MIT 9.520/6.860, Fall 2018
Statistical Learning Theory and Applications

Class 05: Logistic Regression and Support Vector Machines

Lorenzo Rosasco

Last class

Non linear functions using

- features

$$f(x) = w^\top x \mapsto f(x) = w^\top x,$$

- kernels

$$f(x) = w^\top x \mapsto f(x) = \sum_{i=1}^{N} k(x, x_i) c_i.$$

More precisely

- A feature map Φ defines the space $\mathcal{H}_\Phi$ of functions

$$f(x) = w^\top x,$$

  and $k(x, \bar{x}) := \Phi(x)\Phi(\bar{x})$, is pos. def.
- A pos. def. kernels k defines space $\mathcal{H}_k$ of functions

$$f(x) = \sum_{i=1}^{N} k(x, x_i) c_i.$$

  with the reproducing property

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$$

- For every k there is a[1] Φ such that

$$k(x, \bar{x}) = \Phi(x)\Phi(\bar{x}),$$

  and

$$\mathcal{H}_k \simeq \mathcal{H}_\Phi.$$

[1] Indeed, infinitely many.

Today

Beyond least squares

$$(y - f(x))^2 \mapsto \ell(y, f(x)).$$

Today

- Logistic loss.
- Hinge loss.

ERM and penalization

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \|w\|^2, \qquad \lambda \geq 0.$$

- Logistic loss $\mapsto$ logistic regression.
- Hinge $\mapsto$ SVM.

Non linear extensions via features/kernels.

From regularization to optimization

Problem Solve

$$\min_{w \in \mathbb{R}^d} \widehat{L}(w) + \lambda \|w\|^2$$

where

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

Minimization

Assume ℓ convex and continuous, let

$$\widehat{L}_\lambda(w) = \widehat{L}(w) \; + \; \lambda \|w\|^2.$$

- Coercive[2], strongly convex functional
  $\Rightarrow$ a minimizer exists and is unique.

- Computations depends on the considered loss.

---

[2] $\lim_{\|w\| \to \infty} \widehat{L}_\lambda(w) = \infty.$

Logistic regression

$$\widehat{L}_\lambda(w) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-y_i w^\top x_i}) + \lambda\|w\|^2.$$

- $\widehat{L}_\lambda$ is smooth

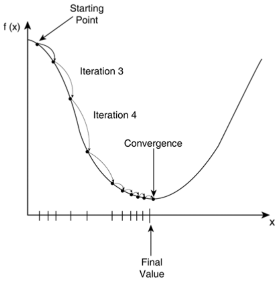
$$\nabla\widehat{L}_\lambda(w) = -\frac{1}{n}\sum_{i=1}^n \frac{x_i y_i}{1 + e^{y_i w^\top x_i}} + 2\lambda w.$$

- Optimality condition gives a nonlinear equation

$$\nabla\widehat{L}_\lambda(w) = 0,$$

so we use gradient methods.

Gradient descent



Let $F : \mathbb{R}^d \to \mathbb{R}$ differentiable, (strictly) convex and such that

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

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

Then

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

converges to the minimizer of F.

Gradient descent for logistic regression

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

Consider

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

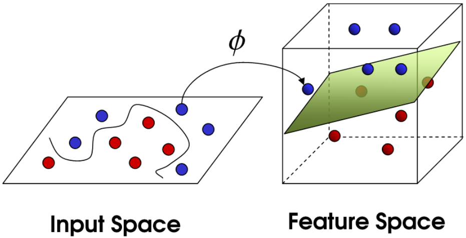
Complexity

Time: $O(ndT)$ for n examples, d dimension, T steps.

Non-linear features

$$f(x) = w^\top x \quad \mapsto \quad f(x) = w^\top \Phi(x),$$

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



Gradient descent for non linear logistic regression

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

Consider

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

Complexity

Time $O(npT)$ for n examples, p features, T steps.

What about kernels?

Representer theorem for logistic regression?

As for least squares,
Show that $w = \sum_{i=1}^n x_i c_i$. i.e.

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

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

Representer theorem for GD & logistic regression

By induction

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

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

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

Proof of the representer theorem for GD & logistic regression

Assume

$$w_t = \sum_{i=1}^n x_i (c_t)_i$$

$$\begin{aligned}
w_{t+1} &= w_t - \frac{1}{B}\left(-\frac{1}{n}\sum_{i=1}^n \frac{x_i y_i}{1+e^{y_i w_t^\top x_i}} + 2\lambda w_t\right) \\
&= \sum_{i=1}^n x_i (c_t)_i - \frac{1}{B}\left(-\frac{1}{n}\sum_{i=1}^n x_i \frac{y_i}{1+e^{y_i (\sum_{j=1}^n x_j (c_t)_j)^\top x_i}} + 2\lambda (\sum_{i=1}^n x_i (c_t)_i)\right) \\
&= \sum_{i=1}^n x_i \left[(c_t)_i - \frac{1}{B}\left(-\frac{1}{n}\frac{y_i}{1+e^{y_i (\sum_{j=1}^n x_j (c_t)_j)^\top x_i}} + 2\lambda (c_t)_i\right)\right].
\end{aligned}$$

Then

$$w_{t+1} = \sum_{i=1}^n x_i (c_{t+1})_i$$

Kernel logistic regression

Given a pos. def. kernel, consider

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

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

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

Complexity
Time: $O(n^2(C_k + T))$ for n examples, C_k kernel evaluation, T steps.

Hinge loss and support vector machines

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

Consider "left" derivative

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

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

$B\sqrt{t}$ is a bound on the subgradient.

Subgradient

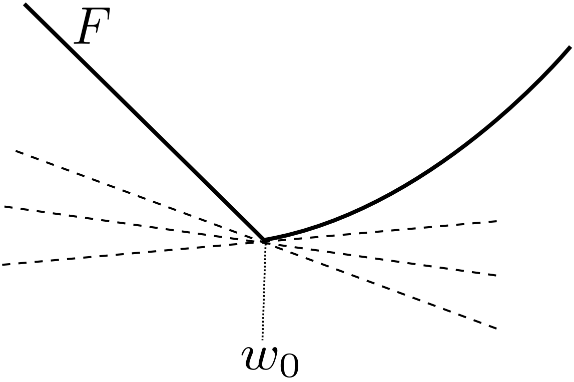
Let $F : \mathbb{R}^p \to \mathbb{R}$ convex,

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 $F : \mathbb{R}^p \to \mathbb{R}$ convex, with subdifferential bounded by B, and $\gamma_t = \frac{1}{B\sqrt{t}}$ then,

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

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

Note: it is not a descent method.

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

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

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

Complexity

Time: $O(ndT)$ for n examples, d dimensions, T steps.

Connection to the perceptron

- Replace the hinge loss with

$$\ell(y, f(x)) = |-yf(x)|_+.$$

- Set $\lambda = 0$.

Reasoning as above we can solve ERM by

$$w_{t+1} = w_t - \frac{1}{B\sqrt{t}}\left(\frac{1}{n}\sum_{i=1}^{n} S_i(w_t) \;+\; 2\lambda w_t\right), \qquad S_i(w_t) = \begin{cases} -y_i x_i & \text{if } y_i w^\top x_i \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

This is a "batch" version of the perceptron of the algorithm,

$$w_{t+1} = w_t - \gamma\left(S_t(w_t)\right), \qquad S_i(w_t) = \begin{cases} -y_t x_t & \text{if } y_t w^\top x_t \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

Nonlinear SVM using features and subgradients

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

Consider

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

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

Complexity

Time $O(npT)$ for n examples, p features, T steps.

What about kernels?

Representer theorem of SVM

By induction

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

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

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

Kernel SVM using subgradient

By induction

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

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

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

and

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

Complexity

Time: $O(n^2(C_k + T))$ for n examples, C_k kernel evaluation, T steps.

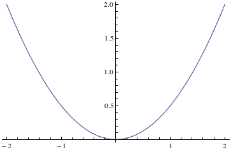
What else

- Why are they called support vector machines?
- And what about the margin and all that?

Optimality condition for SVM

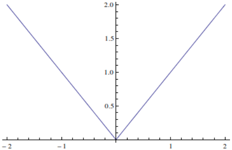
Smooth Convex

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



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

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

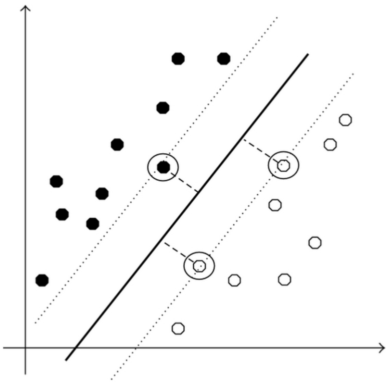
with ℓ^-, ℓ^+ left, right derivatives of $|\cdot|_+$.

A direct computation gives

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

Support vectors

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



Sparsity and SVM solvers

The conditions

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

show that the SVM solution is sparse wrt the training points.

- Classical Quadratic Programming solvers for SVM exploit sparsity.
- Subgradient methods require only matrix vector multiplications, hence are preferable for large scale problems.

And now the margin

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

For $C = \frac{1}{2n\lambda}$, consider the following equivalent formulation

$$\min_{w \in \mathbb{R}^p} C \sum_{i=1}^{n} \xi_i \; + \; \frac{1}{2}\|w\|^2,$$

subj. to for all $i = 1, \dots, n$,

$$\xi_i \geq 0, \qquad y_i w^\top x_i \geq 1 - \xi_i$$

The *slack* variables ξ_i's quantify how much constraints are violated.

Soft and hard margin SVM

This is the classical *soft margin* SVM formulation

$$\min_{w \in \mathbb{R}^p} C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|w\|^2, \qquad \text{subj. to} \quad \xi_i \geq 0, \quad y_i w^\top x_i \geq 1 - \xi_i, \quad \forall\, i = 1, \dots, n.$$

The name comes from considering the limit case $C \to 0$

$$\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|^2, \qquad \text{subj. to} \quad y_i w^\top x_i \geq 1, \quad \forall\, i = 1, \dots, n,$$

called *hard margin* SVM.

Max margin

$$\min_{w \in \mathbb{R}^p} \|w\|^2, \qquad \text{subj. to} \quad y_i w^\top x_i \geq 1, \quad \forall\, i = 1, \dots, n.$$

The above problem has a geometric interpretation.

For linearly separable data

- $2/\|w\|$ is the margin: smallest distance of each class to $w^\top x$.
- The constraint select functions linearly separating the data.

Hard margin SVM: find the max margin solution separating the data.

Summary

- Logistic regression and SVM are instances of penalized ERM.
- Optimization by gradient descent/subgradient method.
- Nonlinear extension using features/kernels.
- Optimality conditions and support vectors.
- Margin .