TUM 2016
Class 2
Tikhonov regularization and kernels

Lorenzo Rosasco
UNIGE-MIT-IIT

July 25, 2016

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).

Empirical Risk Minimization (ERM)

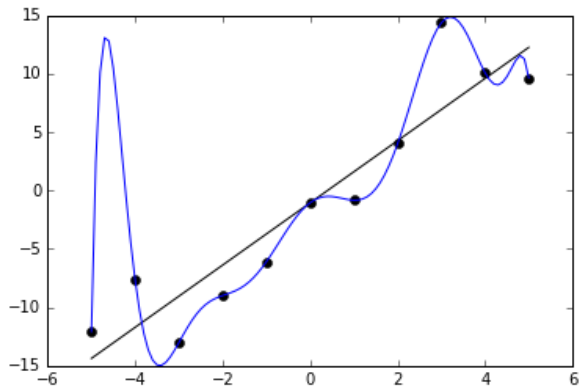
$$\min_{f \in \mathcal{H}} \mathcal{E}(f) \mapsto \min_{f \in \mathcal{H}} \widehat{\mathcal{E}}(f)$$

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

proxy to $\mathcal{E}$

From ERM to regularization

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



Regularization

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

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

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

with

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

Minimization

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

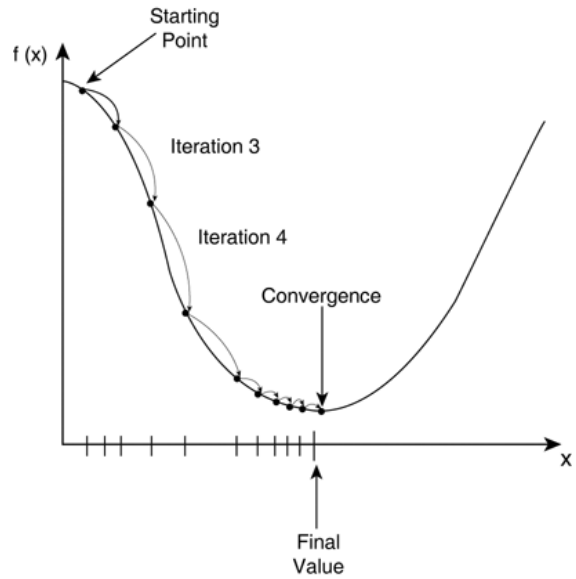
- Strongly convex functional
- Computations depends on the considered loss function

Logistic regression

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

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

Gradient descent



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

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

Complexity

Logistic: $O(ndT)$

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

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

A representer theorem

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

Representer theorem for GD & LR

By induction[1]

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

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

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

---

[1] Proof:

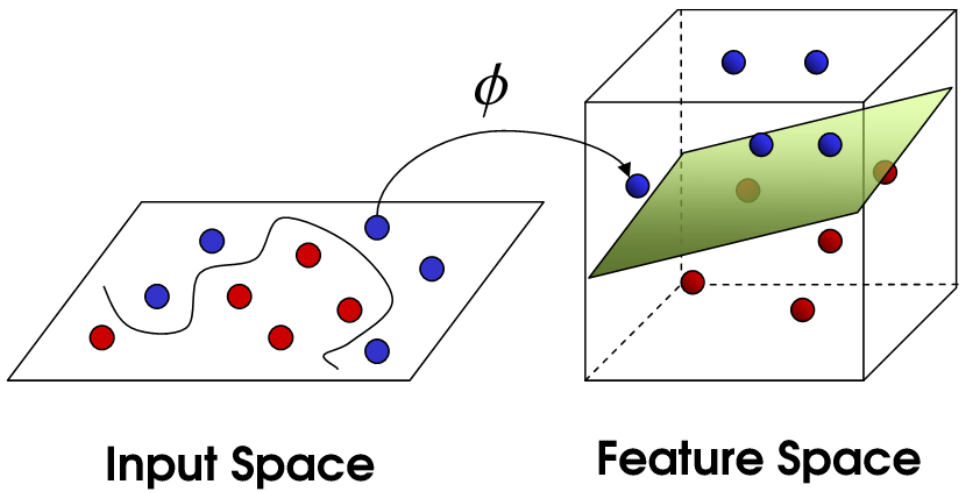
$$w_{t+1} = \underbrace{w_t}_{\sum_{i=1}^{n} x_i (c_t)_i} - \frac{1}{L}\left[-\frac{1}{n}\sum_{i=1}^{n} x_i \frac{y_i}{1+\underbrace{e^{y_i w_t^\top x_i}}_{e^{y_i f_t(x_i)}}} + 2\lambda \underbrace{w_t}_{\sum_{i=1}^{n} x_i (c_t)_i}\right]$$

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



Non-linear features

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

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

Computations

Same up-to the change

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

Representer theorem with non-linear features

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

Rewriting logistic regression and gradient descent

As before,

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

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

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

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

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

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

Subgradient

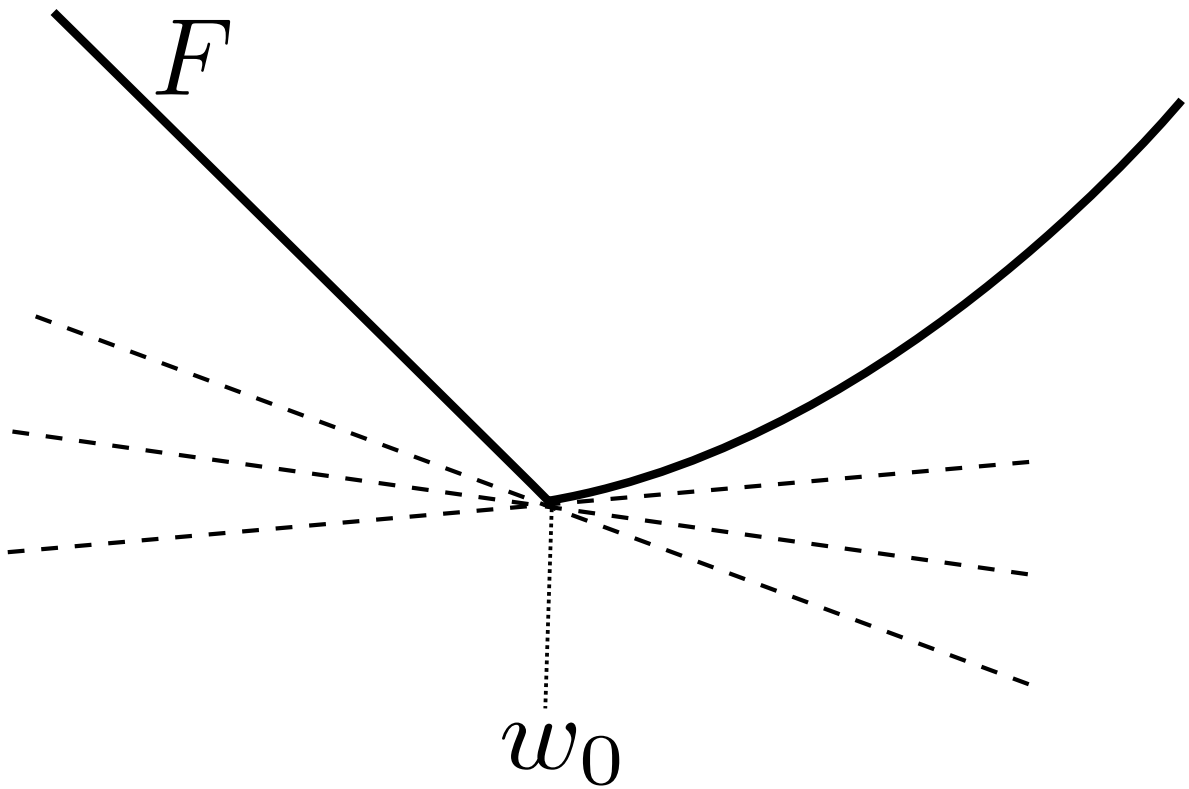
Let $F : \mathbb{R}^p \to \mathbb{R}$ convex,

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 $\gamma_t = 1/t$ 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

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

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

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

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

Nonlinear 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,

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

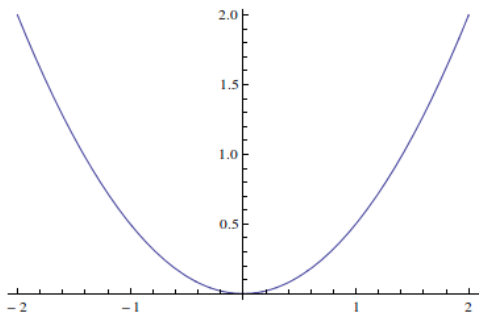
and

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

Optimality condition for SVM

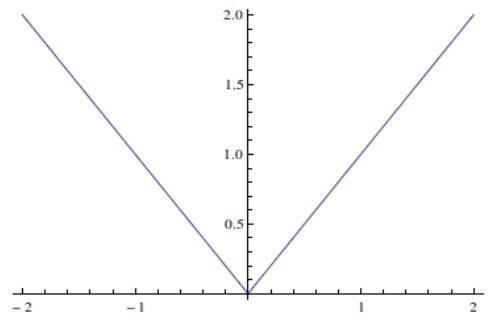
Smooth Convex

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



Non-smooth Convex

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



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

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

$$\dots$$

Optimality condition for SVM (cont.)

The optimality condition can be rewritten as

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

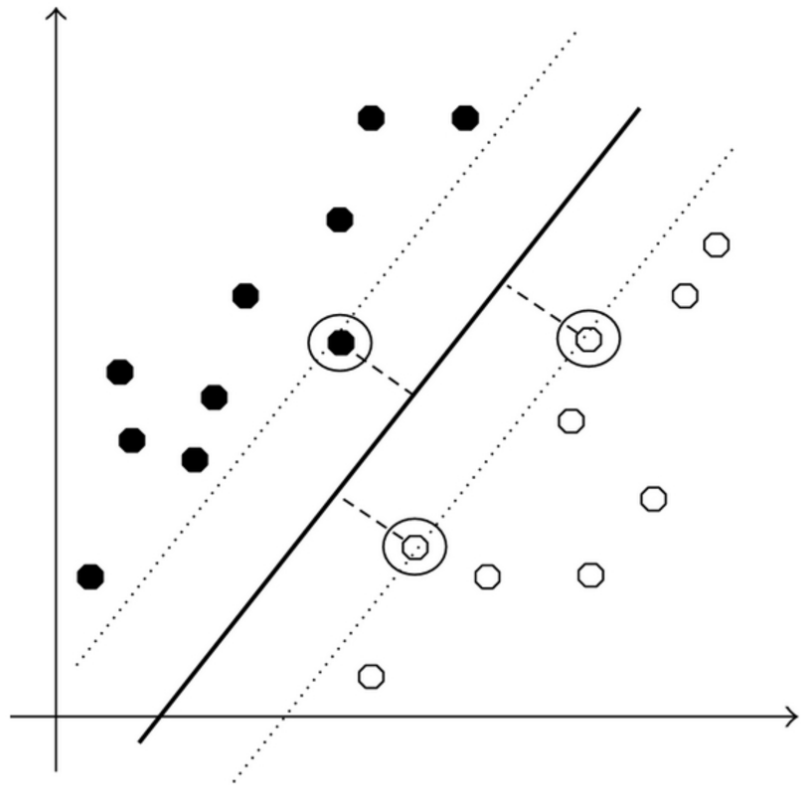
where $c_i = c_i(w) \in [V^-(-y_i w^\top x_i), V^+(-y_i w^\top x_i)]$.

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

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



Square loss

$$(1 - yw^\top x)^2 = (y - w^\top x)^2$$

$$\widehat{\mathcal{E}}_\lambda(w) = \widehat{\mathcal{E}}(w) + \lambda\|w\|^2 \quad \text{with} \quad \widehat{\mathcal{E}}(w) = \frac{1}{n}\|\hat{X}w - \hat{y}\|^2$$

- $\hat{X}$ $n \times d$ data matrix
- $\hat{y}$ $n \times 1$ output vector.

Ridge regression / Tikhonov regression

$$\widehat{\mathcal{E}}_\lambda(w) = \underbrace{\frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda\|w\|^2}_{\text{Smooth and strongly convex}}$$

$$\nabla\widehat{\mathcal{E}}_\lambda(w) = \frac{2}{n}\hat{X}^\top(\hat{X}w - \hat{y}) + 2\lambda w = 0$$

$$\implies (\hat{X}^\top\hat{X} + \lambda n I)w = \hat{X}^\top\hat{y}$$

Linear systems

$$(\hat{X}^\top \hat{X} + \lambda n I)w = \hat{X}^\top \hat{y}$$

- nd^2 to form $\hat{X}^\top \hat{X}$
- roughly d^3 to solve the linear system

Representer theorem for square loss

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

Using SVD of $\hat{X}$...

$$w = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y} = \hat{X}^\top \underbrace{(\hat{X}\hat{X}^\top + \lambda n I)^{-1} \hat{y}}_{c}$$

$$\Longrightarrow \; w = \hat{X}^\top c = \sum_{i=1}^{n} x_i c_i$$

Nonlinear ridge regression

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

$$w = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y}, \quad c = (\hat{X}\hat{X}^\top + \lambda n I)^{-1} \hat{y}$$

- non-linear function

$$x \mapsto \Phi(x) = (\phi_1(x), \dots \phi_n(x)), \quad f(x) = \phi(x)^\top w$$

Interlude: linear systems and stability

$$Aw = y, \qquad A = \mathrm{diag}(a_1, \ldots, a_d), \quad c = \frac{a_1}{a_d} < \infty,$$
$$w = A^{-1}y, \qquad A^{-1} = \mathrm{diag}(a_1^{-1}, \ldots, a_d^{-1})$$

More generally

$$A = U\Sigma U^\top, \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$$
$$A^{-1} = U\Sigma^{-1}U^\top, \quad \Sigma^{-1} = \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_d^{-1})$$

Tikhonov Regularization

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 \quad \mapsto \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \lambda\|w\|^2$$

$$\hat{X}^\top \hat{X} w = \hat{X}^\top \hat{y} \quad \mapsto \quad (\hat{X}^\top \hat{X} + \lambda n I) w = \hat{X}^\top \hat{y}$$

Overfitting and numerical stability

Complexity

Without representer

Logistic: $O(ndT)$ SVM: $O(ndT)$ RR: $O(nd^2 + d^3)$

With representer

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

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

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

So far

- regularization by penalization
- iterative optimization
- linear/non-linear parametric models

What about nonparametric models?

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 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

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

RR

$$(\widehat{K} + \lambda n I)c = \hat{y}$$

But now

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

Nonparametrics and kernels

Number of parameters automatically determined by number of points

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

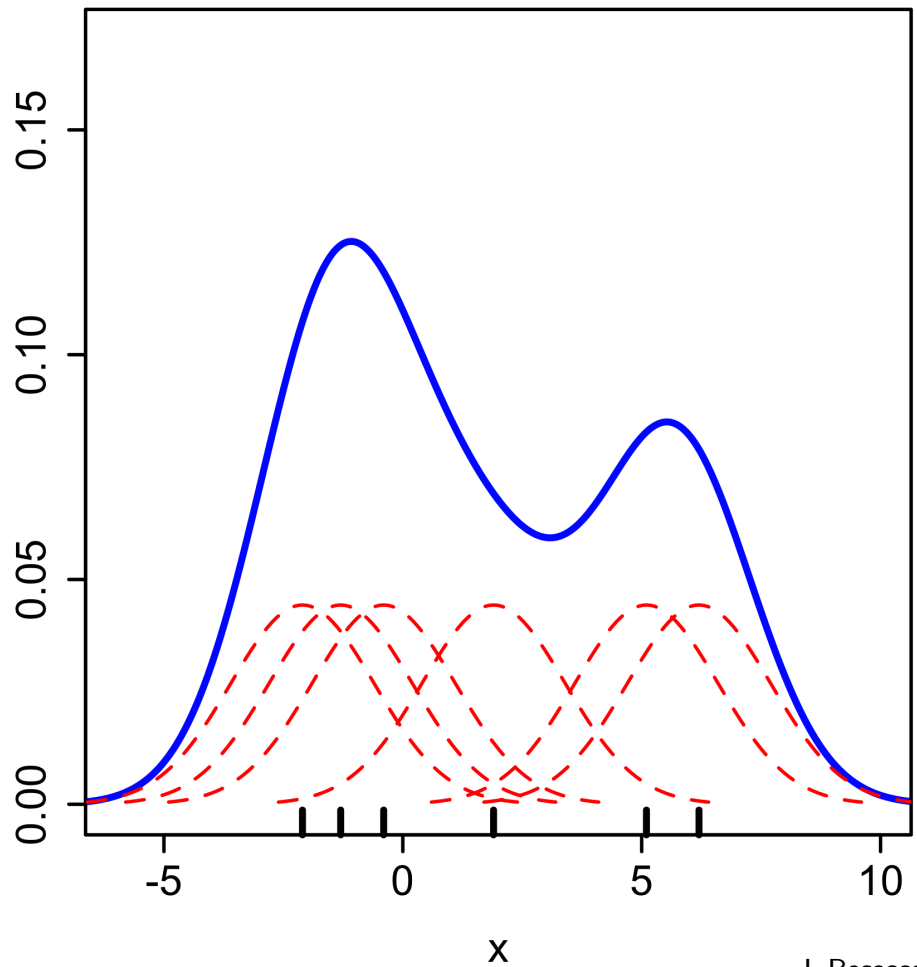
Compare to

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

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

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

for any $n \in \mathbb{N}$, $x_1, \ldots, x_n \in X$, let $\widehat{K}$ be such that

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

then $\widehat{K}$ is positive semidefinite, (eigenvalues ≥ 0)

PD functions and RKHS

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

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

This class

- Learning and regularization: logistic regression, SVM and ridge regression
- Optimization with first order methods
- Linear and Non-linear parametric models
- Non-parametric models and kernels

Next class

Towards large scale:

regularization by

- early stopping
- projection
- randomized subsampling


Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio.
Regularization networks and support vector machines.
Advances in computational mathematics, 13(1):1–50, 2000.


Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Michele Piana, and Alessandro Verri.
Some properties of regularized kernel methods.
Journal of Machine Learning Research, 5(Oct):1363–1390, 2004.