Mathematical Methods for Data Analysis

Massimiliano Pontil

Istituto Italiano di Tecnologia
and
Department of Computer Science
University College London

Learning from data

- Let μ be a probability measure on a set Z
- μ is unknown, but can sample from it

$$z_1, \ldots, z_m \sim \mu$$

- Goal: learn "properties" of μ from the data:
  - ◇ Density estimation
  - ◇ Study "low dimensional" representation of the data
  - ◇ Supervised learning (prediction): $Z = X \times Y$

Supervised learning

$Z = X \times Y$, given data $(x_1, y_1), \ldots, (x_n, y_n) \sim \mu$, find

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2}_{\text{empirical error}}$$

Three key problems:

- Function representation/approximation: which $\mathcal{F}$?

  (Typically $\mathcal{F} = \{\Omega(f) \leq \alpha\}$ with Ω e.g. a norm in a function space

- Numerical optimization: iterative schemes to find $\hat{f}$
  (gradient descent, proximal-gradient methods, stochastic optimization)

- Statistical analysis: derive high probability bound

$$\mathbb{E}\big(y - \hat{f}(x)\big)^2 \leq \min_{f \in \mathcal{F}} \mathbb{E}\big(y - f(x)\big)^2 + \epsilon(n, \delta, \mathcal{F})$$

Regularization

- Difficulty: high dimensional data / complex tasks
- Increasing need for methods which can impose sophisticated form of prior knowledge
- General approach in machine learning and statistics:

$$\underset{f}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda \underbrace{\Omega(f)}_{\text{regularizer}}$$

- Three predominant assumptions:
  - **smoothness:** Ω is the norm in a RKHS
  - **sparsity:** non-differentiable penalties (e.g. ℓ_1 norm)
  - **shared representations:** needs multiple "tasks"

Regularization in reproducing kernel Hilbert spaces

[Aronszajn 1950, Wahba 1990, Cucker & Smale 2002, Schölkopf & Smola, 2002,...]

- Choose a *feature map* $\phi : X \to \ell_2$ and solve:

$$\underset{w \in H}{\text{minimize}} \sum_{i=1}^{n} \left(\langle w, \phi(x_i)\rangle - y_i\right)^2 + \lambda \|w\|_2^2$$

- Regularizer favors smooth functions, e.g. small Sobolev norms

- Define the *kernel function* $K(x, x') = \langle \phi(x), \phi(x')\rangle$

  e.g. the Gaussian: $k(x, x') = e^{-\beta \|x - x'\|^2}$

- Solution has the form $\hat{f}(x) = \sum_{i=1}^{n} c_i K(x_i, x)$

Linear regression and sparsity

[Bickel, Ritos, Tsybakov, 2009, Buhlmann & van de Geer, 20012, Candes and Tao, 2006]

Consider the model

$$y = Xw^* + \xi$$

- $y \in \mathbb{R}^n$ is a vector of observations
- X is a prescribed $n \times d$ data matrix
- $\xi \in \mathbb{R}^m$ is a noise vector (e.g. i.i.d. Gaussian)
- $w^* \in \mathbb{R}^d$ is assumed to be **sparse**

Goal:

- estimate w^* (or its sparsity pattern or its prediction error) from y
- efficient computational schemes for:

$$\underset{w \in H}{\text{minimize}} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \Omega(w)$$

Regularizers for structured sparsity
[Maurer & P., 2012, Micchelli, Morales, P., 2013, McDonald, P. Stamos, 2015]

Exploit additional knowledge on sparsity pattern of w^*:

$$\Omega(w) = \sqrt{\inf_{\theta \in \Theta} \sum_{i=1}^{d} \frac{w_i^2}{\theta_i}}$$

- Constraint set $\Theta \subseteq R^d_{++}$, convex and bounded
- Example: if $\Theta = \{\theta > 0 : \sum_{i=1}^n \theta_i \leq 1\}$ yields the ℓ_1 norm
- Focus on:
  - efficient optimization methods (e.g. proximal gradient methods)
  - statistical estimation bounds (e.g. using Rademacher averages)
  - ongoing applications in neuroimaging

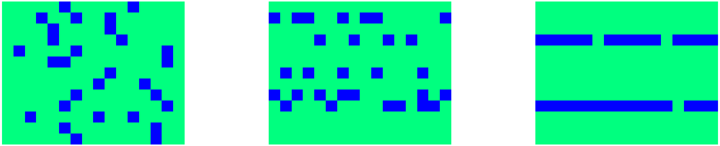
Multi-task learning

$$\min_{w_1,\dots,w_T} \frac{1}{T}\sum_{t=1}^{T} \underbrace{\|X_t w_t - y_t\|^2}_{\text{error task } t} + \lambda \underbrace{\Omega(w_1,\dots,w_T)}_{\text{joint regularizer}}$$

- X_t: $n \times d$ data matrix
- Typical scenario: many tasks but only *few examples* per task: $n \ll d$
- If the tasks are related, learning them **jointly** should perform better than learning each task *independently*
- Several applications: computer vision, neuroimaging, NLP, robotics user modeling, etc.

Multitask regularizers

- *Quadratic:* encourage similarities between tasks (e.g. small variance)
  Can be made more general using RKHS of vector-valued functions [Caponnetto et al., 2008; Carmeli, De Vito, Toigo, 2006]
- *Row sparsity:* few common variables (provably better than Lasso [Lounici, P. Tsybakov, van de Geer, 2011])



- *Spectral:* few common linear features (low rank matrix) [Srebro & Shraibman, 2005, Argyriou, Evgeniou, P. 2006; Maurer and P. 2013]

Matrix completion

- Learn a matrix from a subset of its entry (possibly noisy); see e.g. [Srebro 2004; Candes & Tao, 2008]
- Special case of the above when raws of X_t are elements of the standard basis $\{e_1, \ldots, e_d\}$

$$\min_W \sum_{(i,t)\in S} (Y_{i,t} - W_{i,t})^2 + \lambda\Omega(W)$$

- Ongoing project on online (binary) matrix completion

Lifelong learning

- Human intelligence relies on transferring knowledge learned from previous tasks to learn new tasks
- Online approach: see one tasks at the time, train on past tasks, test on next task
- Interactive learning, e.g. active learning, choose which entries to sample, choose which tasks to learn next
- Nonlinear extension: $\phi : X \to \ell_2$ a prescribed mapping

$$\underset{w_1,\dots w_T \in \ell_2}{\text{minimize}} \sum_{i=1}^{n} \sum_{t=1}^{T} \ell(y_{ti}, \langle w_t, \phi(x_{ti}\rangle) + \lambda \|[w_1, \dots, w_n]\|$$

Vector-valued learning

Choose a class of vector-valued functions:

$$\mathcal{F} \circ \mathcal{G} = \left\{ x \in \ell_2 \mapsto f\,(g\,(x)) \in \mathbb{R}^T : f \in \mathcal{F}, g \in \mathcal{G} \right\},$$

where $g : H \to \mathbb{R}^K$, and $f : \mathbb{R}^K \to \mathbb{R}^T$, found by the method

$$\underset{f \in \mathcal{F},\ g \in \mathcal{G}}{\text{minimize}} \ \sum_{i=1}^{N} \ell(f \circ g(x_i), y_i) + \Omega(f, g)$$

- Includes neural networks with shared hidden layers ("deep nets")
- Loss function includes multitask and multi-category learning
- Includes nuclear or factorization norms [Jameson, 1987]
- Current focus on Rademacher complexity bounds:

$$\frac{1}{N} \mathbb{E} \sup_{f,g} \sum_{i=1}^{N} \epsilon_i \ell(f(x_i), y_i)$$

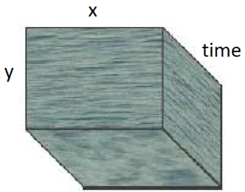
Multilinear models

[Gandy eta la. 2011, Kolda & Bader, 2009,...]

General problem: Learning a tensor from a set of linear measurements

Examples:

- Tensor completion



- Video denoising/completion
- 3D scanning denoising/completion
- Context-aware recomendation
- Entities-relationships learning (NLP)

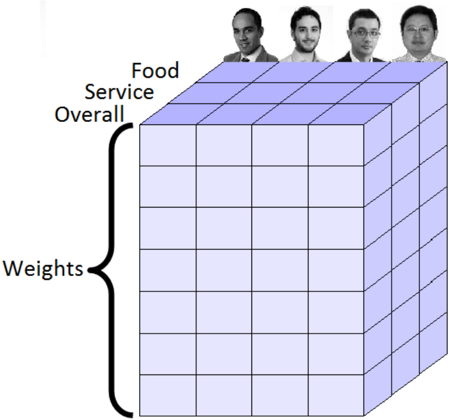
- Multilinear multitask learning

Multilinear multitask learning

[Romera-Paredes et al. 2013]

Tasks are be referenced by multiple indices

E.g: (, Food)



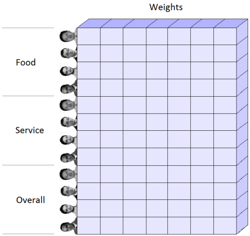
Problem modelling

Want to encourage low rank tensors

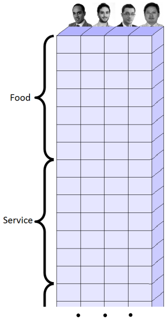
$$\underset{\mathcal{W}}{\operatorname{argmin}}\, E(\mathcal{W}) + \frac{\gamma}{N}\sum_{n=1}^{N} \operatorname{rank}\left(W_{(n)}\right)$$

$W_{(n)}$ is the n-th matricization of the tensor, e.g.:

$W_{(1)}$



$W_{(3)}$



Research interests / PhD projects

- Supervised learning: support vector machines and reproducing kernels
- Study of regularizers for structured sparsity
- Multitask and transfer learning: study assumptions on task relatedness (e.g. learning shared representations)
- Online learning and mistake bounds - connection to lifelong learning
- Statistical learning theory (e.g. study of Rademacher bounds) for competitive vector-valued function classes
- Multilinear models: modelling low rank tensors and convex relaxations
- Sparse coding / dictionary learning (not covered today, ask me if interested)
- Transfer in reinforcement learning (not covered today, ask me if interested)

Plan

- Focus on a specific project for the first 6 months
- Converge to a PhD topic within 9 months
- Can propose your own project
- Interact with postdocs in the group and colleagues at DIMA/DIBRIS/IIT
- Reading groups on specific topics
- 1 year abroad (UCL or to visit other collaborators)

Collaborators (mostly ongoing)

- Mark Herbster (UCL) *online learning*
- Theodoros Evgeniou (INSEAD) *user modelling*
- Cecilia Mascolo (Cambridge) *user modelling*
- Nadia Bianchi-Berthouze (UCL) *affective computing*
- Janaina Mourau-Miranda (UCL) *ML in neuroimaging*
- Alexandre Tsybakov (ENSAE Paris Tech) *statistical estimation*
- Andreas Maurer (Munich) *statistical learning theory*
- Sara van de Geer (ETH Zürich) *sparse estimation*
- Patrick Combettes (Paris 6) *numerical optimization*
- Rapahel Hauser (Oxford) *numerical optimization*
- Charles Micchelli (SUNY Albany) *kernel methods, mathematics*