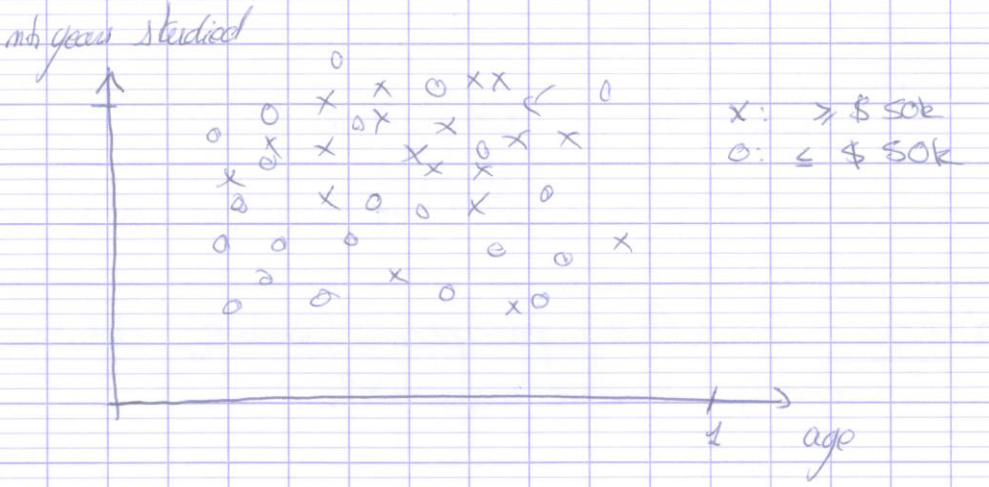
1)

Lecture 5: Learning and optimizat°



$\mathcal{X} = [0,1]^2$; $\mathcal{Y} = \{-1, +1\}$

o x

Aim: find a good classificat° rule $h : \mathcal{X} \to \mathcal{Y}$

Hypothesis: data comes from a probability law $P_{(X,Y)}$

Risk: $R(h) = \mathbb{P}_{(X,Y)}(h(X) \neq Y)$

$\to h^* = \underset{h}{\operatorname{argmin}}\ R(h)$ "Bayes classifier"

$$h^*(x) = \begin{cases} +1 & \text{if } \mathbb{P}(Y=+1 \mid X=x) \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

you don't know $P_{(X,Y)}$ but you have $(X_1, Y_1), \dots (X_m, Y_m)$

$\hat{h}_m = \underset{h}{\operatorname{argmin}}\ \hat{R}_m(h)$ where $\hat{R}_m(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{h(X_i) \neq Y_i\}}$

(• h-nearest-neighbours) $\mathbb{E}_{X_i, Y_i}[\hat{R}_m(h)] = R(h)$

① Assume that the classes are linearly separable

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H} = \{ \operatorname{sign}(\omega^T x) \ ; \ \omega \in \mathbb{R}^2 \}$

$\ln(1+e^{-y\omega x}) = \frac{-y x e^{-y\omega x}}{1+e^{-y\omega x}} = \frac{-y x}{e^{y\omega x}+1}$

$\omega \in \mathbb{R}^p$, $R(\omega) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y_i \omega^t x_i \leq 0\}$

Goal: $\hat{\omega}_m \in \operatorname{argmin} R(\omega)$

pb: $\min_{\omega\in\mathbb{R}^p} \underbrace{\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y_i \omega^t x_i \leq 0\}}_{\ell_R}$

$g_R(\omega)$ not diff. (smooth), not convex.

a) Algo: Rosenblatt's perceptron (Book of Hastie, Friedman, Tibshirani)
→ good for linear separat°.

b) Logistic regression (convexificat° of the Risk!)

Logistic → $\mathbb{P}(Y=1 \mid X=x) = \text{logit}^{-1}(\omega^t x)$

$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$, $\text{logit}^{-1}(u) = \frac{e^u}{1+e^u}$

$\mathbb{P}(Y=-1 \mid X=x) = 1 - \frac{e^{\omega^t x}}{1+e^{\omega^t x}} = \frac{1}{1+e^{\omega^t x}}$

$\mathbb{P}(Y=1 \mid X=x) = \frac{1}{1+e^{-\omega^t x}}$

$g \in \{-1, +1\}$ $\quad \mathbb{P}(Y=g \mid X=x) = \frac{1}{1+e^{-g\omega^t x}}$

ML estimator of ω:

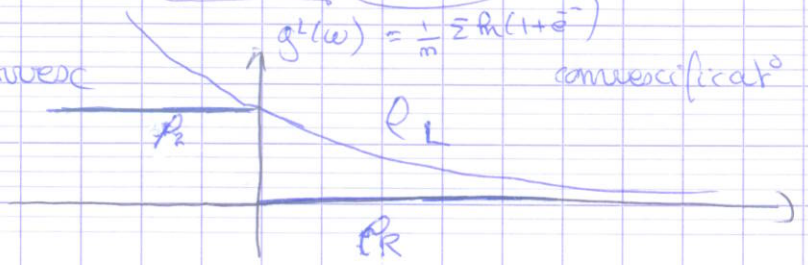
$\hat{\omega}_m^L = \operatorname{argmax}_\omega \log\left(\prod_{i=1}^{m} \mathbb{P}(Y=y_i \mid X=x_i)\right)$

$= \operatorname{argmax}_\omega \sum_{i=1}^{m} \log\left(\frac{1}{1+e^{-y_i\omega^t x_i}}\right)$

$= \operatorname{argmin} \sum_{i=1}^{m} \underbrace{\ln\left(1+e^{-y_i\omega^t x_i}\right)}_{\ell^L}$

$g^L(\omega) = \frac{1}{m}\sum \ln(1+e^{\cdot})$

g^L is C^∞, convex



$$\nabla g^2(w) = \sum_{i=1}^{n} \frac{-x_i y_i}{1+e^{y_i w^T x_i}}$$

② Convex optimization

Reference : . lectures notes by Sebastien Bubeck (Princeton)
. Tutorial at ICML from Bottou, Curtis, Nocedal.

GD: $w_0 \in \mathbb{R}^p$, $w_{k+1} = w_k - \gamma_k \nabla \ell(w_k)$
where $\ell : \mathbb{R}^p \to \mathbb{R}$ the fct to be minimized

Example : $\ell(w) = \frac{1}{2}\|w\|_2^2$
exo: for what values of $\gamma_k = \gamma$ does the GD $\to 0$? What is the optimal value for γ?

$\nabla \ell(w) = w$
$w_0 \in \mathbb{R}^p$ $\quad w_{k+1} = w_k - \gamma w_k = (1-\gamma) w_k$
ie $w_k = (1-\gamma)^k w_0$

if $\gamma > \frac{2}{L}$, $\|w_k\|_2 \to \infty$

if $\gamma = \frac{2}{L}$, $\|w_k\|_2 \to 0$

if $0 < \gamma < 2/L$, $w_k \xrightarrow[k\to\infty]{} 0$, the best γ is $\gamma = \frac{1}{L}$

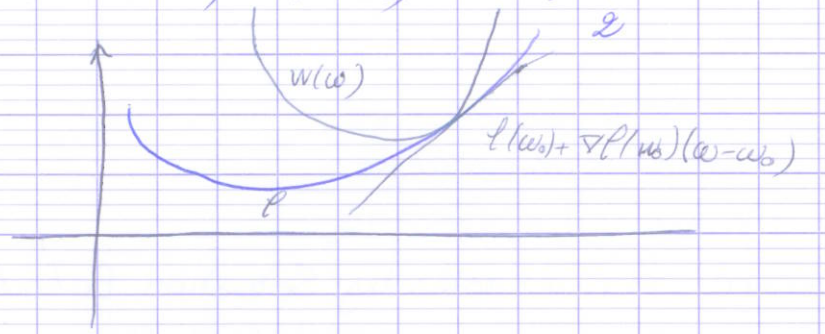
Here $w \mapsto \frac{L}{2}\|w\|_2^2$ has an L-lipschitz gradient

In the sequel, we assume that function ℓ has a L-Lipschitz gradient $\|\nabla \ell(x) - \nabla \ell(y)\| \le L\|x-y\|$ (for example, true with g^2 with $L = \frac{1}{4}$)

Consequence: $\forall \omega_0 \in \mathbb{R}^p$, $\forall \omega \in \mathbb{R}^p$

$$\ell(\omega) \leq \ell(\omega_0) + \nabla \ell(\omega_0)^T (\omega - \omega_0) + \frac{L}{2} \|\omega - \omega_0\|^2$$

$w(\omega)$



Maximizat° - Minimisat°

- $\omega_0 \in \mathbb{R}^p$
- $g_k : \mathbb{R}^p \to \mathbb{R}$

  $g_k(\omega_k) = \ell(\omega_k)$

  $\nabla g_k(\omega_k) = \nabla \ell(\omega_k)$

$\forall \omega$, $g_k(\omega) \geq \ell(\omega)$

- $\omega_{k+1} = \underset{\omega \in \mathbb{R}^p}{\operatorname{argmin}}\ g_k$

If you choose $g_k(\omega) = \ell(\omega_k) + \nabla \ell(\omega_k)^T (\omega - \omega_k) + \frac{L}{2} \|\omega - \omega_k\|^2$

$\nabla g_k(\omega) = \nabla \ell(\omega_k) + L(\omega - \omega_k) = 0$

$$\Leftrightarrow \omega_{k+1} = \omega_k - \frac{1}{L} \nabla \ell(\omega_k)$$

[Proposit° 2.2 in Optimizat° with 1st order Surrogate Fcts by Mairal]

[Nesterov - Complex optimizat°]

Assume that ℓ is convex, and that for some $R > 0$,

$\|\omega - \omega^*\|_2 \leq R$

for all $\omega \in \mathbb{R}^p$ s.t. $\ell(\omega) \leq \ell(\omega_0)$, $\omega_k \in \underset{\omega \in \mathbb{R}^p}{\operatorname{argmin}}\ \ell$

Then $\ell(\omega_k) - \ell(\omega^*) \leq \frac{2L R^2}{k+2}$ $\forall k \geq 1$

Proof: [Nesterov] Define $h_k = g_k - \ell$

Lemma: Then $\ell(w_k) \le \min_{w \in \mathbb{R}^p} \left\{ \ell(w) + \frac{L}{2} \|w - w_{k-1}\|^2 \right\}$

$$\ell(w_k) \le \min_{\alpha \in [0,1]} \ell(\alpha w^* + (1-\alpha) w_{k-1}) + \frac{L}{2} \alpha^2 \|w^* - w_{k-1}\|_2^2$$

convexity of ℓ $\quad \le \min_{\alpha \in [0,1]} \alpha \ell(w^*) + (1-\alpha) \ell(w_{k-1}) + \frac{L\alpha^2}{2} \|w^* - w_{k-1}\|^2$

But (Lemma) the sequence $\ell(w_k)$ is decreasing in k, and has:

$$\ell(w_k) - \ell(w^*) \le \min_{0 \le \alpha \le 1} (1-\alpha)(\ell(w_{k-1}) - \ell(w^*)) + \frac{LR^2\alpha^2}{2}$$

2 cases:

- if $\ell(w_{k-1}) - \ell(w^*) \ge LR^2$

then choosing $\alpha = 1$ yields $\ell(w_k) - \ell(w^*) \le \frac{LR^2}{2}$

- otherwise take $\alpha = \frac{\ell(w_{k-1}) - \ell(w^*)}{LR^2}$

and denote $r_k = \ell(w_k) - \ell(w^*)$

Then:

$$r_k \le r_{k-1} \left(1 - \frac{r_{k-1}}{2LR^2}\right)$$

Trick to upperbound

$$\frac{1}{r_k} = \frac{1}{r_{k-1}} \cdot \frac{1}{1 - \frac{r_{k-1}}{2LR^2}}$$

$$\ge \frac{1}{r_{k-1}} \left(1 + \frac{r_{k-1}}{2LR^2}\right)$$

$$= \frac{1}{r_{k-1}} + \frac{1}{2LR^2}$$

$$\ge \frac{1}{r_1} + \frac{k-1}{2LR^2}$$

$r_k - r_{k-1} \le -\frac{r_k^2}{2LR^2}$

$y' = -\frac{y^2}{2LR^2}$

$\Rightarrow y = \frac{2LR^2}{z}$

$u_k := \frac{1}{r_k}$ or $\Delta_k := u_k - u_{k-1}$ has limit (Cesaro).

Bg case 1: $R_1 \leq \frac{LR^2}{2}$

$$\frac{1}{R_k} \geq \frac{2}{LR^2} + \frac{k-1}{2LR^2} = \frac{k+3}{2LR^2}$$

$$\hookrightarrow R_k \leq \frac{2LR^2}{k+3}$$

□

$|h_k(\omega)| = |g_k^{\omega} - \ell(\omega)|$
$= |\ell(\omega_k) + \nabla\ell(\omega_k)(\omega - \omega^k) + \frac{L}{2}\|\omega - \omega_k\|^2 - \ell(\omega)|$

1) $\ell(\omega) \leq g_k(\omega)$
2) $\ell(\omega_k) - \ell(\omega) + \nabla\ell(\omega_k)(\omega - \omega_k) \leq 0$
True since ℓ is convex (convex $\Leftrightarrow$ def° dérivée directr. + def° convexe)
$\Rightarrow -\frac{L}{2} \leq \ell(\omega_k) - \ell(\omega) + \nabla\ell(\omega_k)(\omega - \omega_k) + \frac{L}{2}$
$\Rightarrow |h_k(\omega)| \leq \frac{1}{2}\|\omega - \omega_k\|^2$

Lemme 2.1: $g_k \geq \ell$ $\quad h_k = g_k - \ell$

- $|h_k(\omega)| \leq \frac{L}{2}\|\omega - \omega_k\|^2$
- $\ell(\omega') \leq \ell(\omega) + \frac{L}{2}\|\omega - \omega_k\|^2$

proof:

$$\ell(\omega_k) \leq g_{k-1}(\omega_k) \leq g_k(\omega) = \ell(\omega) + h_{k-1}(\omega)$$

$$\leq \ell(\omega) + \frac{L}{2}\|\omega - \omega_{k-1}\|^2$$

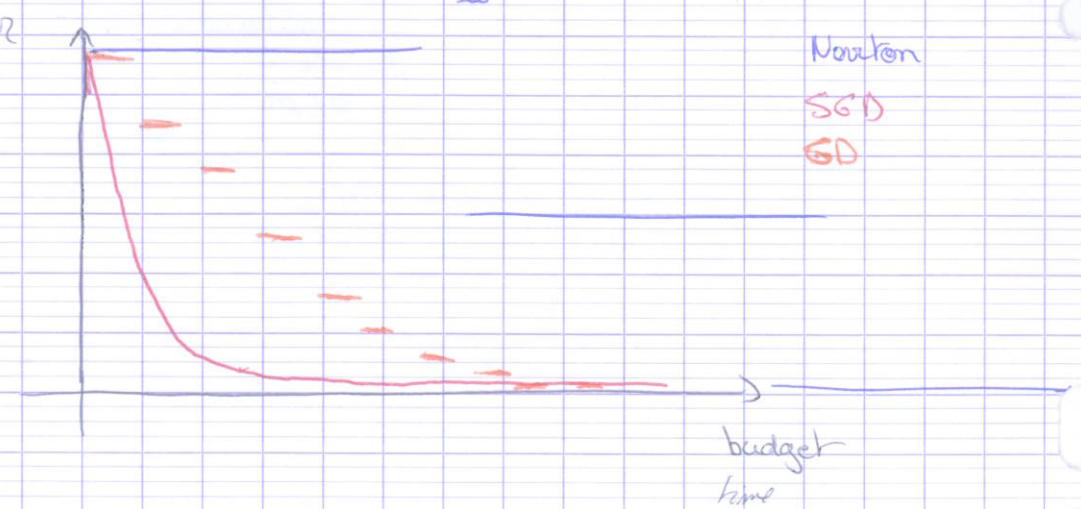
$g_{k-1}(\omega) = \ell(\omega_{k-1}) + \nabla\ell(\omega_{k-1})(\omega - \omega_{k-1}) + \frac{L}{2}\|\omega - \omega_{k-1}\|^2$

Lemma 2.2: The sequence $(\ell(\omega_k))_k$ is non-increasing

proof: $\ell(\omega_k) \leq g_k(\omega_k)$

$\leq g_k(\omega_{k-1}) = \ell(\omega_{k-1})$

□



Other select°:

$$d(\omega, x) = \left| \frac{\omega^t x}{\|\omega\|} \right| , \quad x = x_{\omega^{T\perp}} + x_\omega$$

$$d(\omega, x) = \left| \left(\frac{\omega}{\|\omega\|} \right)^T x_\omega \right|$$

$$= \left| \frac{\omega^t x}{\|\omega\|} \right|$$

$$\Rightarrow \text{Pb} \quad \max_{\omega} \min_{1 \le i \le m} d(\omega, x)$$
$$\text{s.t } \forall i \; y_i \omega^t x_i \ge 1$$

$$\Leftrightarrow \max_{\omega} \frac{1}{\|\omega\|}$$
$$\text{sc } \forall i$$
$$y_i \omega^t x_i \ge 1$$

$$\Leftrightarrow \boxed{\begin{array}{l} \min \frac{1}{2} \|\omega\|^2 \\ \text{sc } \forall i \\ y_i \omega^t x_i \ge 1 \end{array}}$$

SVM (support vector machine)

simpler if ω low dim.

Lagrangian:

$$L(\omega, (\alpha_i)_i) = \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^m \alpha_i (y_i \omega^t x_i - 1)$$

$$\nabla L(\omega) = \omega - \sum_{i=1}^m \alpha_i y_i x_i \quad \Rightarrow \hat{\omega} = \sum_{i=1}^m \alpha_i y_i x_i$$

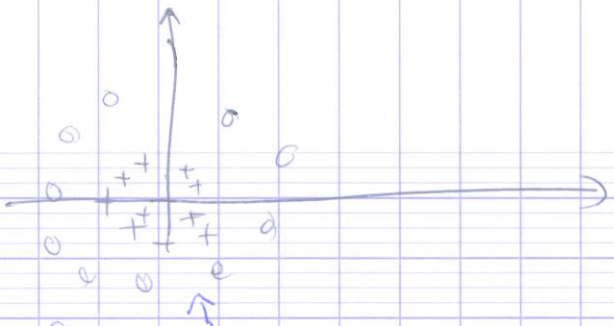
and $\forall i, \; \alpha_i (y_i \omega^t x_i - 1) = 0$

Dual formulation of SVM:

$$\max_{\alpha_1, \dots, \alpha_m} \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i x_i \right\|^2 - \sum_{i=1}^m \alpha_i \left(y_i \left(\sum_j \alpha_j y_j x_j^t \right) x_i \right)$$

simpler if ω high dim $\approx -\alpha^t K \alpha + L^t \alpha$

K, L depends on (x_i) only through the scalar product $(x_i^t x_j)_{i,j}$



In the case $\mathbb{R}^2$:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{pmatrix}$$

$$\Phi : \mathcal{X} \to \mathcal{X}' \subset \mathbb{R}^{p'} \quad p' >> p$$
$$x \mapsto \Phi(x)$$

↪ you only need to compute the $\left(\langle \Phi(x_i), \Phi(x_j) \rangle \right)_{i,j}$ in order to find the SVM solut°

kernel trick

→ you just need to give a fct

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$
$$(x, x') \to k(x, x')$$
$$\| \qquad \langle \Phi(x), \Phi(x') \rangle$$

For example: $k(x, x) = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$

Remark:

$$w_n^S = \operatorname{argmin} \frac{\|w\|^2}{2} + \frac{C}{n} \sum_{i=1}^{n} (1 - y_i w^T x_i)_+$$

↑ penalisat°

$$\ell^S(u) = (1-u)_+$$

Not done: boosting.