SIERRA: Machine Learning

Francis Bach

Laboratoire d’Informatique de l’Ecole Normale Supérieure

Scientific Board of Département d’Informatique, ENS
May 2011
Composition of the team

- Machine learning researchers from WILLOW

- **4 researchers**
  - Sylvain Arlot, CNRS
  - Jean-Yves Audibert, Ecole des Ponts
  - Francis Bach, INRIA
  - Guillaume Obozinski, INRIA

- **3 Post-docs** (Nicolas Le Roux, Mark Schmidt, Simon Lacoste-Julien)

- **11 PhD students** (Louise Benoit, Florent Couzinie-Devy, Loic Février, Edouard Grave, Toby Hocking, Rodolphe Jenatton, Armand Joulin, Augustin Lefèvre, Bamdev Mishra, Anil Nelakanti, Matthieu Solnon)
Machine learning
Computer science and applied mathematics

- Modelisation, prediction and control from training examples

- Theory
  - Analysis of statistical performance

- Algorithms
  - Numerical efficiency and stability

- Applications
  - Computer vision, bioinformatics, neuro-imaging, text, audio
Machine learning
Scientific context

• Proliferation of digital data \( (> 10^{21} \text{ bytes}, +50\% \text{ per year}) \)
  – Personal data (multimedia)
  – Industrial (multimedia, sensors)
  – Scientific (multimedia, sensors, genomics, etc.)

• Need for automated processing of massive data
  – Supervised
  – Unsupervised
Scientific objectives - SIERRA tenet
- Machine learning does not exist in the void
- Specific domain knowledge must be exploited
Scientific objectives - SIERRA tenet
- Machine learning does not exist in the void
- Specific domain knowledge must be exploited

• Scientific challenges
  – Fully automated data processing
  – Incorporating structure
  – Large-scale learning
Scientific objectives - SIERRA tenet
- Machine learning does not exist in the void
- Specific domain knowledge must be exploited

• Scientific challenges
  – Fully automated data processing
  – Incorporating structure
  – Large-scale learning

• Scientific objectives
  – Supervised learning
  – Parsimony
  – Optimization
  – Unsupervised learning
Scientific objectives - SIERRA tenet

- Machine learning does not exist in the void
- Specific domain knowledge must be exploited

**Scientific challenges**

- Fully automated data processing
- Incorporating structure
- Large-scale learning

**Scientific objectives**

- Supervised learning
- Parsimony
- Optimization
- Unsupervised learning

**Interdisciplinary collaborations**

- Computer vision
- Bioinformatics
- Neuro-imaging
- Text, audio, natural language
Scientific objectives - Supervised learning

• Data \((x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \ i = 1, \ldots, n\)

• **Goal**: predict \(y \in \mathcal{Y}\) from \(x \in \mathcal{X}\), i.e., find \(f : \mathcal{X} \rightarrow \mathcal{Y}\)

• Empirical risk minimization

\[
\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|^2
\]

Data-fitting + Regularization

• **Scientific objectives**:
  
  – Studying generalization error (J.-Y. Audibert, S. Arlot)
  
  – Improving calibration (S. Arlot, F. Bach)
  
  – Choosing appropriate representations (F. Bach, G. Obozinski)

  – **Two main types of norms**: \(l_2\) vs. \(l_1\)
Scientific objectives - Parsimony and $\ell_1$-norm

- Data $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}, i = 1, \ldots, n$

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \sum_{j=1}^{p} |w_j|$$

Data-fitting + Regularization

- At the optimum, $w$ is in general **sparse**
1. **Consistent support estimation**: the support is correctly estimated if and only if there are *low correlations* between variables.

- **Goal**: Fix inconsistent situations (most frequent).
Scientific objectives - Parsimony and $\ell_1$-norm

1. **Consistent support estimation**: the support is correctly estimated if and only if there are *low correlations* between variables.
   - **Goal**: Fix inconsistent situations (most frequent).

2. **Exponential number of variables**: under appropriate assumptions, good predictions are possible as long as $\log p = O(n)$
   - **Goal**: Deal with $10^{80}$ variables in practice.
Scientific objectives - Parsimony and $\ell_1$-norm

1. **Consistent support estimation**: the support is correctly estimated if and only if there are *low correlations* between variables.
   - **Goal**: Fix inconsistent situations (most frequent).

2. **Exponential number of variables**: under appropriate assumptions, good predictions are possible as long as $\log p = O(n)$
   - **Goal**: Deal with $10^{80}$ variables in practice.

3. **Beyond the cardinality of the support**: in many application domains, special patterns are preferable.
   - **Goal**: theory, algorithms and applications of *structured sparsity*. 
Structured sparsity

- Sparsity-inducing behavior depends on “corners” of unit balls
Scientific objectives - Optimization

• Key concept in machine learning

• **Convex optimization** (F. Bach, G. Obozinski)
  – Use structure, design algorithms for large scale
  – First-order methods
  – Convex relaxation of combinatorial optimization problems

• **Online optimization** (J.-Y. Audibert, F. Bach)
  – Dealing with millions of observations

• **Multi-armed bandits** (J.-Y. Audibert)
  – Optimization and decision theory
  – Trade-off between exploration and exploitation
Scientific objectives - Optimization

Convex relaxation of combinatorial problems

- **Classical example**: MAXCUT relaxed into semidefinite programming (SDP) optimization problem (Goemans and Williamson, 1994)

- **Discriminative clustering**: find labels $y \in \{-1, 1\}^n$ that minimize

$$\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|^2$$

- Can be relaxed into a SDP (Xu et al., 2004, Bach and Harchaoui, 2007, Joulin et al., 2010)
- Naive computational complexity $= O(n^7)$

- **Efficient manifold optimization over low-rank matrices** (Journée et al., 2009)
Scientific objectives - Optimization

First-order methods for large-scale optimization

• Massive data: large $n$, large $p$, large $k$
  - $n =$ number of observations, $p =$ size of observations
  - $k =$ number of tasks
  - Can only learn with (very) few simple passes over the data
  - Ideal complexity: $O(np + nk)$

• First-order methods (gradient-based)
  - Bad optimization methods..., but good learning methods
  - Stochastic approximation

• Challenge: avoid $O(n^2)$ in nonparametric (nonlinear) problems
Scientific objectives - Optimization

Bandit optimization (J.-Y. Audibert)

• **Goal**: Given set of actions \( A \) and sequential decisions \( a_t \in A \), maximize total reward \( \sum_{t=1}^{T} \mu(a_t) \)
  
  – Reward function \( \mu \) unknown
  – Stochastic vs. adversarial environment
  – Multiple applications

• **Exploration-exploitation dilemma**

• **Scientific objectives**
  
  – Design of robust policies
  – Analysis of more realistic situations
  – Efficient policies in structured situations
Scientific objectives - Unsupervised learning

- Data $x_i \in \mathcal{X}$, $i = 1, \ldots, n$. **Goal:** “Find” structure within data
  - Discrete: clustering (e.g., K-means)
  - Low-dimension: principal component analysis
Scientific objectives - Unsupervised learning

- Data $x_i \in \mathcal{X}, i = 1, \ldots, n$. **Goal:** “Find” structure within data
  - Discrete: clustering (e.g., K-means)
  - Low-dimension: principal component analysis

- **Matrix factorization** (F. Bach, G. Obozinski):
  
  \[
  X = DA
  \]

  - Structure on $D$ and/or $A$
  - Algorithmic and theoretical issues

- **Statistical tests** (S. Arlot, F. Bach):
  - Theoretical analysis and applications
Structured sparse PCA (Jenatton et al., 2010)

• **Goal**: Decompose faces as linear combinations of dictionary elements
Structured sparse PCA (Jenatton et al., 2010)

• Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion
Structured sparse PCA (Jenatton et al., 2010)

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion
Application to face databases

- Quantitative performance evaluation on classification task
Change point detection (Arlot and Celisse, 2010)

- Segmentation of the mean of heteroscedastic data via cross-validation
  - Empirical risk minimization is not consistent
  - Application to comparative genomic hybridization (CGH) data
Change point detection (Arlot and Celisse, 2010)

- Segmentation of the mean of heteroscedastic data via cross-validation
  - Empirical risk minimization is not consistent
  - Application to comparative genomic hybridization (CGH) data
Application domains - Computer vision
Co-segmentation (Joulin et al., 2010)
Application domains - Computer vision
Digital zooming (Couzinie-Devy et al., 2010)
Application domains - Computer vision
Digital zooming (Couzinie-Devy et al., 2010)
Application domains - Audio processing

Source separation (Lefèvre et al., 2010)
Application domains - Audio processing

Musical instrument separation (Lefèvre et al., 2010)

- Unsupervised source separation with group-sparsity prior
  - Top: mixture
  - Left: source tracks (guitar, voice). Right: separated tracks.
Application domains - Bioinformatics

- Matching of graphs of proteins (Zaslavskiy et al., 2009)
- Metastasis prediction from microarray data (Jacob et al., 2009)

- Biological pathways
- Dedicated sparsity-inducing norm for better interpretability and prediction
Application domains - Neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

- “Brain reading”: prediction of (seen) object size

- Multi-scale activity levels through hierarchical penalization
Application domains - Neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization
Application domains - Neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization
Application domains - Neuro-imaging

Structured sparse PCA on resting state activity
(Jenatton et al., 2010)
Application domains - Natural language processing

- Just started a collaboration with NLP group from INRIA/Paris 7 (G. Obozinski)
Teaching

- Essential to a joint INRIA/ENS/CNRS team
  - New curriculum “Math-Info”, first year, ENS
  - Joint class between mathematics and CS department

- Prime access to PhD students
  - Masters M2 MVA, ENS Cachan (2 courses)
  - Masters M2 Probability and Statistics, Université Paris-Sud

- Summer schools and tutorials

- Cours Peccot, Collège de France (S. Arlot, 2011)
Grants

- **European Research Council (ERC):** 1.5 MEuros
  - Starting investigator grant (PI: F. Bach, 2009-2014)

- **Agence Nationale de la Recherche (ANR):** 200 KEuros
  - Projet “blanc” MGA (Mines, Ponts, Telecom, ENS, PI: F. Bach)
  - Projet “blanc” EXPLORA (**SEQUEL, CLASSIC, WILLOW**)
  - Projet “jeunes chercheurs” DETECT (PI: S. Arlot)

- **Digiteo:** 100 KEuros
  - Projet BIOVIZ (INRIA, Institut Curie, PI: F. Bach)
International visibility

• Publications (2007-2010)
  – Conferences (NIPS:12, ICML:8, COLT:3, ECCV:1, CVPR:7, ICCV:3)
  – Journals (JMLR:9, PAMI:1, Ann. Stat:6, EJS:2)

• Editorial activities
  – Journals (editor): JMLR, PAMI, SIAM Imaging

• Workshop organization (NIPS, ICML, Banff)

• European Research Council (ERC) grant
Scientific partners

- Mostly academic institutions
  - Collaborations with industry pursued in an opportunistic manner

- France
  - ENS (CS: Willow, Math, CogSci)
  - Mines, Ponts, Telecom, Agro, Paris-sud, Paris 6, Univ. Lille
  - INRIA Grenoble, INRIA Saclay, INRIA Rennes
  - Institut Curie, Institut Pasteur
  - Xerox Grenoble, EDF R&D, EADS

- Abroad
  - U.C. Berkeley, Princeton Univ., C.M.U.
  - Oxford Univ., University College London, Univ. Liège
  - Technicolor
Future plans

• Core machine learning

• Ongoing interdisciplinary collaborations
  – Computer vision
  – Bioinformatics

• Recent interdisciplinary collaborations
  – Neuro-imaging
  – Audio processing, natural language processing

• Co-advised students

• Recruiting permanent researchers
  – Algorithmic machine learning
SIERRA
Statistical machine learning and parsimony