MAP 553
Apprentissage statistique

Christophe Giraud

Université Paris Sud et Ecole Polytechnique

PC9

Convexification

Notations

Aujourd'hui :

- les étiquettes y_i appartiennent à $\{-1, +1\}$
- les classifieurs h sont à valeurs dans $\{-1, +1\}$:

$$h : \mathcal{X} \to \{-1, +1\}$$

Minimiseur du risque empirique

A partir d'observations $(x_i, y_i)_{i=1,\dots,n}$ et pour un ensemble $\mathcal{H}$ de classifieurs,

$$\hat{h}_{\mathcal{H}} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_n(h), \quad \text{où} \quad \hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\mathbb{R}^+}(-y_i h(x_i))$$

En pratique

1. $\mathcal{H}$ non convexe,
2. $\hat{R}_n(h)$ non convexe.

Temps de calcul prohibitif !

Minimiseur du risque empirique

A partir d'observations $(x_i, y_i)_{i=1,\dots,n}$ et pour un ensemble $\mathcal{H}$ de classifieurs,

$$\hat{h}_{\mathcal{H}} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_n(h), \quad \text{où} \quad \hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\mathbb{R}^+}(-y_i h(x_i))$$

En pratique

1. $\mathcal{H}$ non convexe,
2. $\hat{R}_n(h)$ non convexe.

Temps de calcul prohibitif !

Deux sources de difficultés

1. $\mathcal{H}$ non convexe,
2. $\hat{R}_n(h)$ non convexe.

Convexification du problème

Pour

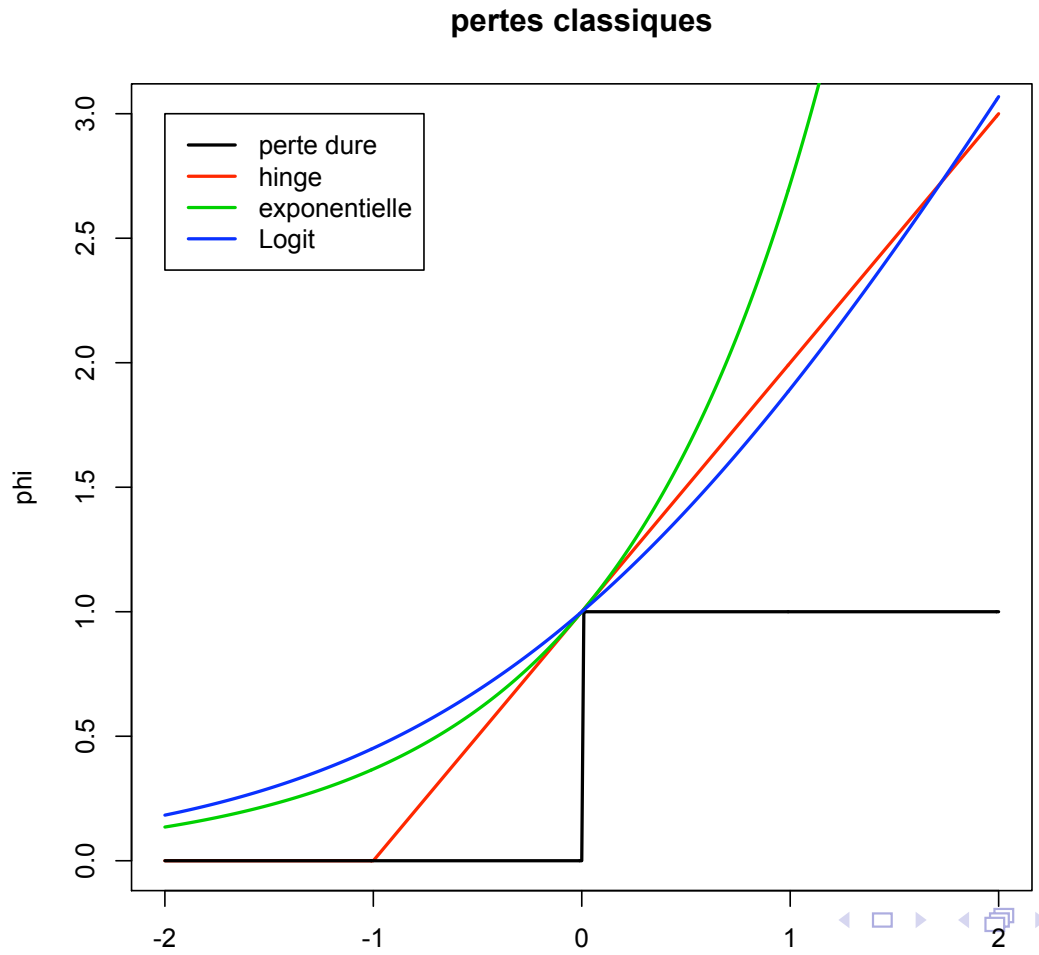
- $\mathcal{F}$ un ensemble convexe de fonction $f : \mathcal{X} \to \mathbb{R}$
- et $\varphi : \mathbb{R} \to \mathbb{R}^+$ convexe croissante

on définit :

$$\hat{h}_{\varphi,\mathcal{F}} = \text{signe}(\hat{f}_{\varphi,\mathcal{F}}) \quad \text{avec} \quad \hat{f}_{\varphi,\mathcal{F}} = \underset{f \in \mathcal{F}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \varphi(-y_i f(x_i))$$

Quelques φ populaires

- **Perte hinge :** $\varphi(x) = (1 + x)_+$
- **Perte exponentielle :** $\varphi(x) = e^x$
- **Perte Logit :** $\varphi(x) = \log_2(1 + e^x)$



Quelques $\mathcal{F}$ populaires

- **Convexification d'un ensemble de classifieurs de bases** $\{h_1, \ldots, h_M\}$ **:**

$$\mathcal{F} = \left\{ f = \sum_{j=1}^{M} \theta_j h_j \ : \ \theta \in \Theta \right\}$$

avec Θ un ensemble convexe de $\mathbb{R}^M$.

- **Boule d'un RKHS** $\mathcal{W}$ **:** pour $R > 0$

$$\mathcal{F} = \{ f \in \mathcal{W} \ : \ |f|_{\mathcal{W}} \leq R \} .$$

SVM

Les SVM correspondent au choix

- $\varphi(x) = (1+x)_+$
- $\mathcal{F} = \{f \in \mathcal{W} \ : \ |f|_{\mathcal{W}} \leq R\}$, avec $\mathcal{W}$ un RKHS et $R > 0$.

SVM : version lagrangienne

le classifieur $\hat{h}_{\varphi,\mathcal{F}}$ est donné par $\hat{h}_{\varphi,\mathcal{F}}(x) = \text{signe}(\hat{f}_{\varphi,\mathcal{F}}(x))$ avec

$$\hat{f}_{\varphi,\mathcal{F}} = \underset{f \in \mathcal{W}}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda |f|_{\mathcal{W}}^2 \right\}$$

Rappels d'optimisation convexe

Considérons $f, -g_1, \ldots, -g_n$ convexes de classe $\mathcal{C}^1$ et

$$\hat{x} = \underset{g_i(x) \geq 0}{\operatorname{argmin}} f(x).$$

Conditions nécessaires de Kuhn-Tucker

Soit

$$L(x, \lambda) = f(x) - \sum_{i=1}^{n} \lambda_i g_i(x).$$

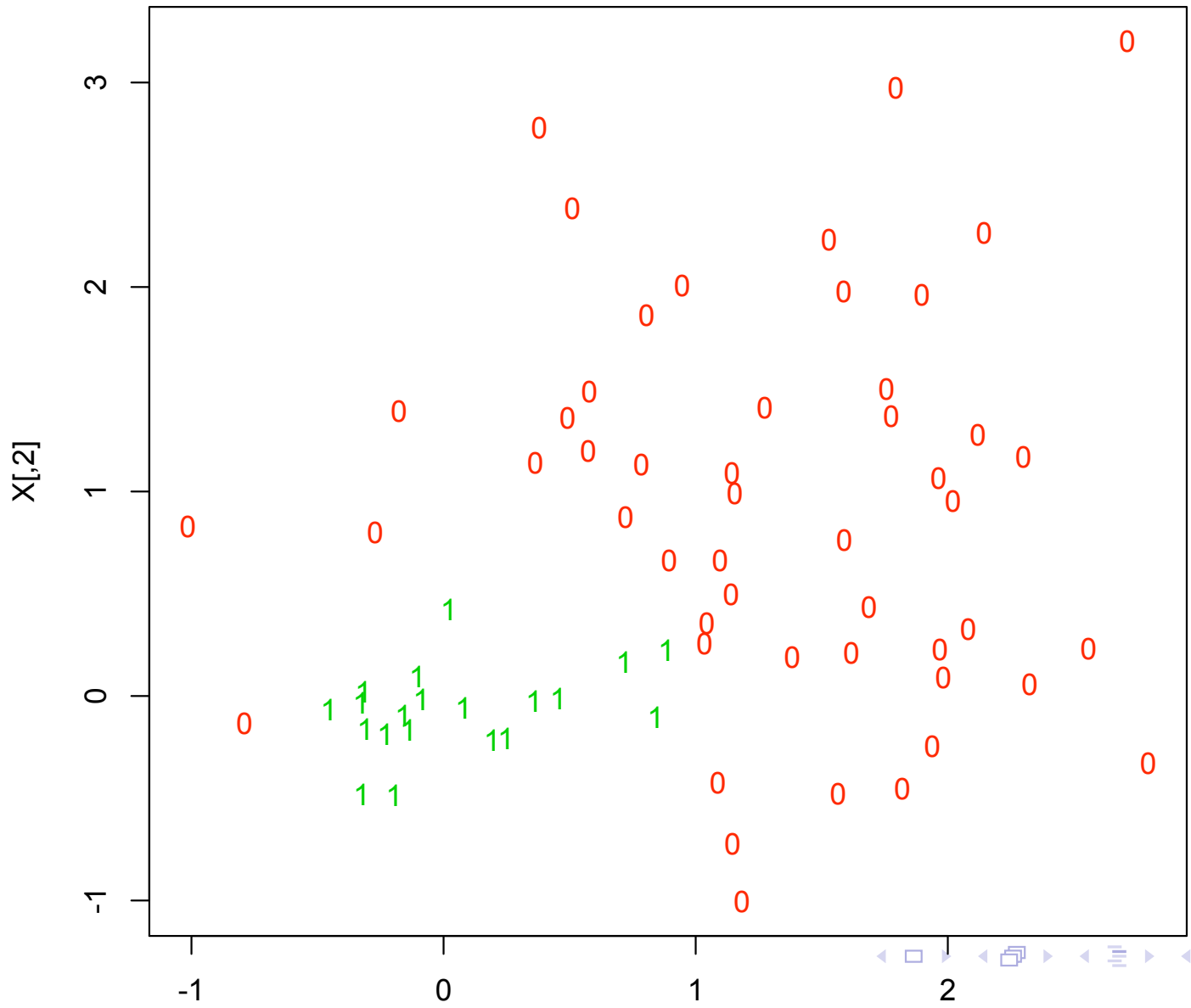
Il existe $\hat{\lambda}$ tel que

1. $\nabla_x L(\hat{x}, \hat{\lambda}) = 0$
2. $\min(\hat{\lambda}_i, g_i(\hat{x})) = 0$ pour $i = 1, \ldots, n$

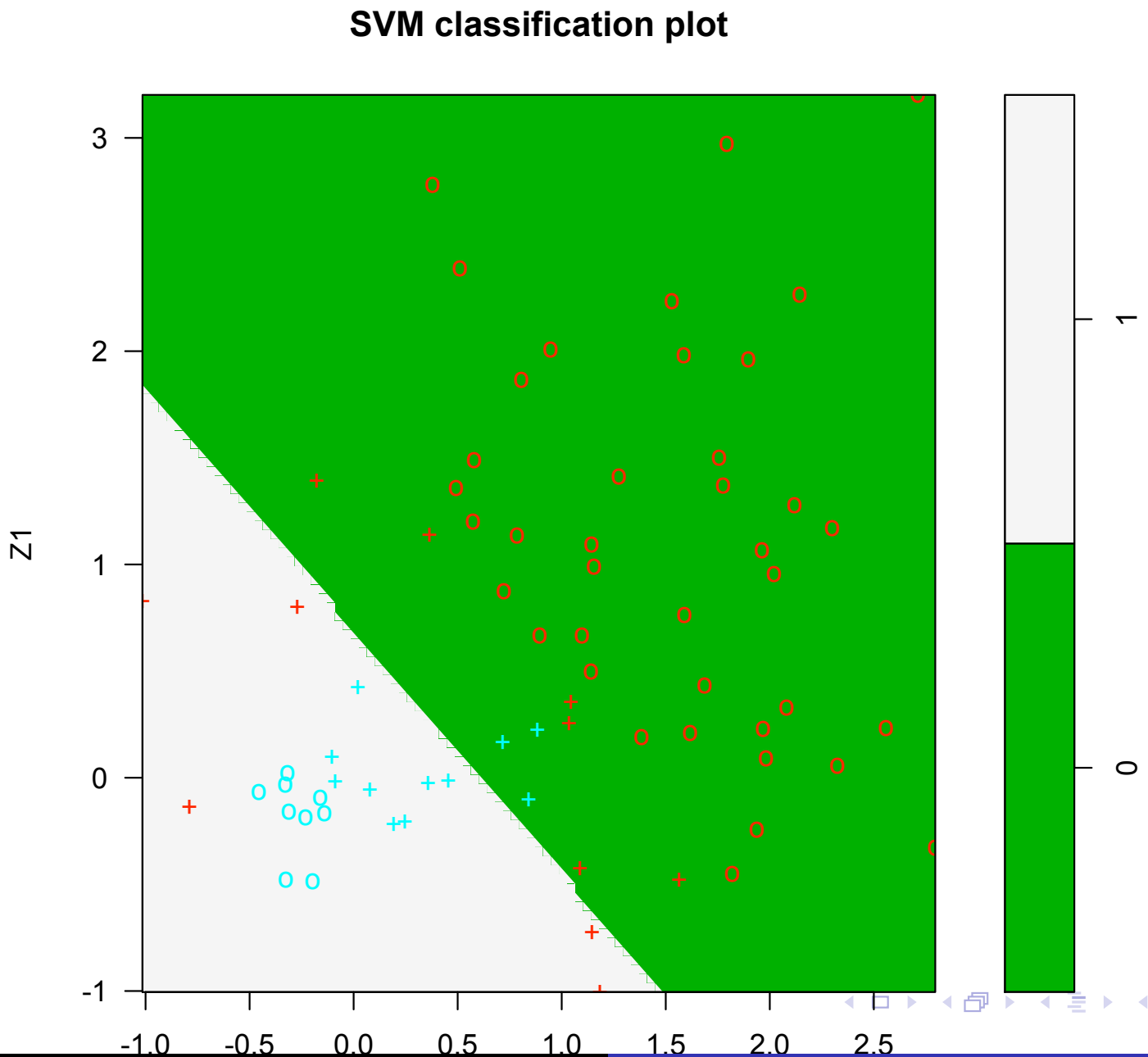
Dualité forte

$$\hat{\lambda} = \underset{\lambda \geq 0}{\operatorname{argsup}} \inf_x L(x, \lambda)$$

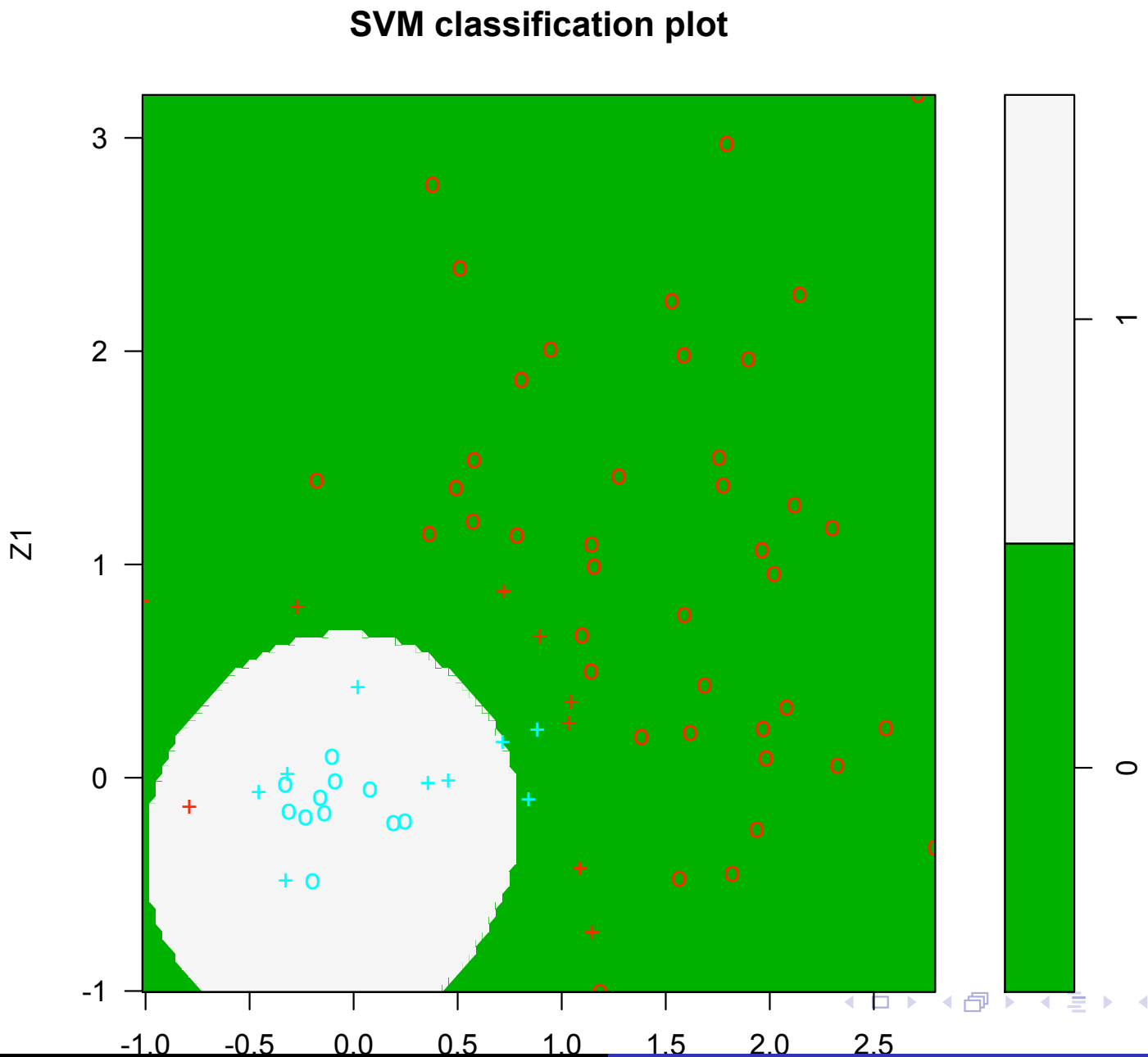
Data :



Linear kernel : les points "+" sont les vecteurs supports



Gaussian kernel : les points "+" sont les vecteurs supports



Dualité forte

$$
\begin{aligned}
(\hat{\alpha}, \hat{\gamma}) &\in \underset{(\alpha,\gamma)\geq 0}{\operatorname{argmax}} \ \min_{\beta,\xi} \left\{ \frac{1}{\lambda}\langle K\beta, \beta\rangle - \langle K\beta, y.\alpha\rangle + \langle \alpha, 1\rangle + \langle \xi, \frac{1}{n} - \alpha - \gamma\rangle \right\} \\
&\in \underset{(\alpha,\gamma)\geq 0}{\operatorname{argmax}} \ \min_{\xi} \left\{ -\frac{1}{4\lambda}\langle K(y.\alpha), y.\alpha\rangle + \langle \alpha, 1\rangle + \langle \xi, \frac{1}{n} - \alpha - \gamma\rangle \right\} \\
&\in \underset{(\alpha,\gamma)\geq 0 \ \&\ \alpha+\gamma=\frac{1}{n}}{\operatorname{argmax}} \left\{ -\frac{1}{4\lambda}\langle K(y.\alpha), y.\alpha\rangle + \langle \alpha, 1\rangle \right\} \\
&\in \underset{0\leq\alpha\leq\frac{1}{n} \ \&\ \gamma=\frac{1}{n}-\alpha}{\operatorname{argmax}} \left\{ -\frac{1}{4\lambda}\langle K(y.\alpha), y.\alpha\rangle + \langle \alpha, 1\rangle \right\}
\end{aligned}
$$

AdaBoost

AdaBoost fournit une solution approximative au problème

$$\hat{f}_{\varphi,\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi(-y_i f(x_i)) \right\}$$

avec

- $\varphi(x) = \exp(x)$
- $\mathcal{F} = \left\{ f = \sum_{j=1}^{M} \theta_j h_j \ : \ \theta \in \mathbb{R}^M \right\}$

« forward learning »

Initialisation : $\hat{f}_0 = 0$

Itérer : For $m = 1, \ldots, M$ do :

$$\hat{f}_m = \hat{f}_{m-1} + \beta_m h_{j_m} \quad \text{où}$$

$$(\beta_m, h_{j_m}) = \underset{\substack{h \in \{h_1, \ldots, h_M\} \\ \beta \in \mathbb{R}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(-y_i\left(\hat{f}_{m-1}(x_i) + \beta h(x_i)\right)\right).$$

AdaBoost

AdaBoost fournit une solution approximative au problème

$$\hat{f}_{\varphi,\mathcal{F}} = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \varphi(-y_i f(x_i)) \right\}$$

avec

- $\varphi(x) = \exp(x)$
- $\mathcal{F} = \left\{ f = \sum_{j=1}^{M} \theta_j h_j \ : \ \theta \in \mathbb{R}^M \right\}$

≪ forward learning ≫

Initialisation : $\hat{f}_0 = 0$

Itérer : For $m = 1, \ldots, M$ do :

$$\hat{f}_m = \hat{f}_{m-1} + \beta_m h_{j_m} \quad \text{où}$$

$$(\beta_m, h_{j_m}) = \underset{\substack{h\in\{h_1,\ldots,h_M\}\\ \beta\in\mathbb{R}}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \varphi\left(-y_i\left(\hat{f}_{m-1}(x_i) + \beta h(x_i)\right)\right).$$

Solution AdaBoost

$$\frac{1}{n}\sum_{i=1}^{n}\varphi\Big(-y_i\big(\hat{f}_{m-1}(x_i)+\beta h(x_i)\big)\Big) = (e^{\beta}-e^{-\beta})\sum_{i=1}^{n} w_i^{(m)}\mathbf{1}_{h(x_i)\neq y_i} + e^{-\beta}\sum_{i=1}^{n} w_i^{(m)}$$

avec $w_i^{(m)} = n^{-1}\exp(-y_i\hat{f}_{m-1}(x_i))$

Solution

Si

$$\mathrm{err}_m(h) = \frac{\sum_{i=1}^{n} w_i^{(m)}\mathbf{1}_{h(x_i)\neq y_i}}{\sum_{i=1}^{n} w_i^{(m)}} > 0 \quad \text{pour tout } h$$

on a

$$h_{j_m} = \underset{h\in\{h_1,\ldots,h_M\}}{\operatorname{argmin}} \ \mathrm{err}_m(h) \quad \text{et} \quad \beta_m = \frac{1}{2}\log\left(\frac{1-\mathrm{err}_m(h_{j_m})}{\mathrm{err}_m(h_{j_m})}\right).$$

Solution AdaBoost

$$\frac{1}{n}\sum_{i=1}^{n}\varphi\Big(-y_i\big(\hat{f}_{m-1}(x_i)+\beta h(x_i)\big)\Big) = (e^{\beta}-e^{-\beta})\sum_{i=1}^{n} w_i^{(m)}\mathbf{1}_{h(x_i)\neq y_i} + e^{-\beta}\sum_{i=1}^{n} w_i^{(m)}$$

avec $w_i^{(m)} = n^{-1}\exp(-y_i\hat{f}_{m-1}(x_i))$

Solution

Si

$$\mathrm{err}_m(h) = \frac{\sum_{i=1}^{n} w_i^{(m)}\mathbf{1}_{h(x_i)\neq y_i}}{\sum_{i=1}^{n} w_i^{(m)}} > 0 \quad \text{pour tout } h$$

on a

$$h_{j_m} = \underset{h\in\{h_1,\ldots,h_M\}}{\operatorname{argmin}} \mathrm{err}_m(h) \quad \text{et} \quad \beta_m = \frac{1}{2}\log\left(\frac{1-\mathrm{err}_m(h_{j_m})}{\mathrm{err}_m(h_{j_m})}\right).$$

Algorithme AdaBoost

AdaBoost

Initialisation : $w_i^{(1)} = 1/n$, pour $i = 1, \ldots, n$

Itérer : For $m = 1, \ldots, M$

$$\begin{aligned} h_{j_m} &= \operatorname*{argmin}_{h \in \{h_1, \ldots, h_M\}} \operatorname{err}_m(h) \\ 2\beta_m &= \log(1 - \operatorname{err}_m(h_{j_m})) - \log(\operatorname{err}_m(h_{j_m})) \\ w_i^{(m+1)} &= w_i^{(m)} \exp(2\beta_m \mathbf{1}_{h_{j_m}(x_i) \neq y_i}), \quad i = 1, \ldots, n \end{aligned}$$

Résultat : $\hat{f}_M(x) = \sum_{m=1}^{M} \beta_m h_{j_m}(x)$.

Bonne fin d'année !