Our Approach to Learning

Curzio Basso
SlipGURU, DISI, Università di Genova
curzio.basso@disi.unige.it

June 8, 2010

Overview

1. **Some notation**
2. **The problems**: regression, classification and more
3. **Past contributions**: regularized kernel methods, inverse problems, feature selection
4. **Current interests**: nonlinear feature selection, dictionary learning, matrix completion

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*)

$$Z = \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y} | i = 1, \ldots, n\},$$

  drawn i.i.d. from an unknown **joint** distribution $\rho(\mathbf{x}, \mathbf{y})$;
- a convex loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such as the squared loss

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

  or the hinge loss

$$\ell(y, y') = \max\{0, 1 - yy'\} = (1 - yy')_+$$

A common language: statistical learning

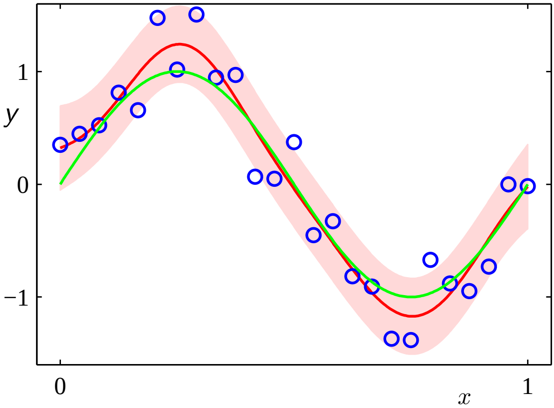
Given a training set Z, we are looking for a function $f_Z : \mathcal{X} \longrightarrow \mathcal{Y}$ with a small expected risk

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

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

Problems: regression

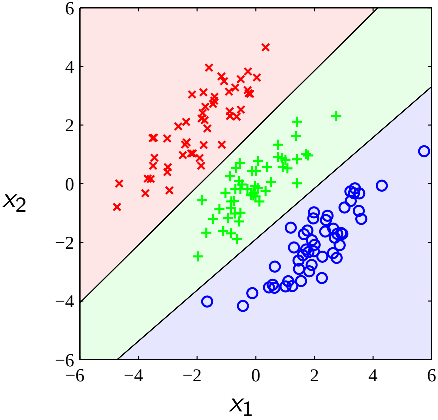
Scalar or vectorial **regression**: $\mathcal{Y} \subseteq \mathbb{R}^k$



kindly stolen from C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Problems: classification

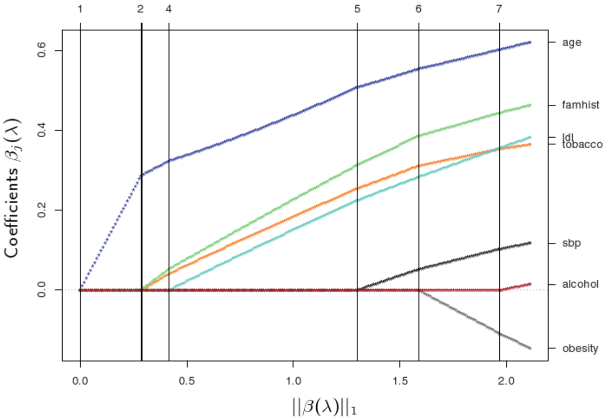
Binary **classification**: $\mathcal{Y} = \{-1, +1\}$



kindly stolen from C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Problems: variable selection

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



kindly stolen from T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2009.

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

$$\mathcal{E}_Z(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i))$$

Regularization of the Empirical Risk

The regularized empirical error is the functional

$$\mathcal{E}_Z{}^{\tau}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)) + \tau\mathcal{R}(f)$$

where

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

  $$\mathcal{H}_K = \{f : \mathcal{X} \to \mathcal{Y} | f(\mathbf{x}) =< f, K(\cdot, \mathbf{x}) >\}$$

  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

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

then

$$f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i K(\mathbf{x}, \mathbf{x}_i) \;\text{ and }\; \|f\|^2_{\mathcal{H}} = \boldsymbol{\beta}^T \mathbf{K}_n \boldsymbol{\beta}$$

Different choices of ℓ yields different recipes for $\boldsymbol{\beta}$.
$\ell(y, y') = (1 - yy')_+$ leads to *Support Vector Machines* (SVM).

- Pontil and Verri, 1998. Properties of Support Vector Machines. *Neural Computation*, 10:977–996.
- De Vito et al., 2004. Some properties of regularized kernel methods. *J. Mach. Learning Res.*, 5:1363-1390.

Learning as an Inverse Problem

The regularized least squares problem

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

is the Tikhonov regularization of $S_n f = \mathbf{y}$ where S_n is the sampling operator defined by $(S_n f)_i = f(\mathbf{x}_i)$.

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

- De Vito et al., 2005. Learning from Examples as an Inverse Problem. *J. Mach. Learning Res.*, 6:883–904.
- Lo Gerfo et al., 2008. Spectral Algorithms for Supervised Learning. *Neural Computation*, 7:1873–1897.

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

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

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

ℓ_1 minimization is equivalent to ℓ_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

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \tau \|\boldsymbol{\beta}\|_1$$

Although ℓ_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 $f_{\boldsymbol{\beta}} = \sum \beta_\gamma \Phi_\gamma$:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f_{\boldsymbol{\beta}}(\mathbf{x}_i))^2 + \tau \|\boldsymbol{\beta}\|_1$$

- Destrero et al., 2009. A sparsity-enforcing method for learning face features. *IEEE Transactions on Image Processing*, 18:188–201.
- Destrero et al., 2009. Feature selection for high dimensional data. *Computational Management Science*, 6:25–40.

Our current favorite topics

Nonlinear Feature Selection

See next slides

Dictionary Learning

See Matteo's talk

Matrix Completion

$$\min_{\mathbf{Z}} \|P(\mathbf{Z} - \mathbf{X})\|_F^2 + \tau \mathrm{Tr}\left[\sqrt{\mathbf{X}^T\mathbf{X}}\right]$$

Nonlinear Variable Selection

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

- ℓ_1 regularization selects the relevant components w.r.t. $\{\Phi_\gamma\}$
- we want to select the relevant variables

Example

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

- Rosasco et al., 2009. Iterative Projection methods for structured sparsity regularization. *MIT-CSAIL-TR-2009-050-CBCL-282*.
- Mosci et al., 2010. A Regularization Approach to Nonlinear Variable Selection. *Proc. of AISTATS 2010*.

A new penalty

Idea: If a function does not depend on a variable, than 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

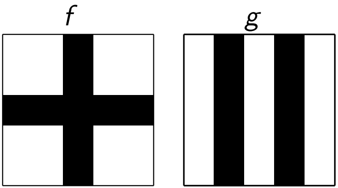
$$\mathcal{R}(f) = \|\beta\|_1.$$

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



$$\begin{array}{ccc} TV(f) & < & TV(g) \\ \mathcal{R}(f) & > & \mathcal{R}(g) \end{array}$$

References

- Pontil and Verri, 1998. Properties of Support Vector Machines. *Neural Computation*, 10:977–996.
- De Vito et al., 2004. Some properties of regularized kernel methods. *J. Mach. Learning Res.*, 5:1363-1390.
- De Vito et al., 2005. Learning from Examples as an Inverse Problem. *J. Mach. Learning Res.*, 6:883–904.
- Lo Gerfo et al., 2008. Spectral Algorithms for Supervised Learning. *Neural Computation*, 7:1873–1897.
- Destrero et al., 2009. A sparsity-enforcing method for learning face features. *IEEE Transactions on Image Processing*, 18:188–201.
- Destrero et al., 2009. Feature selection for high dimensional data. *Computational Management Science*, 6:25–40.
- Rosasco et al., 2009. Iterative Projection methods for structured sparsity regularization. *MIT-CSAIL-TR-2009-050-CBCL-282*.
- Mosci et al., 2010. A Regularization Approach to Nonlinear Variable Selection. *Proc. of AISTATS 2010*.