RegML2017@SIMULA Oslo
Class 2
Tikhonov regularization and kernels

Lorenzo Rosasco
UNIGE-MIT-IIT

May 3, 2017

Learning problem

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

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

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

How can we design reliable algorithms?

This class

Regularization by penalization

Logistic regression

Representer theorems and Kernels

Support vector machines

Empirical Risk Minimization (ERM)

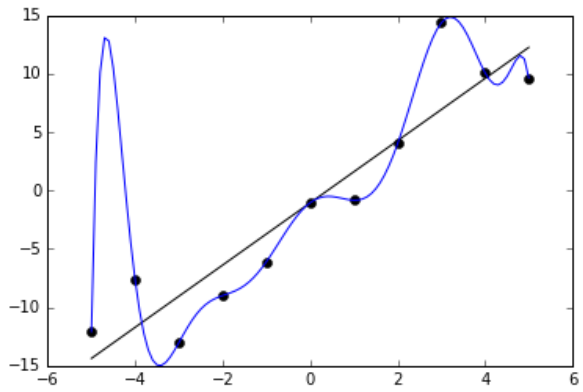
$$\min_{f \in \mathcal{H}} \mathcal{E}(f) \mapsto \min_{f \in \mathcal{H}} \widehat{\mathcal{E}}(f)$$

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

proxy to $\mathcal{E}$

From ERM to regularization

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



Penalization

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

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

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

with

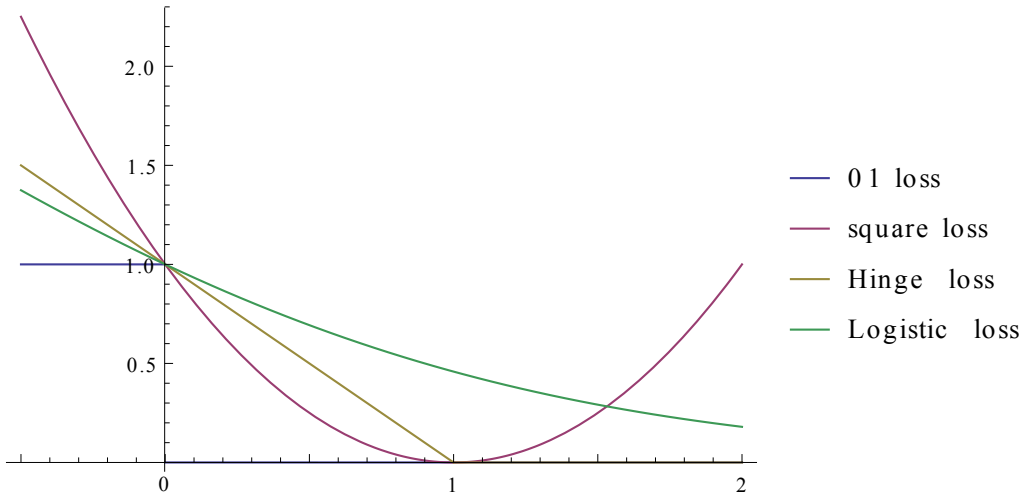
$$\widehat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^{n} L(w^\top x_i, y_i).$$

Minimization

$$\min_{w} \widehat{\mathcal{E}}(w) \; + \; \lambda \|w\|^2$$

- Strongly convex continuous coercive functional
- Computations depends on the considered loss function

Hinge loss & SVM



$$\log\left(1 + e^{-yf(x)}\right)$$

- Linear
- Non linear

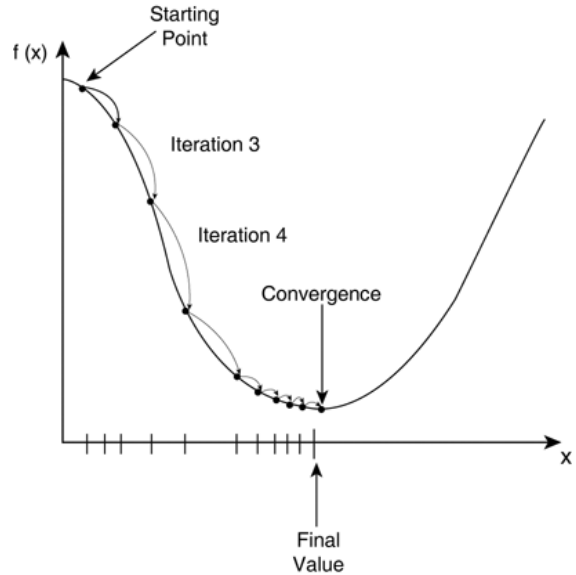
Solution by gradient descent

Logistic regression

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

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

Gradient descent



Let $F : \mathbb{R}^d \to \mathbb{R}$ 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

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

Excercise: compute a good step-size!

Remarks

Complexity: $O(ndT)$

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

What about non-linear functions?

Non-linear features

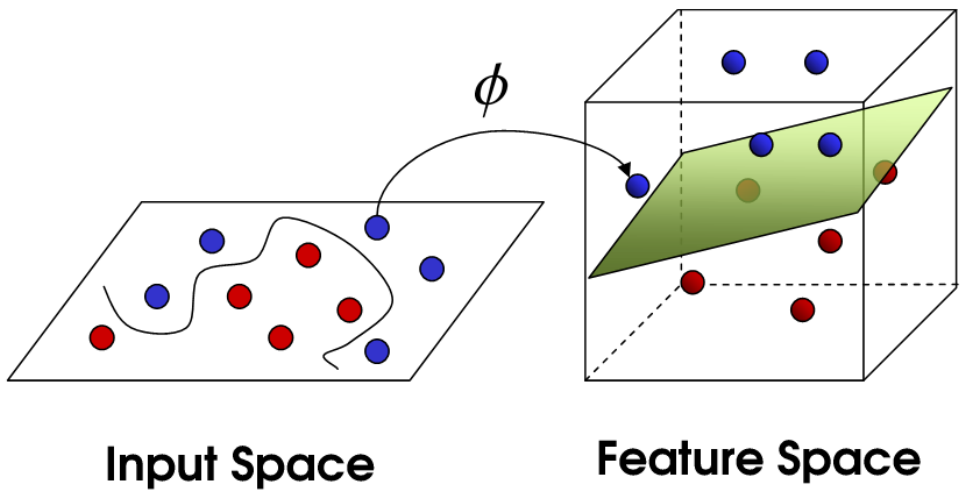
$$f(x) = \sum_{i=1}^{d} w_i x_i \quad \mapsto \quad f(x) = \sum_{i=1}^{p} w_i \phi_i(x_i).$$

Non-linear features

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

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

Model



Gradient descent for LR

Same up-to the change $x \mapsto \Phi(x)$...

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

Consider

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

Gradient descent for LR

Same up-to the change $x \mapsto \Phi(x)$...

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

Consider

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

Complexity is $O(npT)$

Representer theorem and Kernels

Nonparametrics?

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

Representer theorem and Kernels

Nonparametrics?

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

An equivalent formulation for gradient descent provides a way in....

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

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

Show as an excercise!

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

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

Then:

- Replace inner product $x^\top x' \mapsto K(x, x') = \Phi(x)^T \Phi(x)$
- Compute $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ rather than $w \in \mathbb{R}^d$.

From features to kernels

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

Kernels

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

From features to kernels

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

Kernels

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

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

LR with kernels

GD-LR becomes

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

where

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

What are (good) kernels?

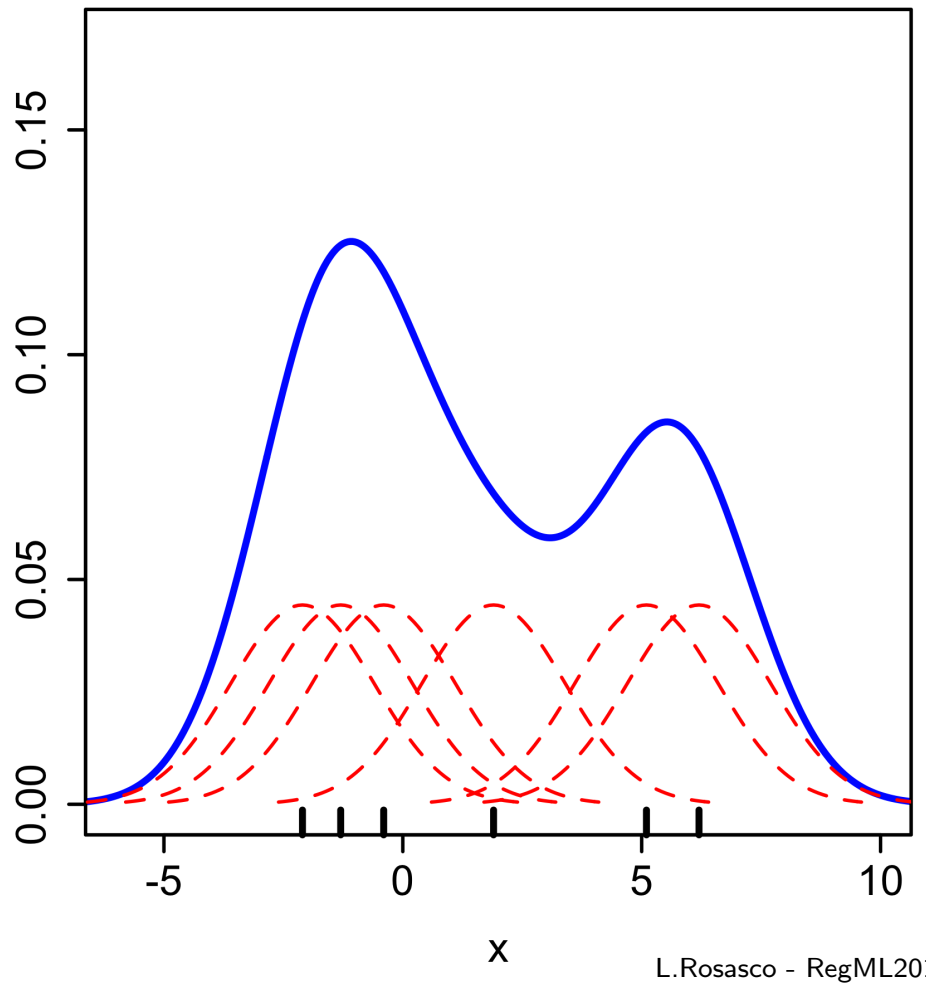
Examples of kernels

- 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 $\gamma > 0$

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?

$$K(x, x')$$

- A similarity measure
- An inner product
- A positive definite function

Positive definite function

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

for any $n \in \mathbb{N}$, $x_1, \ldots, x_n \in X$, 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, RKHS and Gaussian processes

Each PD Kernel defines:

- a function space called Reproducing kernel Hilbert space (RKHS)...

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

- a Gaussian process with covariance function K.

Nonparametrics and kernels

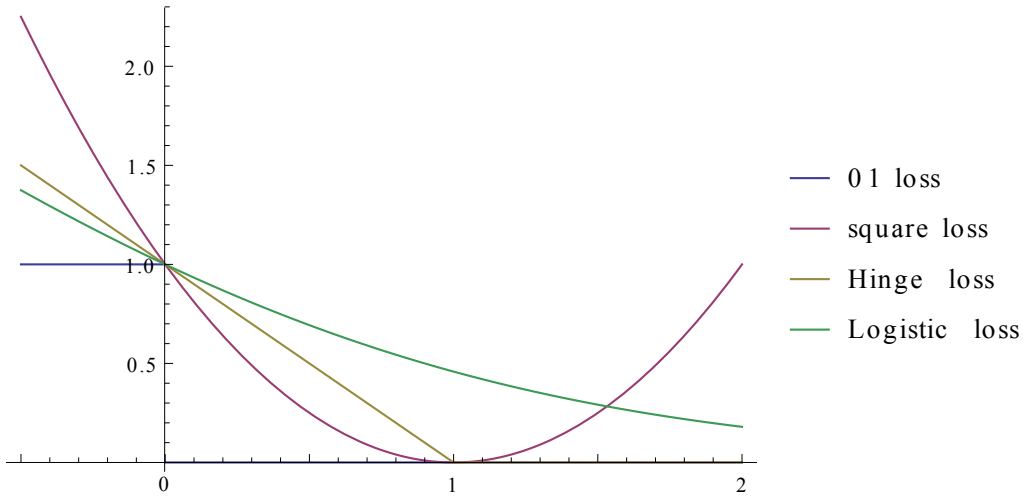
Number of parameters automatically determined by number of points

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

Compare to

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

Hinge loss



$$|1 - yf(x)|_+$$

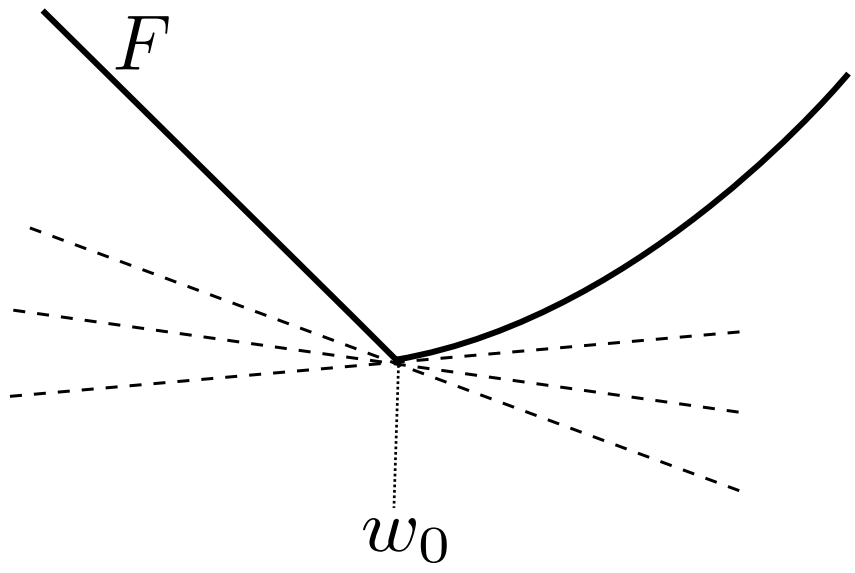
- Linear
- Non linear

Solution by *sub-gradient* method

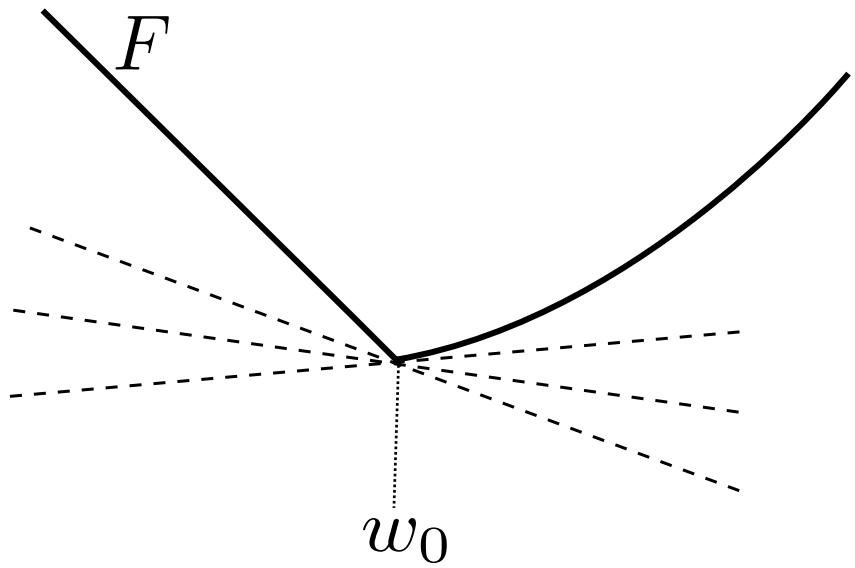
Support vector machine (SVM)

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

Subgradient



Subgradient



Let $F : \mathbb{R}^p \to \mathbb{R}$ 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 $F : \mathbb{R}^p \to \mathbb{R}$ strictly convex, with bounded subdifferential, and $\gamma_t = 1/\sqrt{t}$ 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 "left" derivative

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

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

Remarks

- Traditional solution is solving a quadratic program (QP) ...

Remarks

- Traditional solution is solving a quadratic program (QP) ...
- . . . but it doesn't scale (and is more complicated!).

Remarks

- Traditional solution is solving a quadratic program (QP) ...
- . . . but it doesn't scale (and is more complicated!).
- Easy extension to stochastic gradient.

Remarks

- Traditional solution is solving a quadratic program (QP) ...
- . . . but it doesn't scale (and is more complicated!).
- Easy extension to stochastic gradient.
- Can be seen as a robust, regularized perceptron.

Non linear Subgradient SVM

Same up-to the change $x \mapsto \Phi(x)$,

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

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

Representer theorem for SVM

By induction (exercise)

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

with

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

and

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

where e_i the i-th element of the canonical basis.

Kernel SVM by subgradient

Given a kernel K,

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

with

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

and

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

where e_i the i-th element of the canonical basis.

Complexity

Without representer

Logistic: $O(ndT)$ SVM: $O(ndT)$

With representer

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

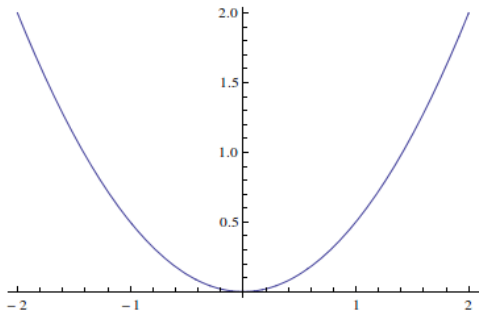
n number of example, d dimensionality, T number of steps

But why are these called *support vector* machines???

Optimality condition for SVM

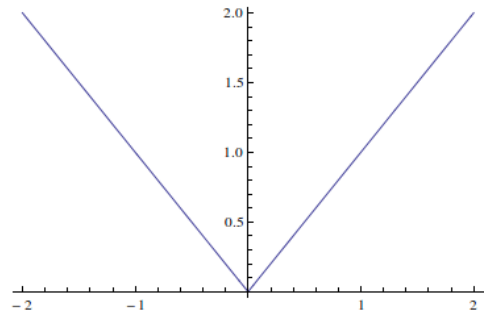
Smooth Convex

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



Non-smooth Convex

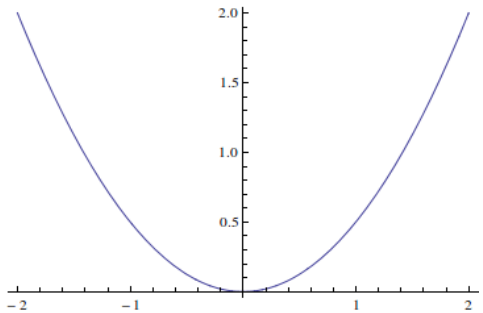
$$0 \in \partial F(w)$$



Optimality condition for SVM

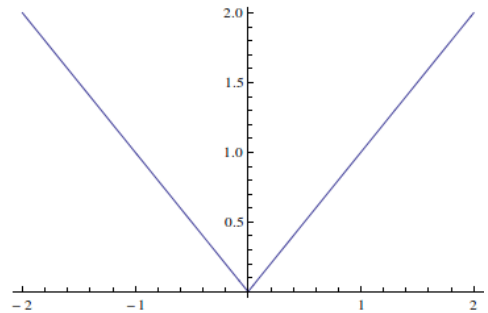
Smooth Convex

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



Non-smooth Convex

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



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

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

Optimality condition for SVM (cont.)

The optimality condition can be rewritten as

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

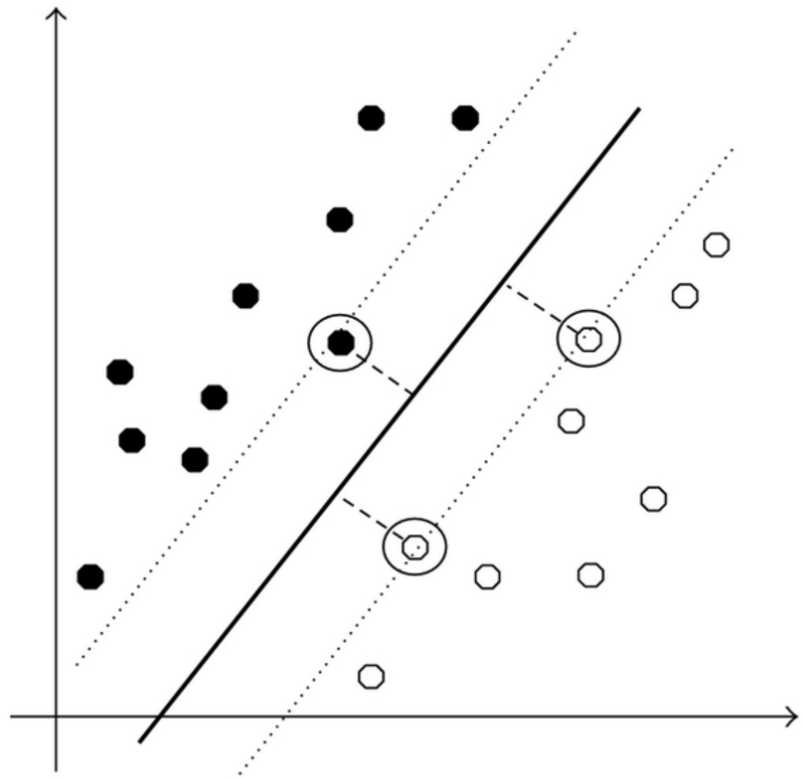
where $c_i = c_i(w) \in [V^-(-y_i w^\top x_i), V^+(-y_i w^\top x_i)]$.

A direct computation gives

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

Support vectors

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



Are loss functions all the same?

$$\min_{w} \widehat{\mathcal{E}}(w) \; + \; \lambda \|w\|^2$$

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

The choice of the loss is problem dependent

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