



CentraleSupélec

# Statistics and Learning

Julien Bect

([julien.bect@centralesupelec.fr](mailto:julien.bect@centralesupelec.fr))

Teaching : CentraleSupélec / dept. of Statistics and Signal Processing

Research : Laboratory of Signals and Systems (L2S)

Lecture 7/9

Classification : logistic regression.  
Risk, hyper-parameters and model selection.

In this lecture you will learn how to...

- ▶ Classify using logistic regression.
- ▶ Define relevant performance measures for classifiers.
- ▶ Estimate a risk (generalization error).
- ▶ Choose the value of hyper-parameters, select a model.

# Lecture outline

## 1 – Classification : régression logistique

1.1 – Objectifs

1.2 – Modèle linéaire pour la classification

1.3 – Estimation des coefficients  $\beta$

1.4 – Evaluation des performances & choix de  $\delta_0$

## 2 – Risk, hyper-parameters and model selection

2.1 – Estimation of the risk (generalization error)

2.2 – Hyper-parameters, model selection

# Lecture outline

## 1 – Classification : régression logistique

### 1.1 – Objectifs

### 1.2 – Modèle linéaire pour la classification

### 1.3 – Estimation des coefficients $\beta$

### 1.4 – Evaluation des performances & choix de $\delta_0$

## 2 – Risk, hyper-parameters and model selection

### 2.1 – Estimation of the risk (generalization error)

### 2.2 – Hyper-parameters, model selection

# Lecture outline

## 1 – Classification : régression logistique

### 1.1 – Objectifs

1.2 – Modèle linéaire pour la classification

1.3 – Estimation des coefficients  $\beta$

1.4 – Evaluation des performances & choix de  $\delta_0$

## 2 – Risk, hyper-parameters and model selection

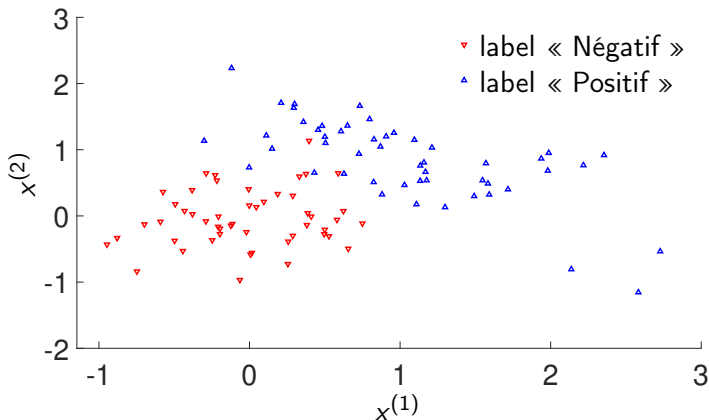
2.1 – Estimation of the risk (generalization error)

2.2 – Hyper-parameters, model selection

## Exemple avec 2 variables explicatives

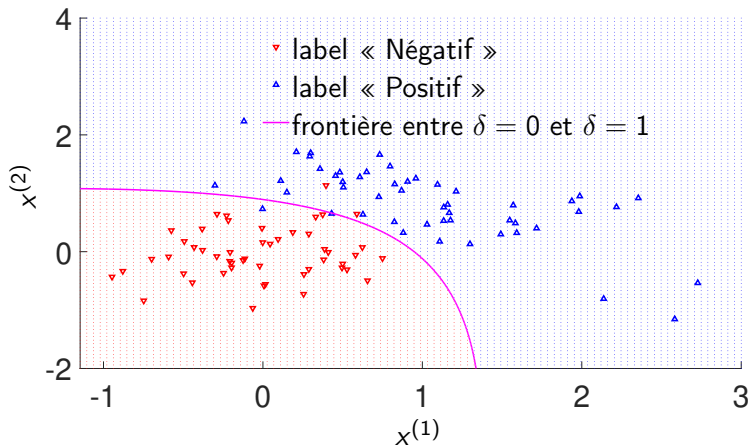
Données utilisées pour l'apprentissage du classifieur :

- ▶  $(X_1, Y_1), \dots, (X_n, Y_n),$
- ▶ ici  $X_i \in \mathbb{R}^2, Y_i \in \{0, 1\}$   
(« 1 » associé au label « Positif », « 0 » au label « Négatif »)



# Résultat : un classifieur

On cherche à obtenir  $h : x \mapsto 0 \text{ or } 1$



## In this lecture you will learn how to...

- ▶ construire des classifieurs à l'aide de la régression logistique
- ▶ définir et évaluer la capacité de généralisation
- ▶ identifier les degrés de liberté
- ▶ comparer des classifieurs

## Cadre mathématique

- ▶  $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$
- ▶  $\mathcal{X} \subset \mathbb{R}^p$
- ▶  $\mathcal{Y} = \{0, 1\}$ 
  - ▮ sauf mention explicite, on s'intéresse à la **classification binaire**
- ▶  $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P^{X, Y}$



# Lecture outline

## 1 – Classification : régression logistique

1.1 – Objectifs

1.2 – Modèle linéaire pour la classification

1.3 – Estimation des coefficients  $\beta$

1.4 – Evaluation des performances & choix de  $\delta_0$

## 2 – Risk, hyper-parameters and model selection

2.1 – Estimation of the risk (generalization error)

2.2 – Hyper-parameters, model selection

# Régression logistique

Malgré le mot « **régression** », il s'agit d'une méthode de **classification** (« régression sur des variables discrètes »)

## Classification binaire

**Rappel.** Si on connaissait  $P^{X,Y}$ , on pourrait calculer, pour une fonction de perte donnée, la fonction de classification optimale.

**Approche suivie.** **modéliser**  $P^{Y|X}$

Ici  $Y|X \sim \text{Bernouilli}$  donc :

**seule**  $P^{Y|X}(Y = 1|X = x)$  **est à modéliser.**

## Modèle

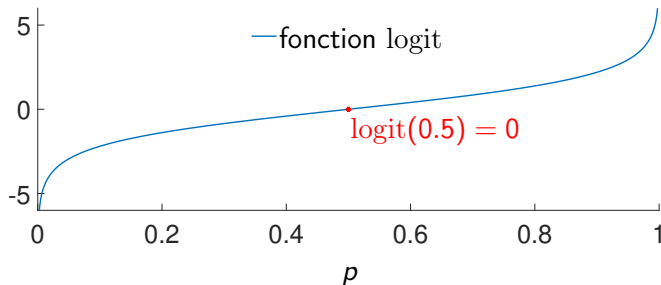
$$P_{\beta}^{Y|X}(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

Remarque : et donc  $P_{\beta}^{Y|X}(Y = 0|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}$

## Fonction logit

$$\begin{aligned}\text{logit} : [0, 1] &\rightarrow \mathbb{R} \\ p &\mapsto \ln\left(\frac{p}{1-p}\right)\end{aligned}$$

On a :  $\text{logit}(P_{\beta}^{Y|X}(Y = 1|X = x)) = \beta_0 + \beta^T x$



⇒ logit réalise une transformation : proba  $p \in [0, 1] \longleftrightarrow \beta_0 + \beta^T x \in \mathbb{R}$

Remarque : équivalence avec une fonction de perte calculée à partir de la divergence de Kullback-Leibler.

## Cas particulier d'un cadre plus large (GLM)

- ▶  $Y|X \sim \text{Bernouilli}(\mathbb{E}_\beta(Y|X))$
  - ▶  $\mathbb{E}_\beta(Y|X)$  s'écrit sous la forme  $g(\mathbb{E}_\beta(Y|X)) = \beta_0 + \beta^\top X$   
(with  $g = \text{logit}$ )
- ⇒ le modèle de RL est un cas particulier du modèle GLM  
( $g$  s'appelle la fonction de lien)

Remarque : tout comme l'est le modèle linéaire gaussien :

- ▶  $Y|X \sim \mathcal{N}(\beta_0 + \beta^\top X, \sigma^2)$
- ▶  $g(\mathbb{E}_\beta(Y|X)) = \beta_0 + \beta^\top X$  with  $g = \text{Id}$

La RL conduit à des prédictions  $P_{\beta}^{Y|X}(Y = 1|X = x) \in [0, 1]$

Pour faire une prédiction dans  $\mathcal{Y} = \{0, 1\}$

Soit  $\delta_0 \in [0, 1]$  (seuil de décision)

On construit la **fonction de décision** :

$$h_{\delta_0} : \mathcal{X} \rightarrow \{0, 1\}$$
$$x \mapsto \begin{cases} 1 & \text{if } P_{\beta}^{Y|X}(Y = 1|X = x) \geq \delta_0 \\ 0 & \text{if } P_{\beta}^{Y|X}(Y = 1|X = x) < \delta_0 \end{cases}$$

$$P_{\beta}^{Y|X}(Y = 1|X = x) = \delta_0 \iff \beta_0 + \beta^T x = \text{logit}^{-1}(\delta_0)$$

⇒ **séparation : hyperplan de  $\mathcal{X}$**

## Proposition

Pour la fonction de perte  $L(y, \tilde{y}) = \mathbb{1}_{\{y \neq \tilde{y}\}}$  :

- ▶ Le risque  $R(h_{\delta_0}) = \mathbb{E}(L(Y, h_{\delta_0}))$  est la **probabilité de mauvais classement**
- ▶ Le **minimum** de  $R(h_{\delta_0})$  par rapport à  $\delta_0$  **est atteint en  $\delta_0 = 0.5$**

# Lecture outline

## 1 – Classification : régression logistique

1.1 – Objectifs

1.2 – Modèle linéaire pour la classification

**1.3 – Estimation des coefficients  $\beta$**

1.4 – Evaluation des performances & choix de  $\delta_0$

## 2 – Risk, hyper-parameters and model selection

2.1 – Estimation of the risk (generalization error)

2.2 – Hyper-parameters, model selection

# Estimateur du maximum de vraisemblance

Allègement des notations :  $x \rightarrow \begin{pmatrix} 1 \\ x \end{pmatrix}$  et  $\beta \rightarrow \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$

$$\Rightarrow P_{\beta}^{Y|X}(Y = 1|X = x) = \frac{\exp(\beta^{\top} x)}{1 + \exp(\beta^{\top} x)}$$

## Vraisemblance

la log-vraisemblance s'écrit :

$$\begin{aligned} \ell(\beta) &= \ln \mathcal{L}(\beta; \underline{x}, \underline{y}) \\ &= \sum_{i=1}^n y_i \beta^{\top} x_i - (1 - y_i) \ln(1 + \exp(\beta^{\top} x_i)) \end{aligned}$$

## Maximisation de $\ell(\beta)$

Elle s'effectue en recherchant  $\beta$  tel que  $\nabla_{\beta} \ell(\beta) = 0$

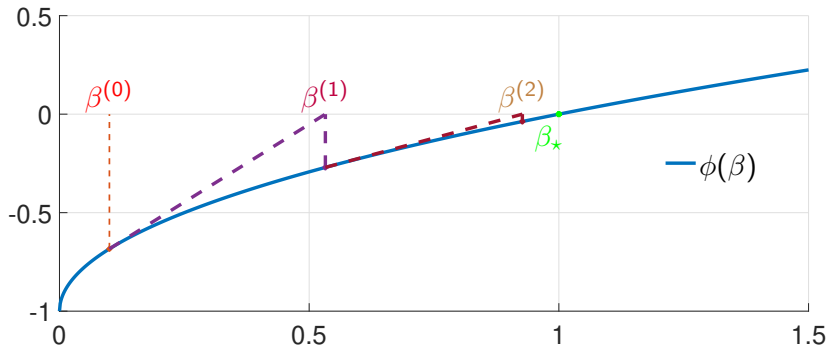
$\Rightarrow$  algorithme de Newton-Raphson

# Algorithme de Newton-Raphson

Soit  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . On cherche  $\beta$  tel que  $\phi(\beta) = 0$

L'algorithme de Newton-Raphson est itératif :

- ▶ initialization :  $\beta^{(0)}$
- ▶ récurrence :  $\beta^{(k+1)} = \beta^{(k)} - \frac{\phi(\beta^{(k)})}{\phi'(\beta^{(k)})}$





## Maximisation de $\ell(\beta)$

Newton-Raphson avec  $\phi : \mathbb{R}^{p+1} \rightarrow \mathbb{R}^{p+1}$  :

Même algorithme avec :

►  $\phi \rightarrow \nabla_{\beta} \ell$

►  $\phi' \rightarrow \nabla_{\beta}^2 \ell$

D'où la récurrence :

$$\beta^{(k+1)} = \beta^{(k)} - \left[ \nabla_{\beta}^2 \ell \left( \beta^{(k)} \right) \right]^{-1} \nabla_{\beta} \ell \left( \beta^{(k)} \right)$$

L'algorithme peut diverger  $\rightarrow$  en pratique on diminue le pas :

$$\beta^{(k+1)} = \beta^{(k)} - \rho_n \left[ \nabla_{\beta}^2 \ell \left( \beta^{(k)} \right) \right]^{-1} \nabla_{\beta} \ell \left( \beta^{(k)} \right)$$

with  $\rho_n > 0$

# Lecture outline

## 1 – Classification : régression logistique

1.1 – Objectifs

1.2 – Modèle linéaire pour la classification

1.3 – Estimation des coefficients  $\beta$

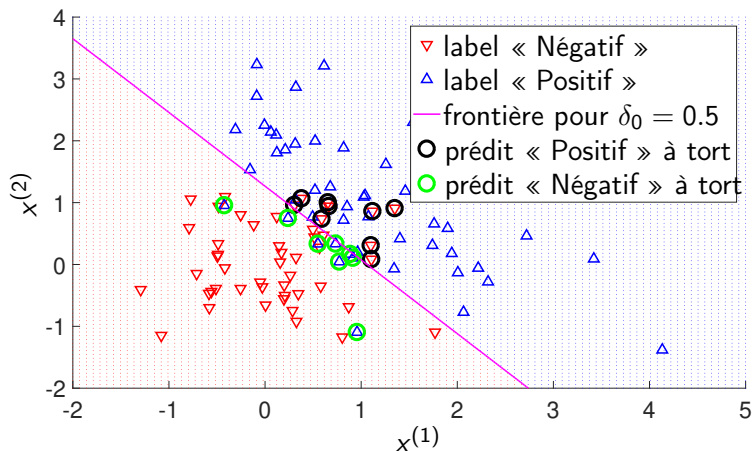
1.4 – Evaluation des performances & choix de  $\delta_0$

## 2 – Risk, hyper-parameters and model selection

2.1 – Estimation of the risk (generalization error)

2.2 – Hyper-parameters, model selection

## RL sur l'exemple avec 2 variables explicatives



### Erreurs de prédiction :

- ▶ prédire « Positif » un objet de classe « Négatif »
- ▶ prédire « Négatif » un objet de classe « Positif »

# Matrice de confusion & grandeurs dérivées

	Réalité Négatif (N)	Réalité Positif (P)
Prédiction Négatif	Vrais Négatifs (VN)	Faux Négatifs (FN)
Prédiction Positif	Faux Positifs (FP)	Vrais Positifs (VP)

## Taux de Vrais Positifs

$$TVP = \frac{VP}{P} = \frac{VP}{VP + FN}$$

Egalement appelé **sensibilité**

## Taux de Vrais Négatifs

$$TVN = \frac{VN}{N} = \frac{VN}{VN + FP}$$

Egalement appelé **spécificité**

# Compromis Taux de Vrais Négatifs / Taux de Vrais Positifs

Si l'étiquette « Positif » correspond à un « défaut » :

- ▶  $1 - TVP$  est le **taux de non-détections**
- ▶  $1 - TVN$  est le **taux de fausses alarmes**

Dans ce cas :

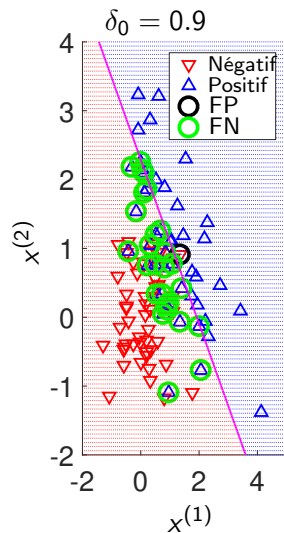
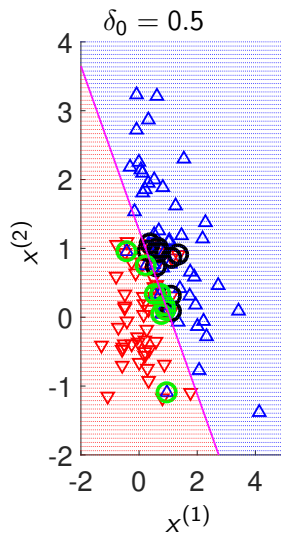
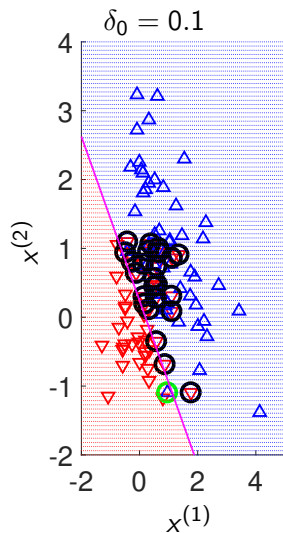
- ▶ le  $TVP$  est généralement la grandeur à privilégier
- ▶ exemple : détection d'une maladie

## Moyen d'action.

Le choix de la valeur de  $\delta_0$  influe sur le compromis  $TVN/TVP$  :

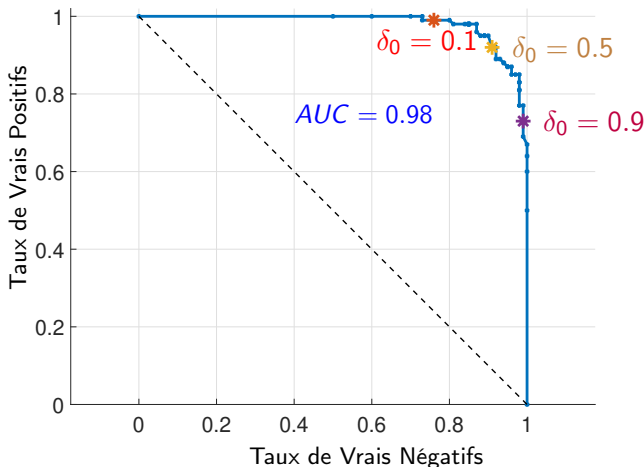
- ▶ rappel :  $h_{\delta_0} = 1$  if  $P_{\beta}^{Y|X}(Y = 1|X = x) \geq \delta_0$
- ▶ quand  $\delta_0 \nearrow$ ,  **$TVN \nearrow$** , and  **$TVP \searrow$**

# Influence de $\delta_0$



# Courbe ROC (Receiver Operating Characteristic)

- ▶ un outil d'**aide à la décision** (choix de  $\delta_0$ )
- ▶ un outil de **comparaison de classifieurs**
- ▶ grandeur dérivée : **AUC** = Area Under Curve



## Gérer le cas où $p$ est grand (grande dimension)

On pénalise directement la log-vraisemblance par une pénalité :

- ▶  $L_1 : \hat{\beta} = \arg \max_{\beta} (\ell(\beta) - \lambda \|\beta\|^2)$
- ▶  $L_2 : \hat{\beta} = \arg \max_{\beta} (\ell(\beta) - \lambda \|\beta\|_1)$

## Gérer le cas multi-classes

Lorsque le nombre de classes est  $\geq 3$ ,

on parle de classification **multi-classes**

Soit  $\{0, 1, \dots, K\}$  l'ensemble des  $K + 1$  labels ( $K + 1$  classes).

On réalise  $K$  RL binaires en considérant une classe (ici « 0 ») comme référence :

$$\begin{cases} \ln \left( \frac{P(Y=\mathbf{1}|X=x)}{P(Y=\mathbf{0}|X=x)} \right) &= \beta_{\mathbf{1},0} + \beta_{\mathbf{1}}^T x \\ \vdots & \\ \ln \left( \frac{P(Y=\mathbf{K}|X=x)}{P(Y=\mathbf{0}|X=x)} \right) &= \beta_{\mathbf{K},0} + \beta_{\mathbf{K}}^T x \end{cases}$$



# Lecture outline

## 1 – Classification : régression logistique

### 1.1 – Objectifs

### 1.2 – Modèle linéaire pour la classification

### 1.3 – Estimation des coefficients $\beta$

### 1.4 – Evaluation des performances & choix de $\delta_0$

## 2 – Risk, hyper-parameters and model selection

### 2.1 – Estimation of the risk (generalization error)

### 2.2 – Hyper-parameters, model selection

# Lecture outline

## 1 – Classification : régression logistique

### 1.1 – Objectifs

### 1.2 – Modèle linéaire pour la classification

### 1.3 – Estimation des coefficients $\beta$

### 1.4 – Evaluation des performances & choix de $\delta_0$

## 2 – Risk, hyper-parameters and model selection

### 2.1 – Estimation of the risk (generalization error)

### 2.2 – Hyper-parameters, model selection

## Problem

Back to the **general setting** (regression/classification).

Let  $\hat{h}$  be a predictor  $\mathcal{X} \rightarrow \mathcal{Y}$  learned from data :

$$\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n)) = \hat{h}(x; \underline{X}, \underline{Y}).$$

Recall that, given a loss function  $L$ , we define the **risk**, or **generalisation error** :

$$\begin{aligned}\mathcal{R}(\hat{h}) &= \mathbb{E} \left( L(Y, \hat{h}(X)) \mid \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) P^{X,Y}(\mathrm{d}x, \mathrm{d}y).\end{aligned}$$

Examples.  $L(y, \tilde{y}) = (y - \tilde{y})^2$ ,  $L(y, \tilde{y}) = |y - \tilde{y}|$ ,  $L(y, \tilde{y}) = \mathbb{1}_{y \neq \tilde{y}}$ , ...

## Problem

How can we **estimate this risk** (which depends on  $P^{X,Y}$ ) ?

## Refresher : empirical risk

We call **empirical risk** the risk

$$\hat{\mathcal{R}}_n = \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) \hat{P}_n(dx, dy) = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{h}(X_i))$$

computed with  $P^{X,Y}$  equal to  $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i, Y_i}$ .

### Question

Is this empirical risk  $\hat{\mathcal{R}}_n$ , in general, a “good” estimator of the true risk  $\mathcal{R}(\hat{h})$ ?



the data is used twice !

**Intuition** : It is « risky » to estimate the risk from the error observed on the same data already used to construct  $\hat{h}$  . . .

## Zoom on an illuminating special case

Consider the case of “ordinary” linear regression :

- ▶  $h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)},$
- ▶ quadratic loss :  $L(y, \tilde{y}) = (y - \tilde{y})^2,$
- ▶  $p + 1 \leq n$  and  $\underline{X}^\top \underline{X}$  an a.s. invertible  $(p + 1) \times (p + 1)$  matrix.

Empirical risk minimization :  $\hat{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}.$

Remark : link between  $\hat{\mathcal{R}}_n$  and the coefficient  $R^2$  of determination :

$$\begin{aligned} R^2 &= 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}^\top X_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{\hat{\mathcal{R}}_n}{\hat{\mathcal{V}}_n(Y)} \quad \text{with } \hat{\mathcal{V}}_n(Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

## Zoom on an illuminating special case (cont'd)

Consider the generalization error wrt responses only :

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n \left( \tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right),$$

with, for all  $i$ ,  $\tilde{Y}_i$  and  $Y_i$  iid conditionnally to  $\underline{X}$ .

### Proposition

Assume that the unknown distribution  $P^{X,Y}$  is such that  $Y_i = \beta^\top X_i + \varepsilon_i$ , with  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ , independent of  $X_i$ . Then

$$\mathbb{E} \left( \tilde{\mathcal{R}}_n \right) = \sigma^2 \left( 1 + \frac{p+1}{n} \right),$$

$$\mathbb{E} \left( \hat{\mathcal{R}}_n \right) = \sigma^2 \left( 1 - \frac{p+1}{n} \right).$$

## Zoom on an illuminating special case (cont'd)

**Interpretation.** On average, the empirical risk under-estimates the generalization error :

$$\mathbb{E} \left( \tilde{\mathcal{R}}_n - \hat{\mathcal{R}}_n \right) = 2 \frac{p+1}{n} \sigma^2 > 0.$$

Another way of looking at this result. Set

$$\eta = \frac{p+1}{n} = \frac{\text{number of coefficients}}{\text{sample size}}.$$

Then

$$\frac{\mathbb{E} \left( \tilde{\mathcal{R}}_n \right)}{\mathbb{E} \left( \hat{\mathcal{R}}_n \right)} = \frac{1 + \eta}{1 - \eta} \xrightarrow[\eta \rightarrow 1]{} +\infty.$$

## Zoom on an illuminating special case (cont'd)

**Proof.** Let us compute first  $\mathbb{E} \left( \tilde{\mathcal{R}}_n \mid \underline{X} \right)$  with (reminder)

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n \left( \tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right).$$

We have  $\mathbb{E} \left( \tilde{Y}_i \mid \underline{X} \right) = \mathbb{E} \left( \hat{\beta}^\top X_i \mid \underline{X} \right) = \beta^\top X_i$ , therefore

$$\begin{aligned} \mathbb{E} \left( \tilde{\mathcal{R}}_n \mid \underline{X} \right) &= \frac{1}{n} \sum_{i=1}^n \mathbb{V} \left( \tilde{Y}_i - \hat{\beta}^\top X_i \mid \underline{X} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left( \underbrace{\mathbb{V} \left( \tilde{Y}_i \mid \underline{X} \right)}_{=\sigma^2} + \underbrace{\mathbb{V} \left( \hat{\beta}^\top X_i \mid \underline{X} \right)}_{=0} \right). \end{aligned}$$



## Zoom on an illuminating special case (cont'd)

We already know that  $\mathbb{V}(\hat{\beta} \mid \underline{X}) = \sigma^2 (\underline{X}^\top \underline{X})^{-1}$ . Therefore :

$$\begin{aligned} \circledast &= \mathbb{V}(\hat{\beta}^\top X_i \mid \underline{X}) \\ &= X_i^\top \mathbb{V}(\hat{\beta} \mid \underline{X}) X_i \\ &= \sigma^2 X_i^\top (\underline{X}^\top \underline{X})^{-1} X_i \\ &= \sigma^2 \text{tr} \left( (\underline{X}^\top \underline{X})^{-1} X_i X_i^\top \right). \end{aligned}$$

By noting that  $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$ , we get :

$$\begin{aligned} \sum_i \mathbb{V}(\hat{\beta}^\top X_i \mid \underline{X}) &= \sigma^2 \text{tr} \left( (\underline{X}^\top \underline{X})^{-1} \sum_i X_i X_i^\top \right) \\ &= \sigma^2 \text{tr}(I_{p+1}) = \sigma^2 (p+1). \end{aligned}$$

## Zoom on an illuminating special case (cont'd)

Thus, we have :

$$\begin{aligned}\mathbb{E}\left(\tilde{\mathcal{R}}_n \mid \underline{X}\right) &= \frac{1}{n} \sum_{i=1}^n \left( \underbrace{\mathbb{V}\left(\tilde{Y}_i \mid \underline{X}\right)}_{=\sigma^2} + \underbrace{\mathbb{V}\left(\hat{\beta}^\top X_i \mid \underline{X}\right)}_{=0} \right) \\ &= \sigma^2 + \sigma^2 \frac{p+1}{n} = \sigma^2 \left(1 + \frac{p+1}{n}\right).\end{aligned}$$

Hence the result :  $\mathbb{E}\left(\tilde{\mathcal{R}}_n\right) = \sigma^2 \left(1 + \frac{p+1}{n}\right)$ .

Prove the second inequality,

$$\mathbb{E}\left(\hat{\mathcal{R}}_n\right) = \sigma^2 \left(1 - \frac{p+1}{n}\right),$$

using Student's theorem (see lecture #6).



# Training set and test set

**Conclusion/extrapolation.** The empirical risk is in general

- ▶ a **downward-biased estimator** of the risk,
- ▶ with a **bias that is increasing when  $p \nearrow$** .

**Solution :** split the data in two sets

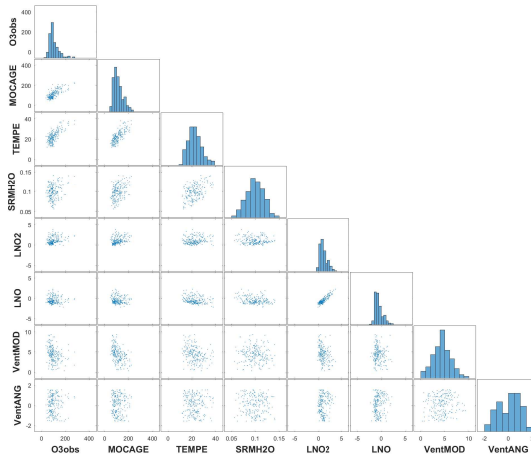
- ▶ **training** data : used to construct  $\hat{h}$ ,
- ▶ **test** data : used to estimate the generalization error.

Example :

**training**  
(e.g., 80%)

**test**  
(20%)

## Exemple “Ozone” (cont'd from lecture #6)



Goal : predict the ozone concentration on day  $t + 1$   
from data available on day  $t$

## “Ozone” example : 70/30

All 7 explanatory variables and their 21 interactions are used.

Result from 10 random splits, 70% / 30% :

$R^2$	$\hat{\mathcal{R}}_n$	$\hat{\mathcal{R}}_n^{\text{test}}$
0.77185	345.1	573.32
0.76831	371.41	496.03
0.77292	343.96	608.62
0.76093	350.53	606.14
0.78584	345.45	669.66
0.75459	399.9	476.61
0.71367	343.72	643.72
0.77689	377.32	524.74
0.8176	317.83	695.86
0.79784	373.18	554.25

# Lecture outline

## 1 – Classification : régression logistique

1.1 – Objectifs

1.2 – Modèle linéaire pour la classification

1.3 – Estimation des coefficients  $\beta$

1.4 – Evaluation des performances & choix de  $\delta_0$

## 2 – Risk, hyper-parameters and model selection

2.1 – Estimation of the risk (generalization error)

2.2 – Hyper-parameters, model selection

## Problem #1 : choosing a « good » family $\mathcal{H}$

**Example.** Selection of  $k$  variables among  $p$ . Let  $J \subset \{1, \dots, p\}$  :

$$h(x) = \beta_0 + \sum_{j \in J} \beta_j x^{(j)}.$$

⇒ Defines a family  $\mathcal{H}_J$  with  $k_J = \text{card}(J) + 1$  parameters.

**Example.** Polynomial (linear!) model in  $x \in \mathbb{R}$ , **degree  $\leq J$**  :

$$h(x) = \beta_0 + \beta_1 x + \dots + \beta_J x^J.$$

⇒ Defines a family  $\mathcal{H}_J$  with  $k_J = J + 1$  parameters.

### Problem : model selection

How to choose the family  $\mathcal{H}_J$  (and, in particular, its « size »  $k_J$ ) ?

Remark : replace  $h(x)$  with  $\ln \frac{h(x)}{1-h(x)}$  for logistic regression.

## Problem #2 : choosing a regularization hyper-parameter

Most methods require some “tuning”...

- ▶ Ridge/LASSO regression :  $\hat{\beta} = \operatorname{argmin} \hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}$ , avec

$$\hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}(\beta) = \hat{\mathcal{R}}_n(\beta) + \lambda \sum_j |\beta_j|^q, \quad q \in \{1, 2\},$$

- ▶ Choosing the number  $k$  of neighbors in a  $k$ -NN model :

$$h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i,$$

with  $\mathcal{V}_{n,k}(x)$  the indices of the  $k$  nearest neighbors of  $x$ .

### Problem : calibration


How to “tune” the value of such hyperparameters ?



# Over-fitting : beware !

## Idea

Choose the family  $\mathcal{H}_J$ , or the hyperparameter  $\lambda$ , in order to **minimize (an estimation of) the generalization error**.

 again, the empirical risk  $\hat{\mathcal{R}}_n$ , estimated on the training data, is not appropriate !

**Example.** Polynomial regression in  $x \in \mathbb{R}$ , **degree  $\leq J$**  :

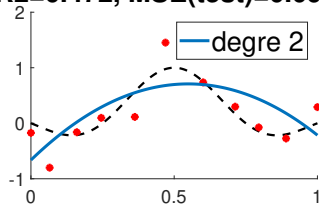
$$h(x) = \beta_0 + \beta_1 x + \dots + \beta_J x^J,$$

with  $J = 2, 5, 8, 11$ .

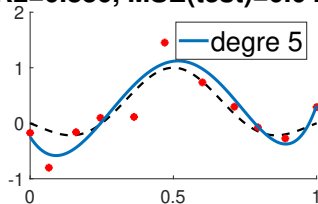
Recall that, in linear regression, the empirical risk has a downward bias proportional to the number of parameters in the model.

## Example : polynomial regression

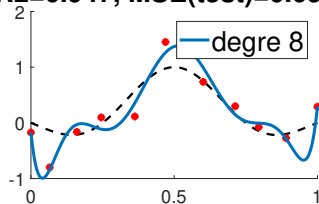
**$R^2=0.472$ ,  $MSE(test)=0.0983$**



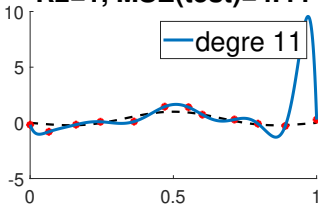
**$R^2=0.836$ ,  $MSE(test)=0.0425$**



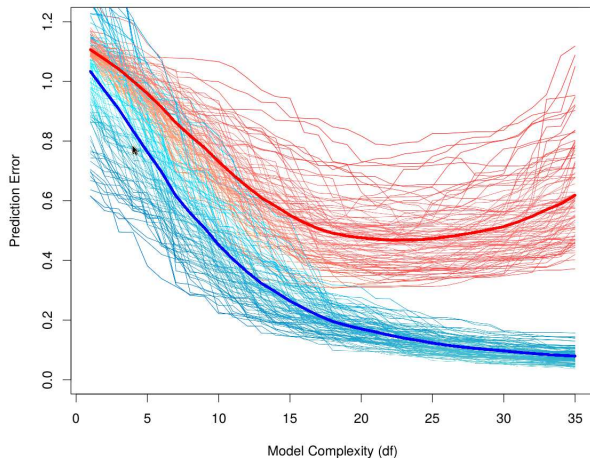
**$R^2=0.947$ ,  $MSE(test)=0.0974$**



**$R^2=1$ ,  $MSE(test)=4.44$**



# Understanding over-fitting : simulations



Blue : empirical risk  $\hat{\mathcal{R}}_n$  / Red : error on the test set

Figure from Hastie, Tibshirani & Friedman (2017).  
*The Elements of Statistical Learning* (12th edition), Springer.

## Let's recapitulate...

**Problem.** We want to estimate the error to choose  $\mathcal{H}$  or  $\lambda$  but...

- ▶ it should be done neither on the **training data**  
( $\Rightarrow$  **over-fitting** problem),
- ▶ nor on the **test data**  
( $\Rightarrow$  **bias** in the final estimation of the generalization error).



# Solution : validation set

Idea : split the data in three sets

- ▶ **training** data : construct  $\hat{h}$  with given  $\mathcal{H}/\lambda$ ,
- ▶ **validation** set : choose  $\mathcal{H}$ ,  $\lambda$ , etc.
- ▶ **test** data : estimate the generalization error.

Simple validation (hold-out)

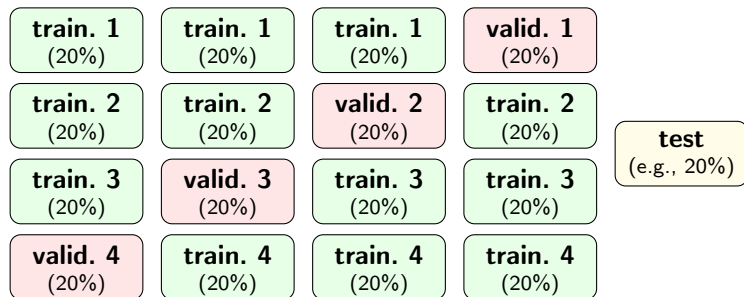
**training**  
(e.g., 60%)

**validation**  
(e.g., 20%)

**test**  
(e.g., 20%)

# Better validation : the cross validation method

**$k$ -fold cross-validation**, here with  $k = 4$  :



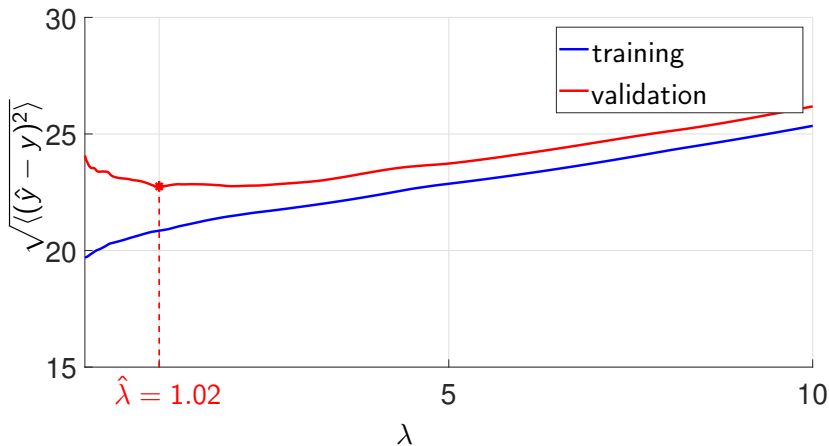
➡ the error is averaged over the  $k$  validation sets.

Special case : **leave-one-out** cross validation

▶  $k = n$  blocks (of size  $n/k = 1$ ).

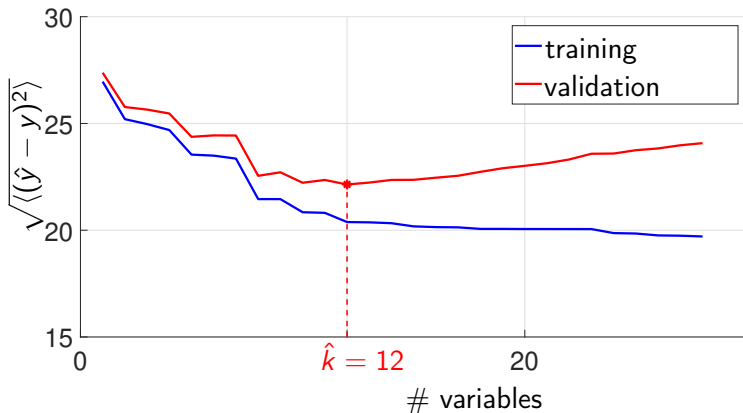
## “Ozone” example : LASSO / choice of $\lambda$

- ▶ Predictor : LASSO regression using all variables and their interactions
- ▶  $\hat{\lambda}$  obtained by CV (LOO)



## “Ozone” example : variable selection

- ▶ Predictor obtained by the ordinary least squares method, on an increasing number of variables  
(linear terms first, then interactions)
- ▶ Validation error : LOO cross validation





## Final remark : another approach to model selection

Assumption : parametric statistical models  $\mathcal{M}_j$  for  $P^{Y|X}$ .

Denote by  $\hat{\theta}_j^{\text{MLE}}$  the MLE of  $\theta$  in model  $\mathcal{M}_j$ .

Then the AIC criterion can also be used for model selection :

$$\hat{j} = \operatorname{argmin} \operatorname{AIC}(j), \quad \operatorname{AIC}(j) = -2 \ln \mathcal{L} \left( \hat{\theta}_j^{\text{MLE}}; \underline{X}, \underline{Y} \right) + 2k_j,$$

with  $k_j$  the number of parameters in model  $\mathcal{M}_j$ .

## “Ozone” exemple : AIC

- Predictor obtained by the ordinary least squares method, on an increasing number of variables  
(linear terms first, then interactions)

