MAP 553
Apprentissage statistique

Christophe Giraud

Université Paris Sud et Ecole Polytechnique

PC8

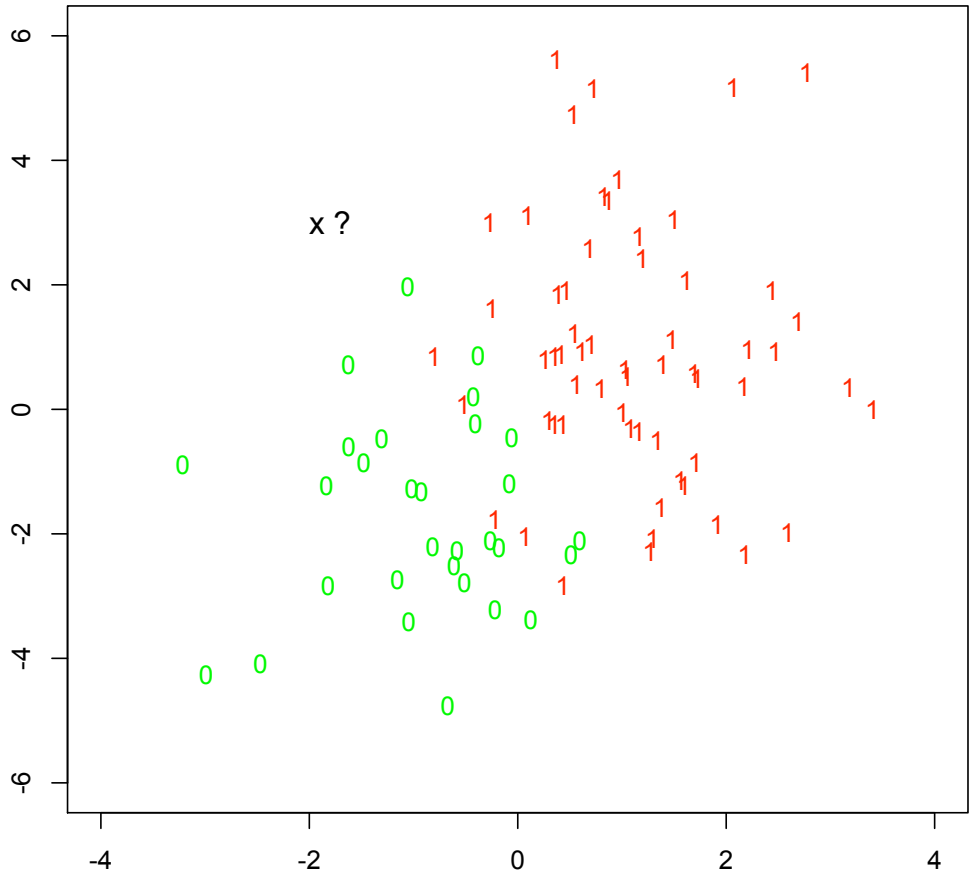
Classification supervisée

Quelques exemples

1. Détection de spam (par analyse du texte)
2. Autorisation d'octroi de carte de crédit (à partir du profil du client : age, revenus, antécédants, sexe, localisation, dettes, etc)
3. Prédire les patients à fort risque dans un service d'urgence (à partir de mesures physiologiques)
4. Détection de clients potentiels (à partir de son profil)
5. Segmentation / catégorisation d'images (face recognition, etc)
6. Classification de cancers à partir de profils mRNA
7. etc

Situation

Observations: points $X_i \in \mathcal{X}$ avec label $Y_i \in \{0, 1\}$ pour $i = 1, \ldots, n$.



Objectif: prédire la classe d'un nouveau point x.

Formalisation

Classifieur: $h : \mathcal{X} \to \{0, 1\}$.

Risque: $R(h) = \mathbb{P}(h(X) \neq Y)$

Classifieur de Bayes: $h_*(x) = \mathbf{1}_{\mathbb{P}[Y=1|X=x]>1/2}$

Problème: loi de (X, Y) inconnue. On ne dispose que d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$.

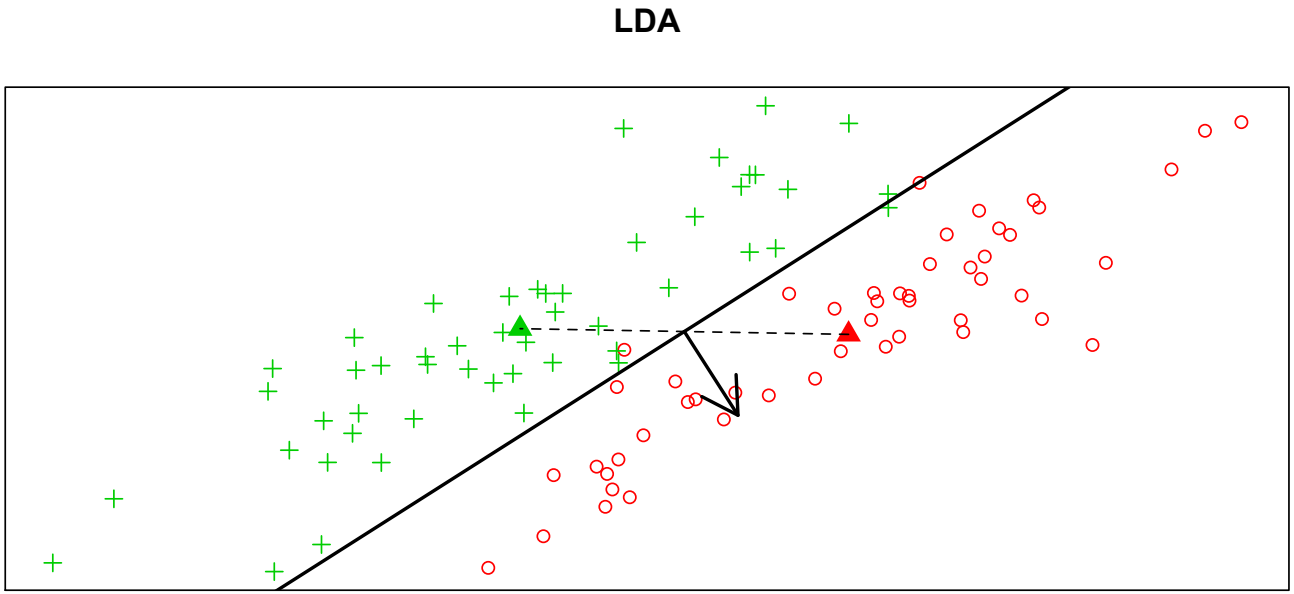
Approche paramétrique

Approche 1: modélisation paramétrique de la loi de (X, Y)

Exemple: Mélanges gaussiens

Modèle

- $\mathbb{P}(Y_i = k) = \pi_k$, pour $k = 0, 1$
- $\text{Loi}(X_i | Y_i = k) = \mathcal{N}(\mu_k, \Sigma_k)$, pour $k = 0, 1$.



Classifieur de Bayes

Lorsque $\Sigma_0 = \Sigma_1 = \Sigma$ on a

$$h_*(x) = 1 \iff \left(x - \frac{\mu_1 + \mu_0}{2}\right)^T \Sigma^{-1}(\mu_1 - \mu_0) > \log(\pi_0/\pi_1).$$

En pratique: on estime μ_0, μ_1 et Σ à partir des données par max de vraisemblance

Approche semi-paramétrique

$h_*(x) = \mathbf{1}_{\mathbb{P}[Y=1|X=x]>1/2}$

Approche 2: modélisation de la loi conditionnelle de Y sachant X

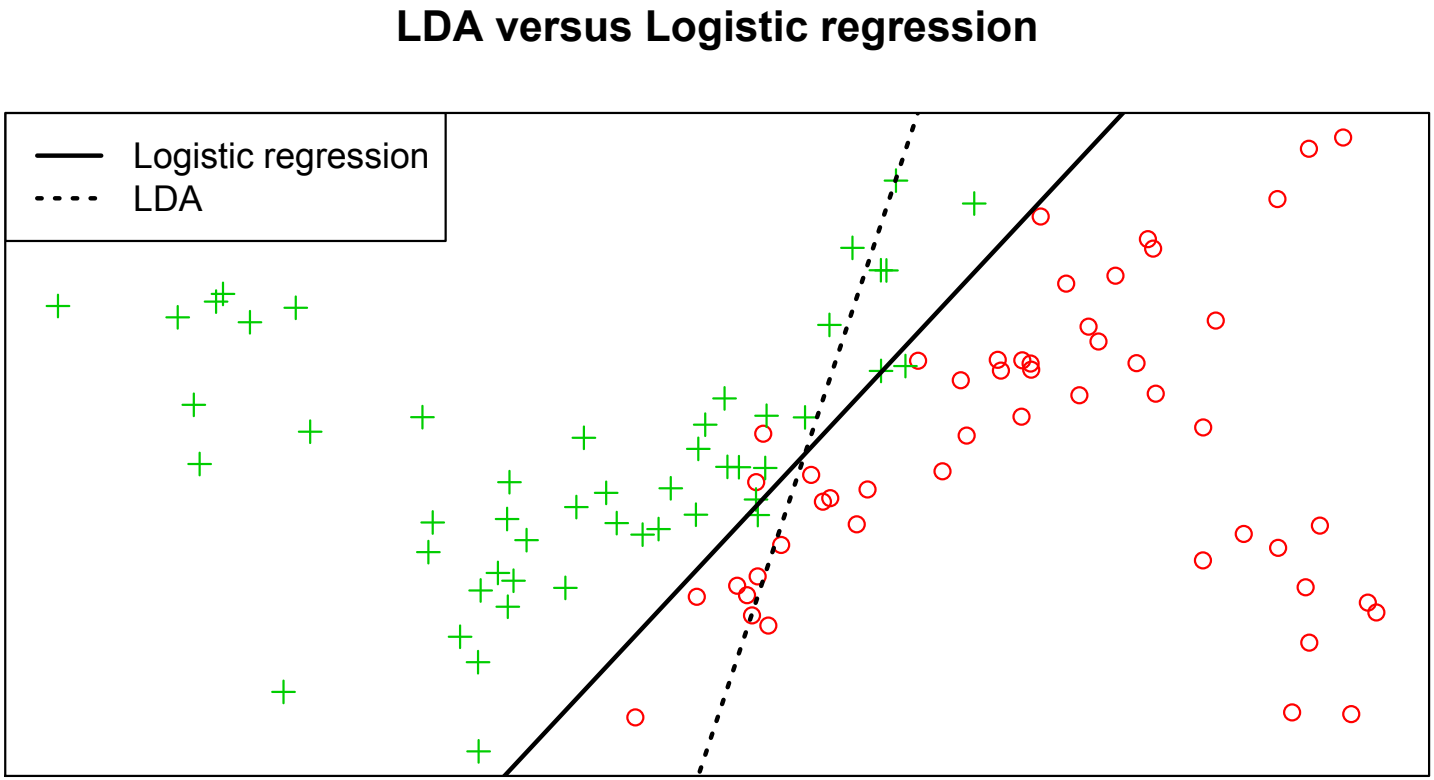
Régression logistique

$$\mathbb{P}[Y = 1|X = x] = \frac{\exp(\alpha + \langle \beta, x \rangle)}{1 + \exp(\alpha + \langle \beta, x \rangle)}$$

Classifieur de Bayes

$$h_*(x) = 1 \iff \alpha + \langle \beta, x \rangle > 0$$

Régression logistique



- On estime α, β à partir des données par max de vraisemblance
- Possibilité de faire de la sélection de variables prédictives avec une pénalité ℓ^1 sur β

Approche non-paramétrique

Modèle: pas de modèle paramétrique pour $\mathbb{P}[Y = 1|X = x]$.

Hypothèse: données (X_i, Y_i) i.i.d. pour $i = 1, \ldots, n$

Minimiseur du risque empirique

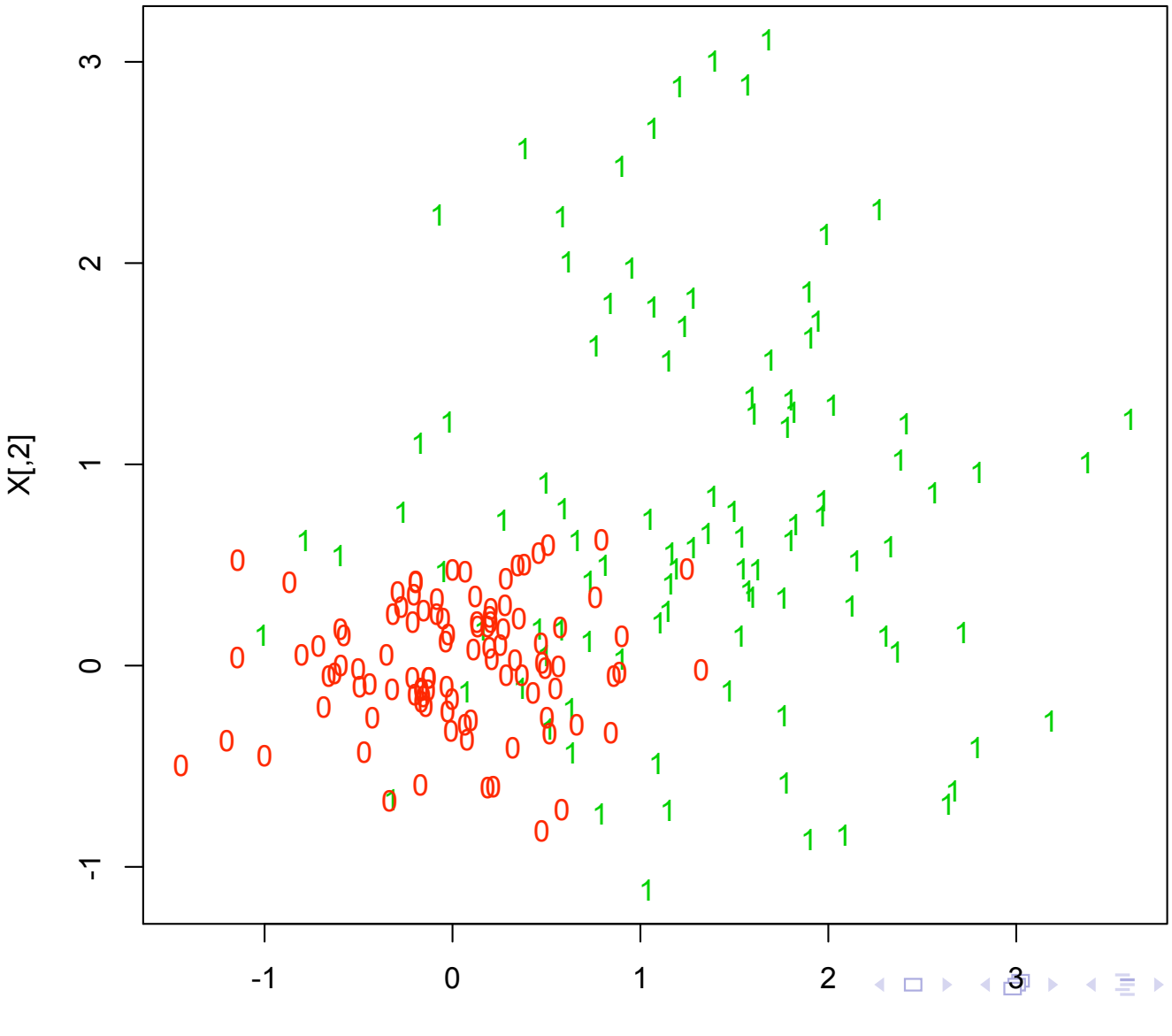
A partir d'observations $(X_i, Y_i)_{i=1,\ldots,n}$ et pour un ensemble $\mathcal{H}$ de classifieurs,

$$\hat{h}_{\mathcal{H}} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_n(h), \quad \text{où} \quad \hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{h(X_i) \neq Y_i}$$

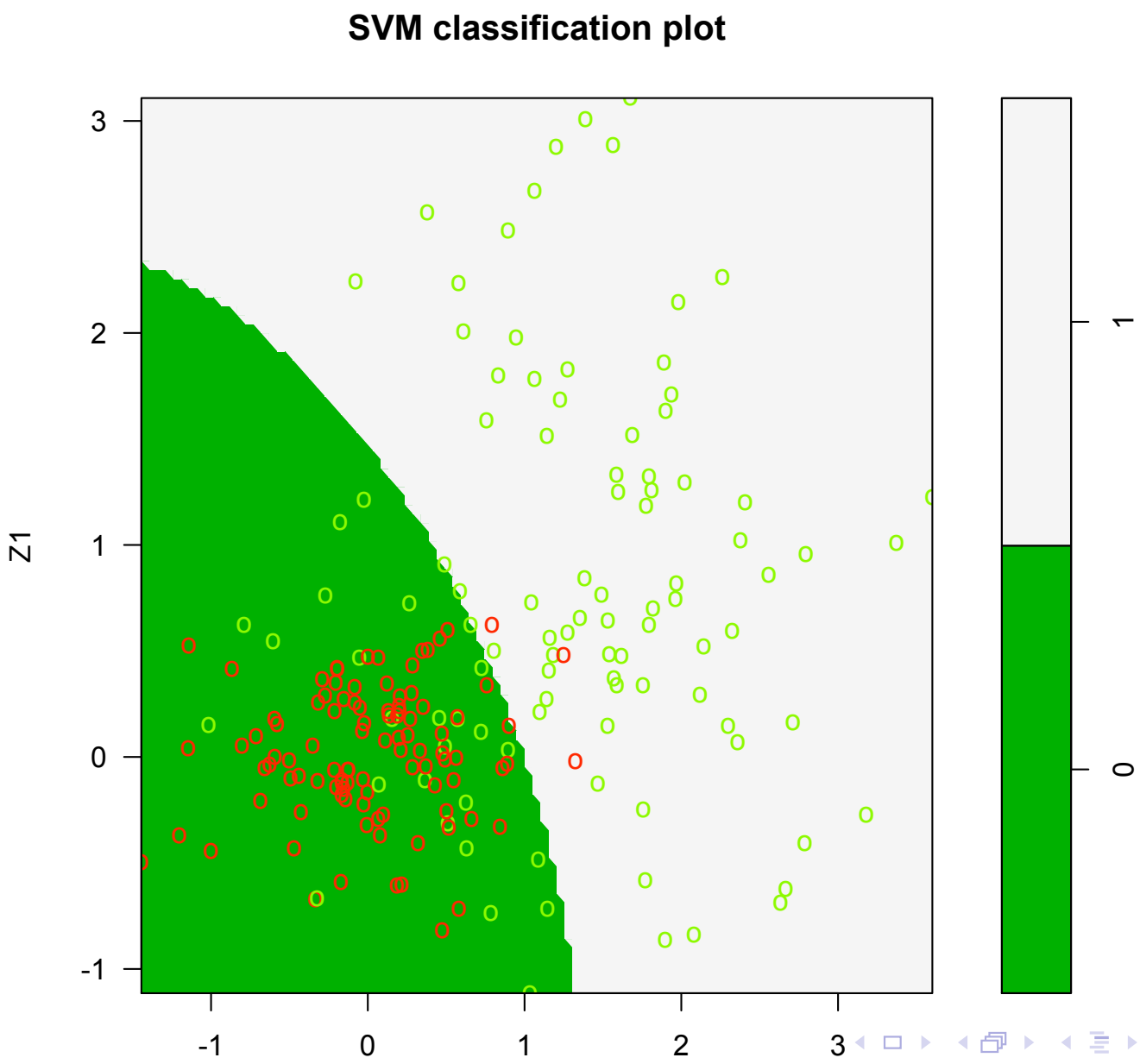
Quel $\mathcal{H}$ choisir?

$\longleftrightarrow$ correspond implicitement au modèle pour $h_*(x)$

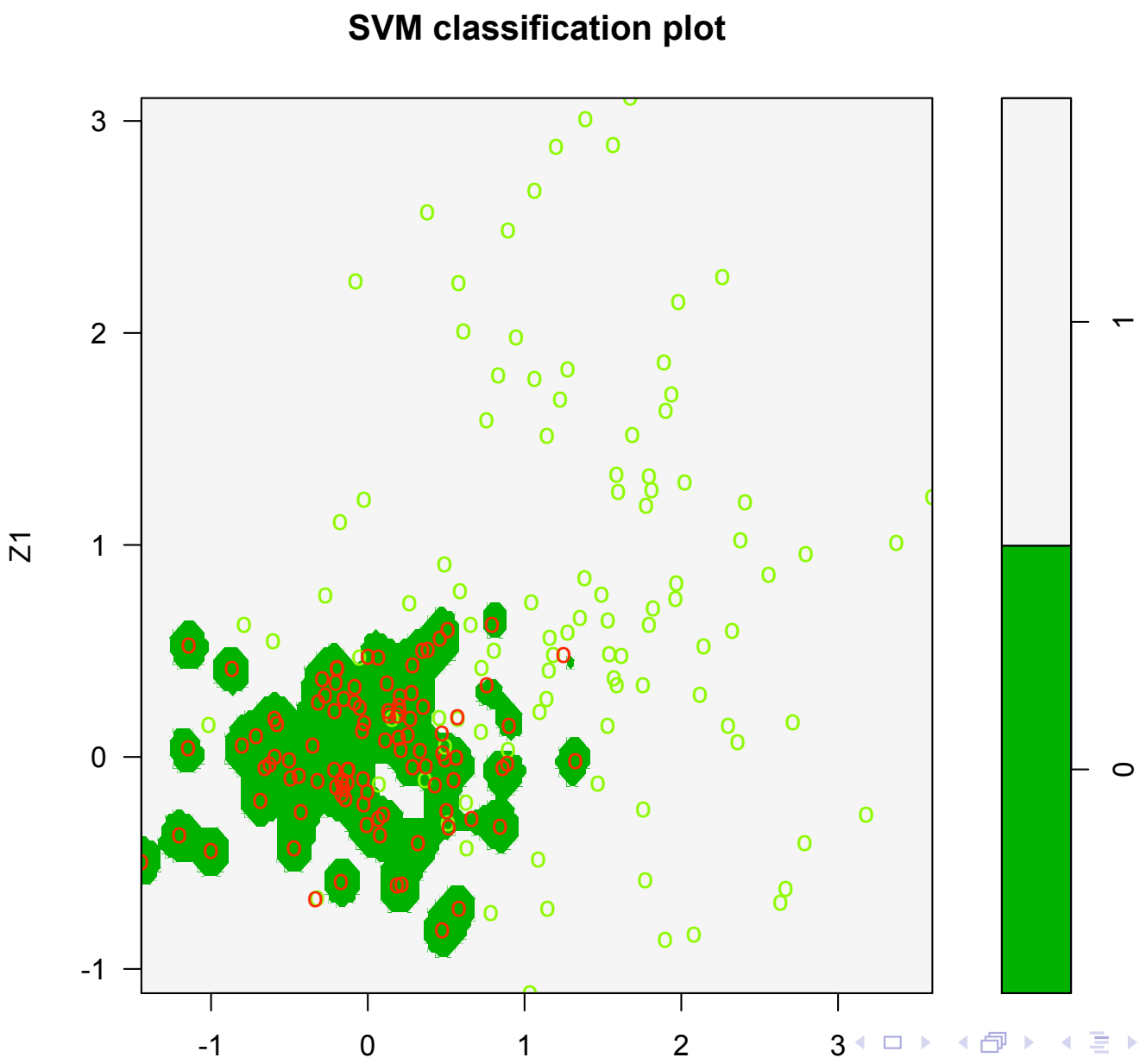
data:



$\mathcal{H}$ **trop petit:** trop rigide.



$\mathcal{H}$ **trop gros:** sur-apprentissage!



Questions

1. Que vaut $R(\hat{h}_{\mathcal{H}})$? Peut-on le comparer à $\min_{h \in \mathcal{H}} R(h)$?

2. Comment choisir entre différentes familles $\mathcal{H}_k$, $k = 1, \ldots, K$?

VC-dimension

Dictionnaire fini

Cas $\mathcal{H} = \{h_1, \dots, h_M\}$: avec probabilité $1 - e^{-L}$

$$R(\hat{h}_{\mathcal{H}}) \leq \min_{j=1,\dots,M} R(h_j) + \sqrt{\frac{2\log(2M) + 2L}{n}}$$

Intuitivement:

$$R(\hat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} R(h) \leq 2 \sup_{h \in \mathcal{H}} |\hat{R}_n(h) - R(h)|$$

avec pour tout $h \in \mathcal{H}$:

$$\hat{R}_n(h) = \frac{1}{n}\mathcal{B}in(n, R(h)) \overset{\text{TCL}}{\approx} R(h) + \frac{Z_h}{\sqrt{n}} + O\left(\frac{1}{n}\right)$$

où $Z_h \sim \mathcal{N}(0, \sigma_h^2)$, avec $\sigma_h^2 = R(h)(1 - R(h)) \leq 1/4$.

Dictionnaire infini

Coefficient d'éclatement

n-ième coefficient d'éclatement d'une famille $\mathcal{H}$ de classifieurs:

$$\mathbb{S}_{\mathcal{H}}(n) = \max_{x_1,\dots,x_n \in \mathcal{X}} \text{Card}\big\{(h(x_1),\dots,h(x_n)) : h \in \mathcal{H}\big\} \leq 2^n.$$

$\mathbb{S}_{\mathcal{H}}(n)$ =nombre maximum de "classifications de n points" possibles à partir des classifieurs dans $\mathcal{H}$

VC-dimension

VC-dimension de $\mathcal{H}$:

$$V_{\mathcal{H}} = \sup\{n \in \mathbb{N} : \mathbb{S}_{\mathcal{H}}(n) = 2^n\}.$$

$V_{\mathcal{H}}$ est donc le nombre maximum de points que $\mathcal{H}$ peut "éclater".

Théorème 3.6

Avec probabilité $1 - e^{-L}$ on a

$$R(\hat{h}_{\mathcal{H}}) \leq \min_{h \in \mathcal{H}} R(h) + 4\sqrt{\frac{2V_{\mathcal{H}} \log(n+1)}{n}} + \sqrt{\frac{2L}{n}}$$

et

$$|R(\hat{h}_{\mathcal{H}}) - \hat{R}_n(\hat{h}_{\mathcal{H}})| \leq 2\sqrt{\frac{2V_{\mathcal{H}} \log(n+1)}{n}} + \sqrt{\frac{L}{2n}}.$$

Sélection de dictionnaire

Sélection de $\mathcal{H}_k$: minimisation du risque structurel

$$\hat{k} = \underset{k=1,\dots,K}{\operatorname{argmin}} \left\{ \hat{R}_n(\hat{h}_{\mathcal{H}_k}) + \operatorname{pen}(\mathcal{H}_k) \right\}$$

où

$$\operatorname{pen}(\mathcal{H}_k) = 2\sqrt{\frac{2V_{\mathcal{H}_k} \log(n+1)}{n}}\,.$$

Corollaire

Avec probabilité $1 - e^{-L}$ on a

$$R(\hat{h}_{\mathcal{H}_{\hat{k}}}) \leq \min_k \left\{ R(\hat{h}_{\mathcal{H}_k}) + 4\sqrt{\frac{2V_{\mathcal{H}_k} \log(n+1)}{n}} \right\} + \sqrt{\frac{2L + 2\log(K)}{n}}$$

Preuve: Notons $\hat{h}_k = \hat{h}_{\mathcal{H}_k}$ et $\Delta_K(L) = \sqrt{\frac{L+\log(K)}{2n}}$.
D'après le Théorème 3.6, avec probabilité $1 - e^{-L}$ on a

$$|R(\hat{h}_k) - \hat{R}_n(\hat{h}_k)| \leq \text{pen}(\mathcal{H}_k) + \Delta_K(L) \quad \text{pour tout } k = 1, \ldots, K.$$

Il en découle: avec probabilité $1 - e^{-L}$ on a

$$\begin{aligned} R(\hat{h}_{\hat{k}}) &\leq \hat{R}_n(\hat{h}_{\hat{k}}) + \text{pen}(\mathcal{H}_{\hat{k}}) + \Delta_K(L) \\ &\leq \min_k \{\hat{R}_n(\hat{h}_k) + \text{pen}(\mathcal{H}_k)\} + \Delta_K(L) \\ &\leq \min_k \{R(\hat{h}_k) + 2\text{pen}(\mathcal{H}_k)\} + 2\Delta_K(L) \end{aligned}$$

□

Sélection de $\mathcal{H}$ en pratique

V-fold Cross-Validation

En pratique: subdiviser les données en V groupes

1. apprendre $\hat{h}_{\mathcal{H}_k}$ sur $V-1$ groupes "train"
2. tester $\hat{h}_{\mathcal{H}_k}$ sur le groupe "test" restant
3. recommencer en permutant les groupes "train" et "test"
4. garder le $\hat{h}_{\mathcal{H}_k}$ ayant le plus petit taux d'erreur sur les V tests.

Exemple: 5-fold CV

train	train	train	train	test
train	train	train	test	train
train	train	test	train	train
train	test	train	train	train
test	train	train	train	train