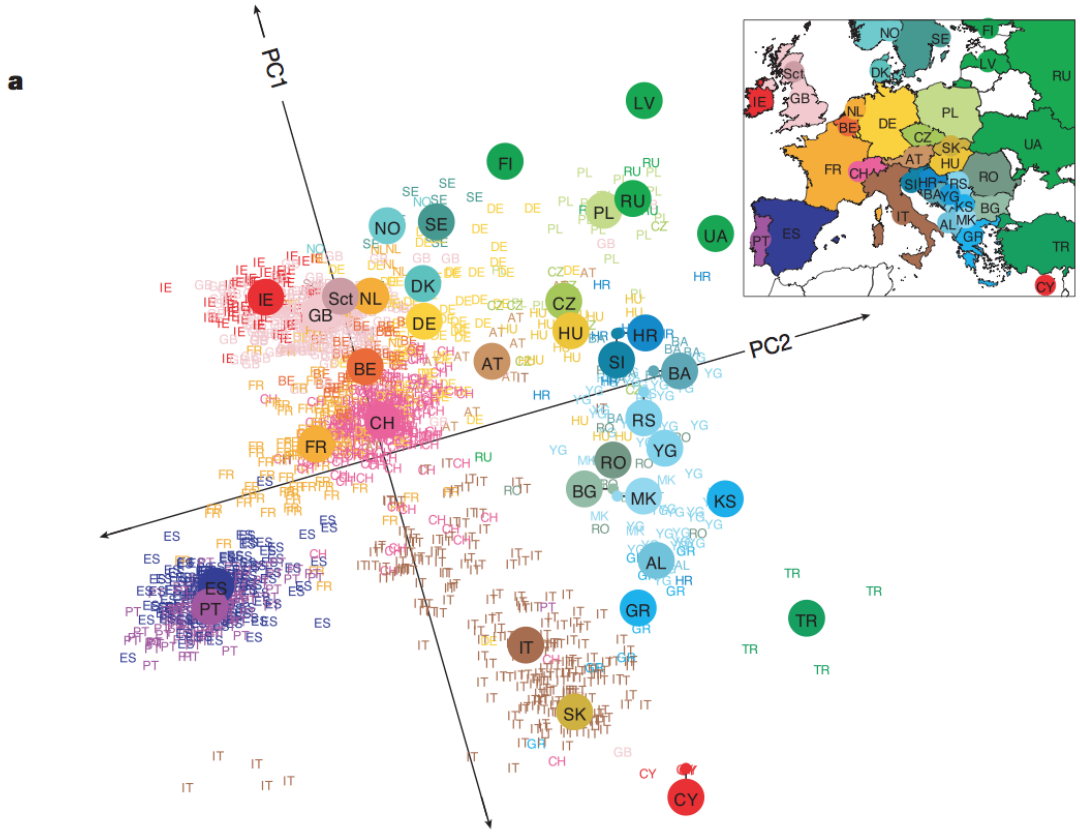
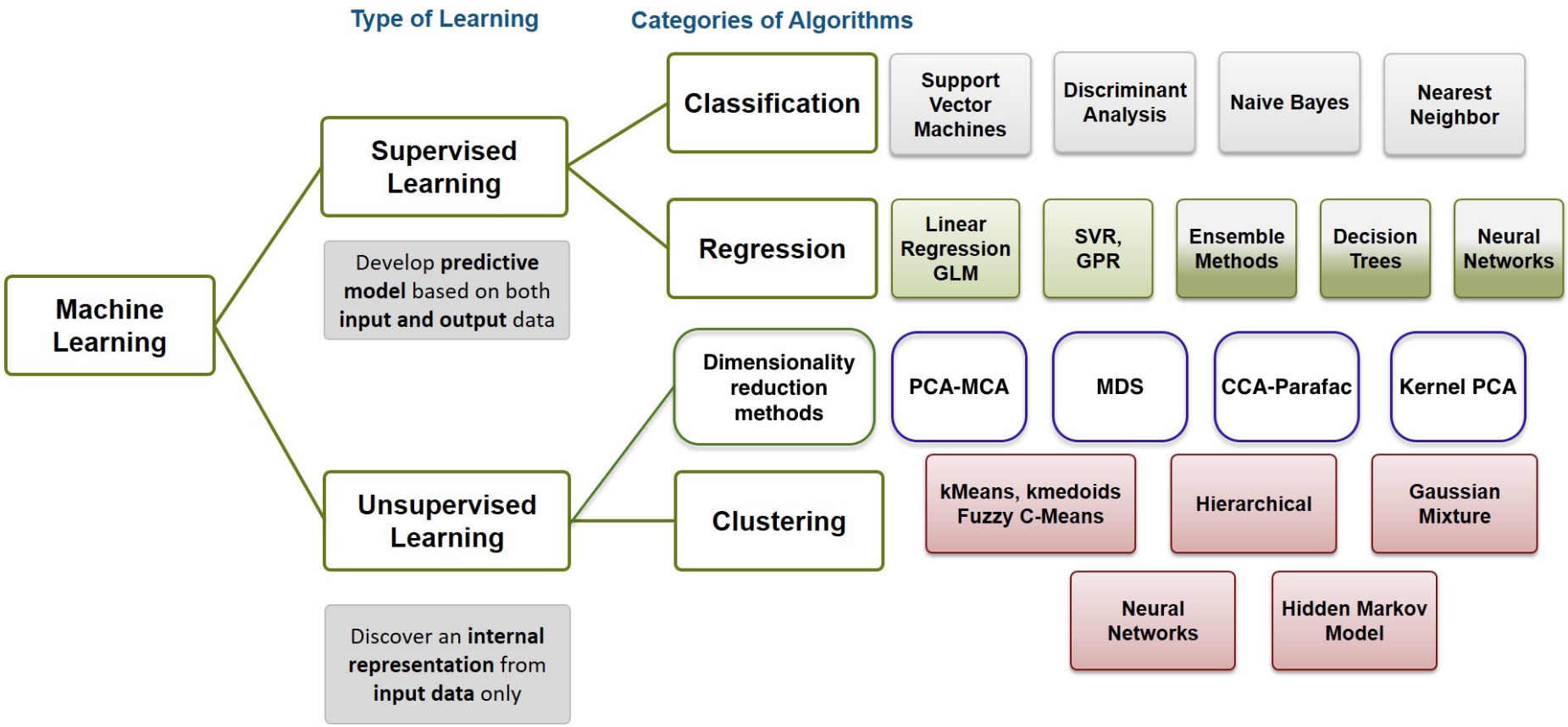


Figure: PCA output source: Nature "Gene Mirror Geography Within Europe", 2008

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



Supervised vs Unsupervised Learning

Supervised Learning

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

$\rightsquigarrow$ The goal is clear: predict y based on x (regression, classification)

Unsupervised Learning

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

$\rightsquigarrow$ The goal is less well defined, and *validation* is questionable

Supervised vs Unsupervised Learning

Supervised Learning

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

$\rightsquigarrow$ The goal is clear: predict y based on x (regression, classification)

Unsupervised Learning

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

$\rightsquigarrow$ 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 (improve performances)

Application in statistics and 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 $\{1, \ldots, K\}$ where K is a fixed number of clusters.

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

Motivations

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

Goals of MAP573

Comprehensive introduction to unsupervised learning

- Dimension Reduction
- Clustering
- $+$ handling missing data

Practical skills for data/exploratory analysis

- by applying classical unsupervised approaches and their recents developments
- by performing complete analyses via projects
- To develop critical evaluation

Outline

Part 1: 2 sessions to get started with R

$\rightsquigarrow$ 2 lectures set of tutorials, 2 homework assignments

Part 2: 4 sessions on dimension reduction and clustering

$\rightsquigarrow$ 4 lectures, 2 tutorials, 2 labs, 4 homework assignments

Part 3: Projects follow-up

$\rightsquigarrow$ 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)