

RegML 2016
Class 2
Tikhonov regularization and kernels

Lorenzo Rosasco
UNIGE-MIT-IIT

June 27, 2016

Learning problem

Problem For $\mathcal{H} \subset \{f \mid f : X \rightarrow Y\}$, solve

$$\min_{f \in \mathcal{H}} \mathcal{E}(f), \quad \int d\rho(x, y) L(f(x), y)$$

given $S_n = (x_i, y_i)_{i=1}^n$ (ρ , fixed, unknown).

Empirical Risk Minimization (ERM)

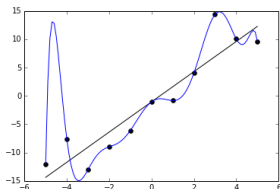
$$\min_{f \in \mathcal{H}} \mathcal{E}(f) \mapsto \min_{f \in \mathcal{H}} \widehat{\mathcal{E}}(f)$$

$$\widehat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^n L(f(x_i), y_i)$$

proxy to \mathcal{E}

From ERM to regularization

ERM can be a bad idea if n is “small” and \mathcal{H} is “big”



Regularization

$$\min_{f \in \mathcal{H}} \hat{\mathcal{E}}(f) \quad \mapsto \quad \min_{f \in \mathcal{H}} \hat{\mathcal{E}}(f) + \lambda \underbrace{R(f)}_{\text{regularization}}$$

λ regularization parameter

Examples of regularizers

Let

$$f(x) = \sum_{j=1}^p \phi_j(x)w_j$$

▶ ℓ_2

$$R(f) = \|w\|^2 = \sum_{j=1}^p |w_j|^2,$$

▶ ℓ_1

$$R(f) = \|w\|_1 = \sum_{j=1}^p |w_j|,$$

▶ Differential operators

$$R(f) = \int_X \|\nabla f(x)\|^2 d\rho(x),$$

▶ ...

From statistics to optimization

Problem Solve

$$\min_{w \in \mathbb{R}^p} \hat{\mathcal{E}}(w) + \lambda \|w\|^2$$

with

$$\hat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^n L(w^\top x_i, y_i).$$

Minimization

$$\min_w \hat{\mathcal{E}}(w) + \lambda \|w\|^2$$

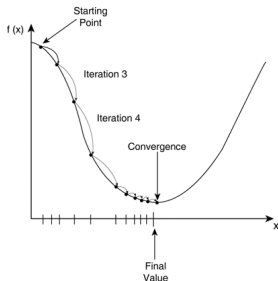
- ▶ Strongly convex functional
- ▶ Computations depends on the considered function

Logistic regression

$$\hat{\mathcal{E}}_\lambda(w) = \underbrace{\frac{1}{n} \sum_{i=1}^n \log(1 + e^{-y_i w^\top x_i})}_{\text{smooth and strongly convex}} + \lambda \|w\|^2.$$

$$\nabla \hat{\mathcal{E}}_\lambda(w) = -\frac{1}{n} \sum_{i=1}^n \frac{x_i y_i}{1 + e^{y_i w^\top x_i}} + 2\lambda w$$

Gradient descent



Let $F : \mathbb{R}^n \rightarrow \mathbb{R}$ twice differentiable, (strictly convex) and such that

$$\|\nabla F(w) - \nabla F(w')\| \leq L\|w - w'\|$$

(e.g. $\sup_w \underbrace{\|H(w)\|}_{\text{hessian}} \leq L$)

Then

$$w_0 = 0, \quad w_{t+1} = w_t - \frac{1}{L}\nabla F(w_t),$$

converges to the minimizer of F .

Gradient descent for LR

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \log(1 + e^{-y_i w^\top x_i}) + \lambda \|w\|^2$$

Consider

$$w_{t+1} = w_t - \frac{1}{L} \left[-\frac{1}{n} \sum_{i=1}^n \frac{x_i y_i}{1 + e^{y_i w_t^\top x_i}} + 2\lambda w_t \right]$$

Hinge loss and support vector machines

$$\hat{\mathcal{E}}_\lambda(w) = \underbrace{\frac{1}{n} \sum_{i=1}^n |1 - y_i w^\top x_i|_+ + \lambda \|w\|^2}_{\text{non-smooth \& strongly-convex}}$$

Consider “left” derivative

$$w_{t+1} = w_t - \frac{1}{L\sqrt{t}} \left(\frac{1}{n} \sum_{i=1}^n S_i(w_t) + 2\lambda w_t \right)$$

$$S_i(w) = \begin{cases} -y_i x_i & \text{if } y_i w^\top x_i \leq 1 \\ 0 & \text{otherwise} \end{cases}, \quad L = \sup_{x \in X} \|x\| + 2\lambda.$$

Subgradient

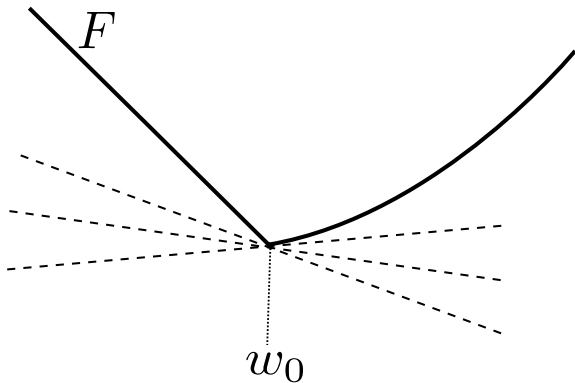
Let $F : \mathbb{R}^p \rightarrow \mathbb{R}$ convex,

Subgradient

$\partial F(w_0)$ set of vectors $v \in \mathbb{R}^p$ such that, for every $w \in \mathbb{R}^p$

$$F(w) - F(w_0) \geq (w - w_0)^\top v$$

In one dimension $\partial F(w_0) = [F'_-(w_0), F'_+(w_0)]$.



Subgradient method

Let $F : \mathbb{R}^p \rightarrow \mathbb{R}$ strictly convex, with bounded subdifferential, and $\gamma_t = 1/t$ then,

$$w_{t+1} = w_t - \gamma_t v_t$$

with $v_t \in \partial F(w_t)$ converges to the minimizer of F .

Subgradient method for SVM

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n |1 - y_i w^\top x_i|_+ + \lambda \|w\|^2$$

Consider

$$w_{t+1} = w_t - \frac{1}{t} \left(\frac{1}{n} \sum_{i=1}^n S_i(w_t) + 2\lambda w_t \right)$$

$$S_i(w_t) = \begin{cases} -y_i x_i & \text{if } y_i w^\top x_i \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Complexity

Logistic: $O(ndT)$

SVM: $O(ndT)$

n number of examples, d dimensionality, T number of steps

What if $n \ll d$? Can we get better complexities?

Representer theorems

Idea Show that

$$f(x) = w^\top x = \sum_{i=1}^n x_i^\top x c_i, \quad c_i \in \mathbb{R}.$$

i.e. $w = \sum_{i=1}^n x_i c_i$.

Compute $c = (c_1, \dots, c_n) \in \mathbb{R}^n$ rather than $w \in \mathbb{R}^d$.

Representer theorem for GD & LR

By induction

$$c_{t+1} = c_t - \frac{1}{L} \left[-\frac{1}{n} \sum_{i=1}^n \frac{e_i y_i}{1 + e^{y_i f_t(x_i)}} + 2\lambda c_t \right]$$

with e_i the i -th element of the canonical basis and

$$f_t(x) = \sum_{i=1}^n x^\top x_i (c_t)_i$$

Representer theorem of SVM

By induction

$$c_{t+1} = c_t - \frac{1}{t} \left(\frac{1}{n} \sum_{i=1}^n S_i(c_t) + 2\lambda c_t \right)$$

with e_i the i -th element of the canonical basis,

$$f_t(x) = \sum_{i=1}^n x^\top x_i (c_t)_i$$

and

$$S_i(c_t) = \begin{cases} -y_i e_i & \text{if } y_i f_t(x_i) < 1 \\ 0 & \text{otherwise} \end{cases} .$$

Complexity

Without representer

Logistic: $O(ndT)$

SVM: $O(ndT)$

With representer

Logistic: $O(n^2(d + T))$

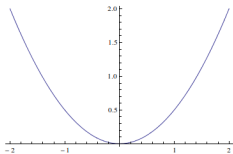
SVM: $O(n^2(d + T))$

n number of example, d dimensionality, T number of steps

Optimality condition for SVM

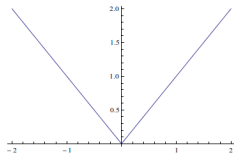
Smooth Convex

$$\nabla F(w_*) = 0$$



Non-smooth Convex

$$0 \in \partial F(w)$$



$$0 \in \partial F(w_*) \Leftrightarrow 0 \in \partial |1 - y_i w^\top x_i|_+ + \lambda 2w$$

$$\Leftrightarrow w \in \partial \frac{1}{2\lambda} |1 - y_i w^\top x_i|_+.$$

Optimality condition for SVM (cont.)

The optimality condition can be rewritten as

$$0 = \frac{1}{n} \sum_{i=1}^n (-y_i x_i c_i) + 2\lambda w \quad \Rightarrow \quad w = \sum_{i=1}^n x_i \left(\frac{y_i c_i}{2\lambda n} \right).$$

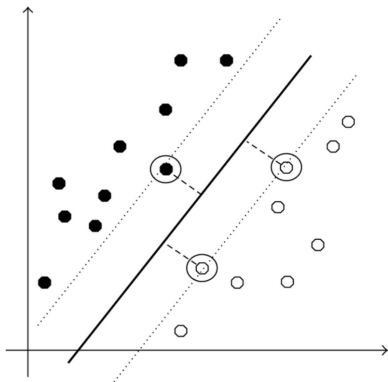
where $c_i = c_i(w) \in [V^-(-y_i w^\top x_i), V^+(-y_i w^\top x_i)]$.

A direct computation gives

$$\begin{array}{lll} c_i = 1 & \text{if} & yf(x_i) < 1 \\ 0 \leq c_i \leq 1 & \text{if} & yf(x_i) = 1 \\ c_i = 0 & \text{if} & yf(x_i) > 1 \end{array}$$

Support vectors

$$\begin{aligned} c_i = 1 & \quad \text{if } yf(x_i) < 1 \\ 0 \leq c_i \leq 1 & \quad \text{if } yf(x_i) = 1 \\ c_i = 0 & \quad \text{if } yf(x_i) > 1 \end{aligned}$$



Are loss functions all the same?

$$\min_w \hat{\mathcal{E}}(w) + \lambda \|w\|^2$$

- ▶ each loss has a different target function. . .
- ▶ . . . and different computations

The choice of the loss is problem dependent

So far

- ▶ regularization by penalization
- ▶ iterative optimization
- ▶ linear models

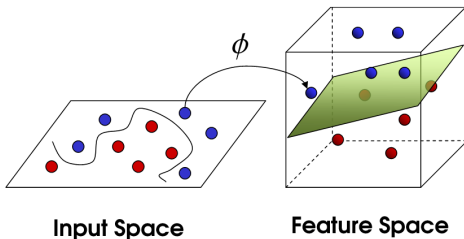
What about non-linear and nonparametric models?

Non-linear features

$$f(x) = \sum_{i=1}^d w_i x_i \quad \mapsto \quad f(x) = \sum_{i=1}^p w_i \phi_i(x_i).$$

$$\Phi(x) = (\phi_1(x), \dots, \phi_p(x)),$$

Model



Non-linear features

$$f(x) = \sum_{i=1}^d w_i x_i \quad \mapsto \quad f(x) = \sum_{i=1}^p w_i \phi_i(x_i).$$

$$\Phi(x) = (\phi_1(x), \dots, \phi_p(x)),$$

Computations

Same up-to the change

$$x \mapsto \Phi(x)$$

Representer theorem with non-linear features

$$f(x) = \sum_{i=1}^n x_i^\top x c_i \quad \mapsto \quad f(x) = \sum_{i=1}^n \Phi(x_i)^\top \Phi(x) c_i$$

Rewriting logistic regression and gradient descent

By induction

$$c_{t+1} = c_t - \frac{1}{L} \left[-\frac{1}{n} \sum_{i=1}^n \frac{e_i y_i}{1 + e^{y_i f_t(x_i)}} + 2\lambda c_t \right]$$

with e_i the i -th element of the canonical basis and

$$f_t(x) = \sum_{i=1}^n \Phi(x)^\top \Phi(x)_i (c_t)_i$$

Rewriting SVM by subgradient

By induction

$$c_{t+1} = c_t - \frac{1}{t} \left(\frac{1}{n} \sum_{i=1}^n S_i(c_t) + 2\lambda c_t \right)$$

with e_i the i -th element of the canonical basis,

$$f_t(x) = \sum_{i=1}^n \Phi(x)^\top \Phi(x)_i (c_t)_i$$

and

$$S_i(c_t) = \begin{cases} -y_i e_i & \text{if } y_i f_t(x_i) < 1 \\ 0 & \text{otherwise} \end{cases}.$$

From features to kernels

$$f(x) = \sum_{i=1}^n x_i^\top x c_i \quad \mapsto \quad f(x) = \sum_{i=1}^n \Phi(x_i)^\top \Phi(x) c_i$$

Kernels

$$\Phi(x)^\top \Phi(x') \mapsto K(x, x')$$

$$f(x) = \sum_{i=1}^n K(x_i, x) c_i$$

LR and SVM with kernels

As before:

LR

$$c_{t+1} = c_t - \frac{1}{L} \left[-\frac{1}{n} \sum_{i=1}^n \frac{e_i y_i}{1 + e^{y_i f_t(x_i)}} + 2\lambda c_t \right]$$

SVM

$$c_{t+1} = c_t - \frac{1}{t} \left(\frac{1}{n} \sum_{i=1}^n S_i(c_t) + 2\lambda c_t \right)$$

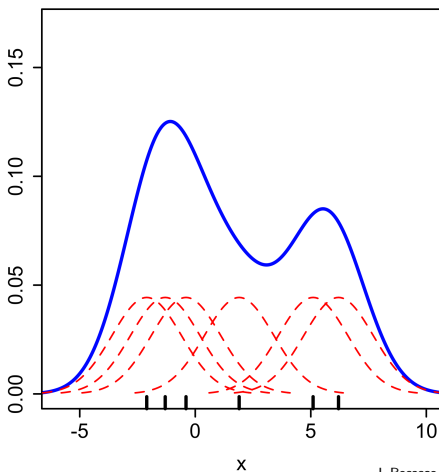
But now

$$f_t(x) = \sum_{i=1}^n K(x, x_i) (c_t)_i$$

Examples of kernels

- ▶ Linear $K(x, x') = x^\top x'$
- ▶ Polynomial $K(x, x') = (1 + x^\top x)^p$, with $p \in \mathbb{N}$
- ▶ Gaussian $K(x, x') = e^{-\gamma \|x - x'\|^2}$, with $\gamma > 0$

$$f(x) = \sum_{i=1}^n c_i K(x_i, x)$$



Kernel engineering

Kernels for

- ▶ Strings,
- ▶ Graphs,
- ▶ Histograms,
- ▶ Sets,
- ▶ ...

What is a kernel?

$$K(x, x')$$

- ▶ Similarity measure
- ▶ Inner product
- ▶ Positive definite function

Positive definite function

$K : X \times X \rightarrow \mathbb{R}$ is *positive definite*, when

for any $n \in \mathbb{N}$, $x_1, \dots, x_n \in X$, let K_n such that

$$K_n \in \mathbb{R}^{n \times n}, \quad (K_n)_{ij} = K(x_i, x_j),$$

then K_n is positive semidefinite, (eigenvalues ≥ 0)

PD functions and RKHS

Each PD Kernel defines a function space called Reproducing kernel Hilbert space (RKHS)

$$\mathcal{H} = \overline{\text{span} \{K(\cdot, x) \mid x \in X\}}.$$

Nonparametrics and kernels

Number of parameters automatically determined by number of points

$$f(x) = \sum_{i=1}^n K(x_i, x)c_i$$

Compare to

$$f(x) = \sum_{j=1}^p \phi_j(x)w_j$$

This class

- ▶ Learning and Regularization: logistic regression and SVM
- ▶ Optimization with first order methods
- ▶ Linear and Non-linear parametric models
- ▶ Non-parametric models and kernels

Next class

Beyond penalization

Regularization by

- ▶ projection
- ▶ early Stopping