Statistical learning: some principles and applications

Adeline LECLERCQ SAMSON

Massih-Reza AMINI
Some vocabulary in statistical learning

1. Learning a decision rule
   - Prediction of an (labeled) outcome based on observed variables
   - Prediction of the outcome from new observations

2. Clustering
   - Creation of groups of similar individuals / objects / variables
   - Research of patterns

3. Learning a model from the data
   - Physical / biological models
   - Estimation of the parameters, shape of the model

4. Learning associations, statistical tests
   - Correlation between variables
   - Interactions in a network

5. Data visualization
   - Exploration of the data, outliers detection, errors
   - Reduction of the dimension
Challenges in statistical learning

- **High dimension**
  - A lot of variables per individual
  - *Aim: learn the effect of all these variables* even when the number of individuals / units remains small

- **Repeated measures**
  - Repeated trials
  - Longitudinal measurements: several measures in time
  - *Aim: learn the variability in the process and the evolution in time*

- **Structured data**
  - Connected objects: a function measured with high frequency
  - Functional data
  - *Aim: learn a function* (infinite dimension) and not only some parameters
Main approaches in statistical learning

Parametric versus Non Parametric

- **Non parametric**
  - No prior knowledge of a model, of a relationship between variables
  - Objectives: learn the model, the distribution of the variables, the network

- **Parametric**
  - Models with meaningful parameters: chemical interactions, physical laws, biological transformation
  - Prior models: linear regression, reliability
  - Objectives: numerical optimization of a criteria

Challenges

- **Parsimony**: high dimension but maybe few significant signals
- **Optimization**: new technics to optimize complex criteria/space
- **Distributed calcul**: scalability of the solution
1. Decision rule, classification

- **Supervised learning**: class labels are provided
- **Aim**: learn a classifier to predict class labels of novel data

- **Statistical tools**
  - Logistic regression (parametric)
  - K-nearest neighbors (non-parametric)
  - Decision tree (non-parametric)
1. Decision rule

**Decision tree**

- **Decision rule**
  - $V_1 \geq 0.5$
  - $V_2 \geq 0.75$
  - $V_2 \geq 0.67$
  - $V_1 < 0.2$
  - $V_2 \geq 0.25$
  - $V_1 \geq 0.5$
  - $V_2 \geq 0.75$
  - $V_2 \geq 0.87$
  - $V_2 < 0.85$
  - $V_1 < 0.75$
  - $V_2 \geq 0.85$
  - $V_1 < 0.88$
  - $V_1 < 0.84$
  - $V_2 < 0.022$
  - $V_1 \geq 0.93$
  - $V_2 \geq 0.019$
  - $V_1 \geq 0.69$
  - $V_1 < 0.67$
  - $V_2 < 0.015$
  - $V_1 \geq 0.099$
  - $V_1 < 0.1$
  - $V_1 \geq 0.13$
  - $V_1 \geq 0.099$
  - $V_1 < 0.1$
  - $V_1 \geq 0.13$
  - $V_1 < 0.1$

---

A. Samson

Grenoble, 7/06/2018
1. Decision rule

K-nearest neighbors
Examples

- **Advanced personalized medicine**

  - Predict the best treatment from the knowledge of biomarkers measured at an initial clinical visit
  - Logistic regression and decision tree
1. Decision rule

Examples

- Manufacturing industry
  - High production costs
  - Some non-conformity at the end of the production process
  - Large number of sensors

- Decision tree to predict early in the production process which piece is likely to be not conform
- Reduce the proportion of non-conformity
2. Clustering

- Unsupervised learning: no class label is given

- Aim
  - Creation of groups of similar individuals / objects / variables;
  - Understanding the structure underlying the data

- Statistical tools
  - K-means (non-parametric)
  - Mixture model (parametric)
  - Bi-clustering, Stochastic Block Model (parametric)
2. Clustering

Examples

- **Consumption curves**
  - A curve per consumer
  - Prediction of the future consumption

- Clustering and identification of profiles by mixture model of functional data
- Prediction based on these clusters
Examples

- Consumption curves
  - A curve per consumer
  - Prediction of the future consumption

- Clustering and identification of profiles by mixture model of functional data
- Prediction based on these clusters
Bi-clustering
Stochastic block model
Examples

- **Autonomous Vehicle**
  - Precision of the position
  - Sequence of images

- Segmentation of the images by Stochastic Block Model
- Prediction of the position
3. Learning a model

- **Aim**
  - Fit the data with a model
  - Regression model

- **Statistical tools**
  - Differential equations
  - Point process
  - Estimation of the parameters (parametric)
    - Maximum likelihood
    - Bayesian
    - Penalization with high dimension
  - Estimation of the model (non-parametric)
    - Splines, Wavelets, Fourier
Examples

- Immunotherapy

- Efficacy of the treatment
- Optimization of the treatment, in terms of dose and time to treatment

A. Samson

Grenoble, 7/06/2018
Examples

- **Maintenance, reliability**
  - Maintenance optimization
  - Imperfect maintenance

- Virtual age modeling, point processes
- Optimization of the next preventive maintenance
4. Learning associations, statistical tests

**Aim**
- Correlation between variables,
- Learning communities and networks

**Statistical tools**
- Correlation tests, multiple tests
- Graphical models
4. Learning associations

Examples

- Genomics
  - Effect of the pollution on epigenetics and baby growth

- Associations tests and multiple testing
- Mediation to infer causality
4. Learning associations

Examples

- **Neurosciences**
  - Understanding the connexions in the brain
  - Longitudinal, functional data through EEG, MEG data
4. Learning associations

Statistical learning frameworks

Open source development frameworks available

Possible to use in industry
Take-Home message

Core-idea of statistical learning
- Variety of (industrial) problems
- Variety of statistical questions
- Variety of learning approaches
- Machine learning: decision rule, clustering
- Statistical learning: model learning, association/correlation

A strategy that is effective across different disciplines
- Health
- Environment
- Energy
- Marketing
- Manufacturing industry
Some directions of ongoing research

- Structured data, connected objects
- Social networks, interaction graph
- High dimension
- Optimization with constraints
- Distributed computation