Overview

1. Some notation

2. The problems: regression, classification and more

3. Past contributions: regularized kernel methods, inverse problems, feature selection

A common language: statistical learning

Supervised setting. The ingredients:

- input and output spaces $\mathcal{X}$ and $\mathcal{Y}$, often $\mathbb{R}^d$ and $\mathbb{R}$;
- a collection of examples (training set)

$$Z = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} | i = 1, \ldots, n\},$$

drawn i.i.d. from an unknown joint distribution $\rho(x, y)$;

- a convex loss function $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such as the squared loss

$$\ell(y, y') = \|y - y'\|_2^2$$

or the hinge loss

$$\ell(y, y') = \max\{0, 1 - yy'\} = (1 - yy')_+$$
A common language: statistical learning

Given a training set $Z$, we are looking for a function $f_Z : \mathcal{X} \longrightarrow \mathcal{Y}$ with a small expected risk

$$\mathcal{E}(f_Z) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, f_Z(x)) d\rho(x, y)$$

- $f_Z$ is called **estimator**
- $Z \mapsto f_Z$ is the **learning algorithm**
Problems: regression

Scalar or vectorial regression: \( \mathcal{Y} \subseteq \mathbb{R}^k \)

Problems: classification

Binary classification: $\mathcal{Y} = \{-1, +1\}$

Problems: variable selection

**Feature and variable selection**: classification or regression under an additional constraint requiring a **sparse** solution

Problems

Scalar or vectorial regression: $\mathcal{Y} \subseteq \mathbb{R}^k$

Binary classification: $\mathcal{Y} = \{-1, +1\}$

Feature and variable selection: classification or regression under an additional constraint requiring a sparse solution

Unsupervised learning (e.g. data clustering, matrix completion, dictionary learning): no $\mathcal{Y}$
Regularization of the Empirical Risk

The empirical error is the functional

\[ \mathcal{E}_Z (f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \]
Regularization of the Empirical Risk

The regularized empirical error is the functional

\[ \mathcal{E}_Z^\tau(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \tau \mathcal{R}(f) \]

where

- \( \mathcal{R} \) is a convex penalty term, weighted by \( \tau > 0 \)
- \( f \in \mathcal{H} \) (hypothesis space), often a RKHS

\[ \mathcal{H}_K = \{ f : \mathcal{X} \to \mathcal{Y} | f(x) = \langle f, K(\cdot, x) \rangle \} \]

induced by a symmetric, positive semi-definite function \( K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) (a kernel);

Learning attempts at solving

\[ f_Z^\tau \in \arg \min_{f \in \mathcal{H}} \mathcal{E}_Z^\tau(f) \]
Regularized Kernel Methods

In fact one can show that given

\[ \ell(y, y') \text{ convex, } f \in \mathcal{H}_K \text{ and } R(f) = \| f \|^2_{\mathcal{H}} \]

then

\[ f(x) = \sum_{i=1}^{n} \beta_i K(x, x_i) \text{ and } \| f \|^2_{\mathcal{H}} = \beta^T K_n \beta \]

Different choices of \( \ell \) yields different recipes for \( \beta \).

\[ \ell(y, y') = (1 - yy')_+ \] leads to Support Vector Machines (SVM).


- De Vito et al., 2004. Some properties of regularized kernel methods.
Learning as an Inverse Problem

The regularized least squares problem

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \tau \| f \|_{\mathcal{H}}^2$$

is the Tikhonov regularization of $S_n f = y$ where $S_n$ is the sampling operator defined by $(S_n f)_i = f(x_i)$.

This connection allowed us to use theoretical results and algorithms from the theory of inverse problems, e.g. spectral algorithms.


Variable Selection

Problem setting

- **Assumption**: there exists an optimal function $f^*$ depending on a small number of input variables
- **Goal**: detect the relevant variables
Variable Selection

Problem setting

- **Assumption**: there exists an optimal function $f^*$ depending on a small number of input variables
- **Goal**: detect the relevant variables

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \tau \|\beta\|_0$$

Theorem (Donoho ’01; Candes-Romberg-Tao ’06)

\(\ell_1\) minimization is equivalent to \(\ell_0\) minimization, i.e. performs variable selection
Variable Selection

Problem setting

- **Assumption**: there exists an optimal function $f^*$ depending on a small number of input variables
- **Goal**: detect the relevant variables

\[
\min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \tau \|\beta\|_1
\]

Although $\ell_1$-norm is **non-differentiable**, the functional is **convex** and iterative algorithms have been developed to solve such problems (proximal methods).
Feature Selection

The approach can be extended to generalized linear models

\[ f_\beta = \sum \beta_\gamma \Phi_\gamma : \]

\[
\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\beta(x_i))^2 + \tau \|\beta\|_1
\]


Our current favorite topics

Nonlinear Feature Selection

See next slides

Dictionary Learning

See Matteo’s talk

Matrix Completion

$$\min_Z \|P(Z - X)\|_F^2 + \tau \text{Tr} \left[ \sqrt{X^T X} \right]$$
Nonlinear Variable Selection

$$\min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\beta(x_i))^2 + \tau \| \beta \|_1, \quad \text{with } f = \sum \beta_\gamma \Phi_\gamma.$$  

- $\ell_1$ regularization selects the relevant components w.r.t. $\{ \Phi_\gamma \}$
- we want to select the relevant variables

Example

$\mathcal{H} = \{ f(x, y) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 \}$. 

A new penalty

Idea: If a function does not depend on a variable, then the corresponding partial derivative is 0 so that its norm is 0.

We propose

$$\mathcal{R}_n(f) = \sum_{j=1}^{d} \|\hat{D}_j(f)\|_n := \sum_{j=1}^{d} \sqrt{\frac{1}{n} \sum_{i=1}^{n} |\partial_j f(x_i)|^2},$$

where $\hat{D}_j f = ((\partial_j f)(x_1), \ldots, (\partial_j f)(x_n))$.

If $f(x) = \beta \cdot x$, then $\partial_j f(x) = \beta_j$, and

$$\mathcal{R}(f) = \|\beta\|_1.$$
Comparison with the total variation

\[ TV(f) = \sum_{i,j=1}^{n} \sqrt{\partial_1 f(i,j)^2 + \partial_2 f(i,j)^2}, \quad f \in \mathbb{R}^{n \times n} \]

\[ \mathcal{R}_n(f) = \sqrt{\frac{1}{n} \sum_{i,j=1}^{n} \partial_1 f(i,j)^2} + \sqrt{\frac{1}{n} \sum_{i,j=1}^{n} \partial_2 f(i,j)^2} \]

Example

\[ TV(f) \quad \mathcal{R}(f) \quad TV(g) \quad \mathcal{R}(g) \]
References