MAP 569
Machine Learning II

Christophe Giraud

PC2

Supervised Classification

Supervised Classification

Alternative modeling

Parametric modeling

Modeling of the X_i by a mixture of Gaussians $\Longrightarrow$ LDA.

Semi-parametric modeling

Modeling of the distribution on Y given X $\Longrightarrow$ logistic regression.

Non-parametric modeling

For a given set $\mathcal{H}$ of classifiers, take the empirical risk minimizer

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

Empirical risk minimizer

For some observations $(x_i, y_i)_{i=1,\dots,n}$ and a set $\mathcal{H}$ of classifiers

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

In pratice

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

Prohibitive computational complexity !

Empirical risk minimizer

For some observations $(x_i, y_i)_{i=1,\dots,n}$ and a set $\mathcal{H}$ of classifiers

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

In pratice

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

Prohibitive computational complexity !

Two issues

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

Convexification

For

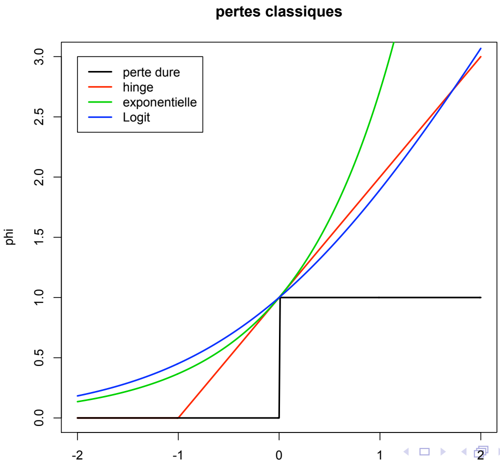
- $\mathcal{F}$ a convex set of functions $f : \mathcal{X} \to \mathbb{R}$
- and $\varphi : \mathbb{R} \to \mathbb{R}^+$ convex and non-decreasing

we define

$$\hat{h}_{\varphi,\mathcal{F}} = \mathrm{sign}(\hat{f}_{\varphi,\mathcal{F}}) \quad \text{with} \quad \hat{f}_{\varphi,\mathcal{F}} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varphi(-y_i f(x_i))$$

Some popular φ

- **Hinge loss :** $\varphi(x) = (1+x)_+$
- **Exponential loss :** $\varphi(x) = e^x$
- **Logit loss :** $\varphi(x) = \log_2(1+e^x)$



Some popular $\mathcal{F}$

- **Linear classifier :** $\mathcal{F} = \{\langle w, . \rangle : \|w\| \leq R\}$ (exercise 1)

- **Convex hull of some basic classifiers** $\{h_1, \ldots, h_M\}$ **:**

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

  with Θ a convex subset of $\mathbb{R}^M$. (later)

- **Ball of a RKHS** $\mathcal{W}$ **:** for $R > 0$

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

  (next week)

Support Vector Machine

SVM

SVM corresponds to

- $\varphi(x) = (1 + x)_+$
- $\mathcal{F} = \{\langle w, . \rangle \ : \ \|w\| \leq R\}$, with $R > 0$.

SVM : Lagrangian version

The classifier $\hat{h}_{\varphi,\mathcal{F}}$ is defined by $\hat{h}_{\varphi,\mathcal{F}}(x) = \text{sign}(\langle \widehat{w}, x \rangle)$ with

$$\widehat{w} = \underset{w \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \langle w, x_i \rangle)_+ + \lambda \|w\|^2 \right\}$$

Reminder on convex optimization

Let $f, -g_1, \ldots, -g_n$ be $\mathcal{C}^1$ convex functions and

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

Karush-Kuhn-Tucker necessary conditions

Set

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

There exists $\hat{\lambda}$ such that

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

Strong duality

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

Geometric interpretation

We have shown that

$$\hat{f}_{\varphi,\mathcal{F}}(x) = \langle \widehat{w}, x \rangle, \text{ with } \widehat{w} = \sum_{i=1}^{n} \hat{\beta}_i x_i$$

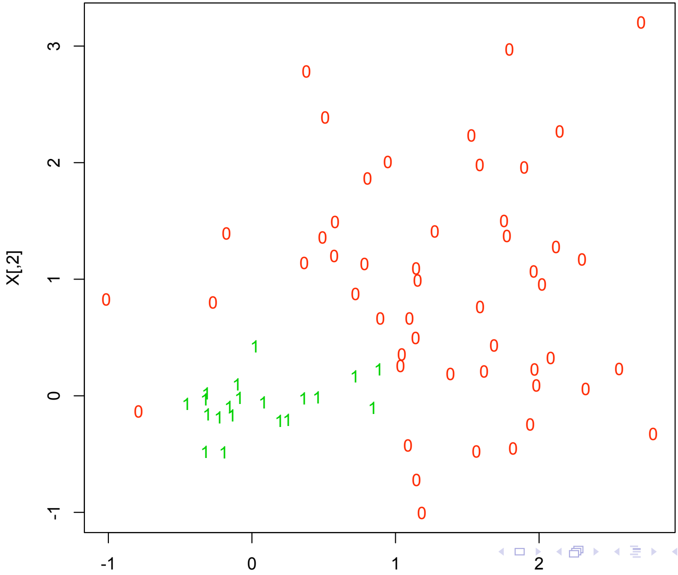
where

KKT

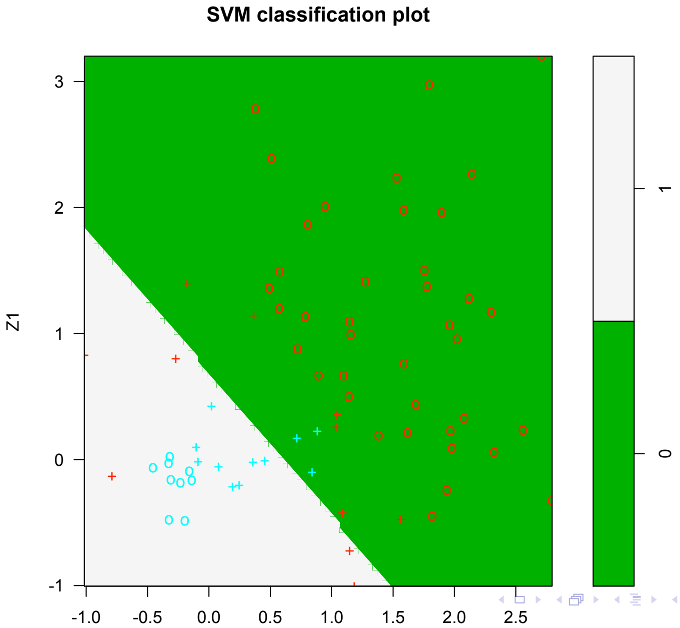
- if $y_i \hat{f}(x_i) > 1$ then $\hat{\beta}_i = 0$,
- if $y_i \hat{f}(x_i) < 1$ then $\hat{\beta}_i = y_i/(2\lambda n)$,
- if $y_i \hat{f}(x_i) = 1$ then $0 \leq \hat{\beta}_i y_i \leq 1/(2\lambda n)$,

Geometric interpretation ?

Data :



les points "+" sont les vecteurs supports



Strong duality

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

Selection of λ

V-fold Cross-Validation

Recipe : split the data into V groups

1. learn $\hat{h}_\lambda$ on $V-1$ "training" groups
2. test $\hat{h}_\lambda$ on the remaining "test" group
3. iterate by permuting the "train" and "test" groups
4. keep $\hat{h}_\lambda$ with the smallest average misclassification error on the V tests.

Example : 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

Elastic Net

Regression setting

Linear model

$$y_i = \langle \beta^*, x_i \rangle + \epsilon_i \quad \text{i=1,...,n}$$

Vectorial writing

$$Y = \mathbf{X}\beta^* + \epsilon$$

Least squares

Least-squares

$$\widehat{\beta}^{LS} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - \mathbf{X}\beta\|^2$$

Issues

1. no unique solution if $p > n$
2. if $\mathrm{cov}(\epsilon) = \sigma^2 I_n$ then the average error

$$\mathbb{E}[\|\widehat{\beta}^{LS} - \beta^*\|^2] = \mathbf{Tr}((\mathbf{X}^T\mathbf{X})^{-1})\sigma^2$$

can be huge. ☹

Sparse regression

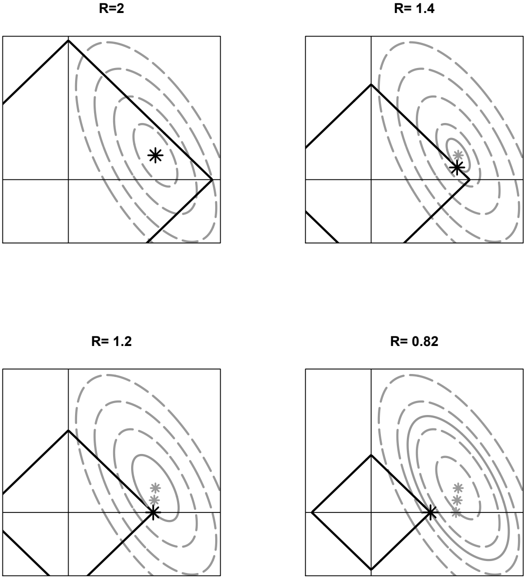
Sparse regression paradigm

Only a few features matters : β^* is sparse or is close to a sparse vector.

Lasso

$$\widehat{\beta}_\mu \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta) \quad \text{with} \quad \mathcal{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \mu|\beta|_{\ell^1}.$$

Singularities of the ℓ^1-ball induce feature selection



Elastic net

Issue

If two (or more) important features are strongly correlated, then only one of the two will be selected. ☹

Elastic Net

Recipe : add a ℓ^2 penalty

$$\widehat{\beta}_{\lambda,\mu} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta) \quad \text{with} \quad \mathcal{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2 + \mu|\beta|_{\ell^1}.$$

Illustration

Intermediate between Ridge and Lasso