Apprentissage numérique (AN)
Séance 1

François-Xavier Dupé, Valentin Emiya
prénom.nom@lif.univ-mrs.fr
http://www.lif.univ-mrs.fr/~vemiya/teaching/AN/

M2 IF/GSI/BDA - Aix-Marseille Université
2011-2012

Module AN : objectifs

- Pratique et théorie en apprentissage numérique
  - BDA : projets applicatifs
  - GSI : projets théorie/application
  - IF : étude d'articles
- Cohésion / 3 filières & compétences différentes
  - grande richesse : favoriser les échanges
  - (in)compréhension : travailler la communication

Module AN : organisation

- 5 séances du 8/11 au 6/12 = 1 intro + 4 suivi
- BDA :
  - projets « reconnaissance faciale (visages) »
  - 2 groupes de 5 personnes
- GSI :
  - Projets techniques d'apprentissage
  - 3 groupes de 2-3 personnes
- M2 IF : étude 1 articles / personne

Module AN : suivi et évaluation

Suivi en séance :

- BDA : 3 présentations type prestataire-client
- GSI et IF : 1 présentation/séance avancement + suivi personnalisé

Evaluation : en plus du suivi en séance,

- tous : 1 rapport à rendre ~ 10 janvier
- IF : 1 soutenance en janvier

Plan de la séance 1

- Présentation du module et des sujets
- Apprentissage supervisé :
  - Principe, formalisation, concepts théoriques de base
  - Aperçu de quelques techniques : kNN, SVM
  - Validation d'un apprentissage
  - Principe ERM, compromis biais-variance
- Thématique de l'année : « sélection de modèle »

Un exemple introductif

- Problème : Quelle est le chiffre *a* qui prolonge la séquence :
  - *1235 ... a*

Un exemple introductif

- Quelques solutions valides :
  - a=6. Argument : c'est la suite des entiers sauf 4.
  - a=7. Argument 1 : c'est la suite des nombres premiers.
  - a=7. Argument 2 : suite binaire 1(1), 10(2), 11(3), 101(5), 111(7), 1011(11), 1111(15), 10111(23), 11111(31)...
  - a=8. Argument : c'est la suite de Fibonacci.
  - a=2π. Argument : la liste ordonnée des racines du polynôme : $x^5 - (11+a)x^4 + (41+11a)x^3 - (61-41a)x^2 + (30+61a)x - 30a$ qui est le développement de $(x-1)(x-2)(x-3)(x-5)(x-a)$ $\Longrightarrow$ a peut être n'importe quel nombre réel supérieur ou égal à 5)
- Généralisation : il est facile de montrer que n'importe quel nombre est la suite d'une séquence de nombre...

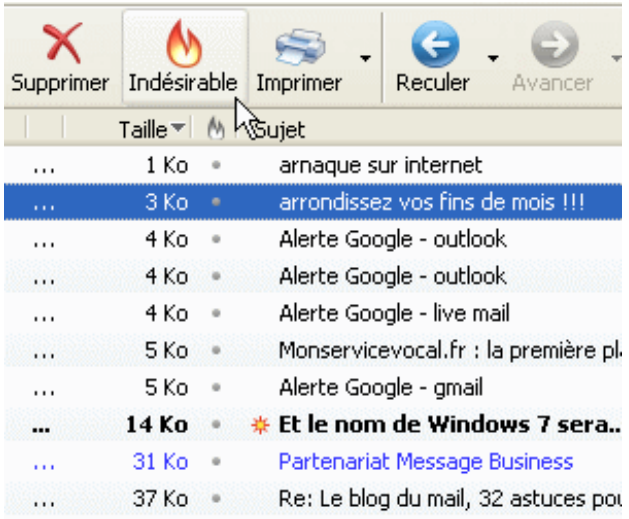
  → Quel modèle choisir ?

De vrais exemples (1)

Classification supervisée

But : écarter automatiquement les annonces publicitaires et autres messages non sollicités.

Données : des messages (x_i) dont on sait s'ils sont des SPAMs ou non (y_i binaire).

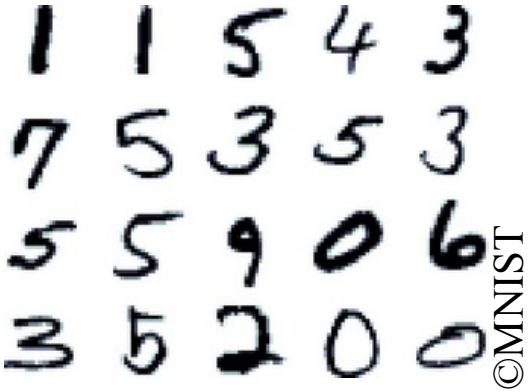


Objectif : construire un ***classifieur***, capable d'attribuer une de ces deux classes à un nouveau document.

But : reconnaissance de chiffres manuscrits.

Données : des chiffres écrits sur une rétine de 16x16 pixels, associés à une classe parmi {0, 1, ... , 9}

Objectif : attribuer la bonne classe (problème de ***reconnaissance des formes***, *pattern recognition*).



De vrais exemples (2)

Régression supervisée

But : Prédire la température, la pression atmosphérique, le taux d'ozone ou la vitesse du vent.

Données : Numériques (ex. : capteurs de température) ou symboliques (temps de la veille).

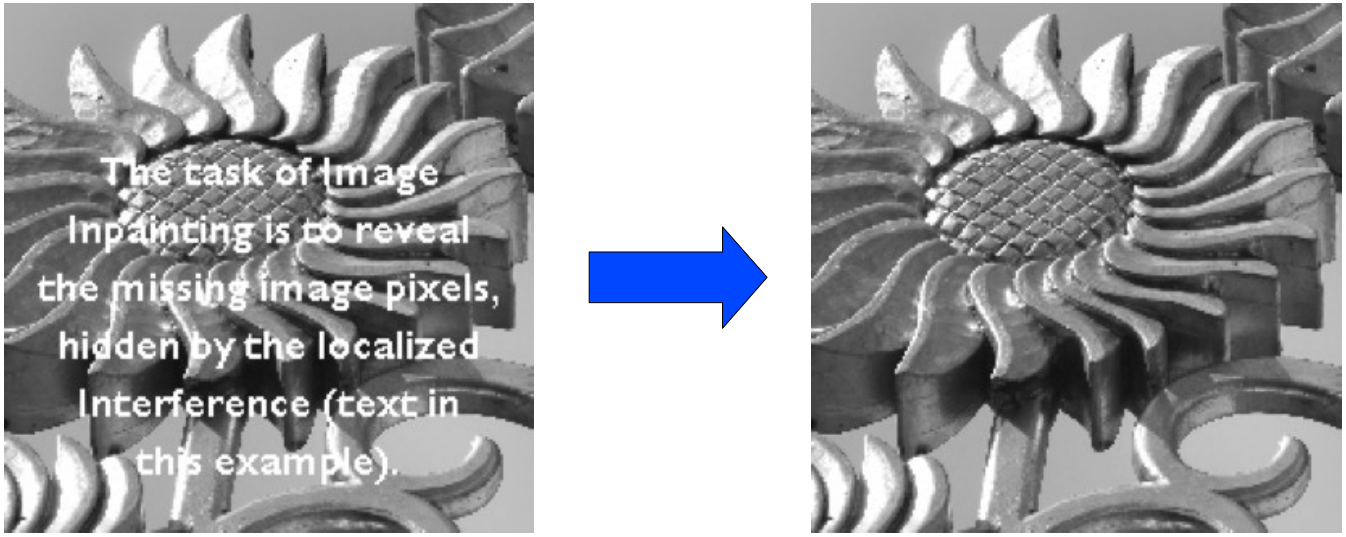
But : Dans le problème de détection des SPAMs, associer à un nouveau document la ***probabilité*** que ce soit un SPAM.

But : Prédire le pronostic vital d'un patient à partir de différents paramètres cliniques.

De vrais exemples (3)

Traitement d'image : l'*inpainting*

But : reconstruire une partie manquante dans une image.



Apprentissage : position x_i et intensité y_i des pixels connus $i \in I$.

Objectif : estimer l'intensité y_i à partir de la position x_i des pixels inconnus $i \in \bar{I}$.

De vrais exemples (4)

Reconnaissance de visages

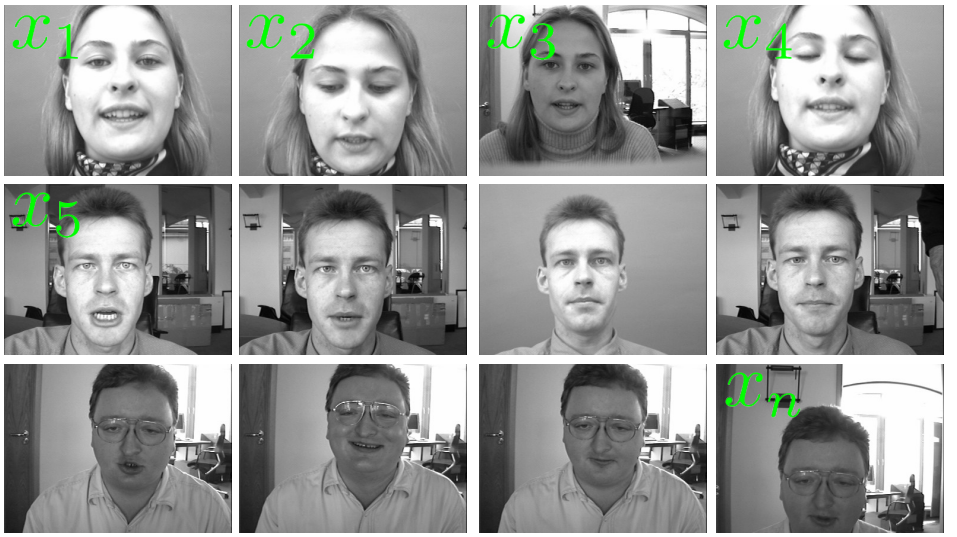
But : reconnaître une personne dans une image.

Apprentissage : images x_i et personnes y_i, $1 \leq i \leq n$, dans une base de *n* images annotées :

$y_1 = \ldots = y_4 = p_1$

$y_5 = \ldots = y_8 = p_2$

$y_9 = \ldots = y_{12} = p_3$



©BioID Face Database

Objectif : identifier une personne à partir d'une nouvelle image *x*.

$x =$  $y?$

Références Bibliographiques

- Apprentissage Artificiel, *Antoine Cornuéjols et Laurent Miclet.*
- The elements of statistical Learning, *Hastie, Tibshirani et Friedman.*
- Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations, *Witten et Frank*, auteurs de ***Weka*** *http://www.cs.waikato.ac.nz/ml/weka*.

Présentation des sujets : IF

- B. Efron, T. Hastie, I. Johnstone, R. Tibshirani,
  Least Angle Regression,
  The Annals of Statistics, Inst. of Math. Stat., No 32, 2004.

- P. Stoica, Y. Selen,
  Model-order selection: a review of information criterion rules,
  IEEE Signal Processing Magazine, No 21, 2004.

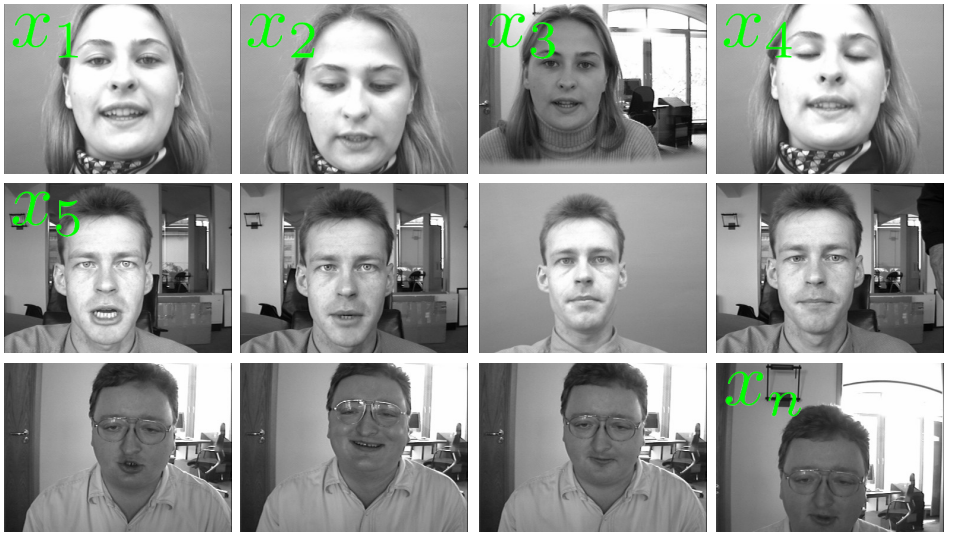
- K. A. Clarke,
  The Effect of Priors on Approximate Bayes Factors from MCMC Output,
  Tech. Rep., 2000.

- A. Gálvez, A. Iglesias,
  Efficient particle swarm optimization approach for data fitting with free knot B-splines
  Computer-Aided Design, No 43, 2011.

Présentation des sujets : BDA

Reconnaissance faciale

Scénario industriel...



©BioID Face Database

Présentations :

- 15/11 : compréhension du sujet, état de l'art
- 22/11 : présentation de la mise en œuvre choisie
- 29/11 : récapitulatif, verrous, endroits délicats

Présentation des sujets : GSI

- Validation croisée pour la sélection de modèles :
  - Théorie
  - Implémentations rapides
- Tests d'hypothèses et classification :
  - Théorie
  - Utilisation en classification : choix des modèles
- Utilisation des kNN dans le cas de grandes bases de données
  - Etudes des différentes techniques et des cas d'utilisation

Plan

- Présentation du module et des sujets
- Apprentissage supervisé :
  - Principe, formalisation, concepts théoriques de base
  - Aperçu de quelques techniques : kNN, SVM
  - Validation d'un apprentissage
  - Principe ERM, compromis biais-variance
- Thématique de l'année : « sélection de modèle »

2 types d'apprentissage

- **Apprentissage supervisé :** les exemples d'apprentissage x possèdent l'information à apprendre y. On cherche une loi de dépendance sous-jacente $y=f(x)$.
  - Si f est une *fonction continue*
    - Régression
    - Estimation de densité
  - Si f est une *fonction discrète*
    - Classification
  - Si f est une *fonction binaire* (booléenne)
    - Apprentissage de concept

2 types d'apprentissage

- **Apprentissage supervisé :** les exemples d'apprentissage x sont disponibles avec l'information à apprendre y. On cherche une loi de dépendance sous-jacente $y=f(x)$.

- **Apprentissage non-supervisé :** les exemples d'apprentissage x sont disponibles sans information supplémentaire, « apprentissage sans professeur ». On cherche des régularités ou structures sous-jacentes :
  - **Clustering : découvrir les catégories et les règles de catégorisation**
  - **Estimation de la densité de probabilité p(X)**
  - **Séparation aveugle de sources**

Formaliser : quoi ? comment ?

- Comment modéliser les données ?

  → *distribution statistique de (x,y)*

- Que veut-on apprendre ?

  → *une fonction f telle que y=f(x)*

- Comment évaluer la performance de l'apprentissage obtenu ?

  → *notion de risque R(f) (=erreur)*

Formalisation : les données

Notations :

- **entrées** : vecteurs de n **attributs**
  $$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{X} = \mathbf{X}_1 \times \ldots \times \mathbf{X}_n$$
- **sorties** $y \in \mathbf{Y}$

Exemple

Entrée à $n = 2$ attributs :

âge $\mathbf{X}_1 = [0; 120]$, fumeur $\mathbf{X}_2 = \{\text{oui}, \text{non}\}$

Sortie : $\mathbf{Y} = \{\text{patient\_a\_risque}, \text{patient\_sans\_risque}\}$.

Dépendance entrée/sortie : le risque cardiaque est-il lié à l'âge et au fait de fumer ?

Formalisation : les données

- Notations :
  - **entrées** : vecteurs de n **attributs**
    $$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{X} = \mathbf{X}_1 \times \ldots \times \mathbf{X}_n$$
  - **sorties** $y \in \mathbf{Y}$
- **Modèle statistique des données** :
  variables aléatoires $(X, Y) \in \mathbf{X} \times \mathbf{Y}$
  distribuées selon $P(X, Y)$ (inconnu en g[al])
- Base d'apprentissage, « **échantillon** » :
  ensemble de l **exemples** $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$
  avec $(\mathbf{x}_i, y_i)$ **i.i.d.** selon $P(X, Y)$
- Test, application : autres tirages i.i.d. selon $P(X, Y)$
  $$\{(\mathbf{x}_{l+1}, y_{l+1}), (\mathbf{x}_{l+2}, y_{l+2}), \ldots\}$$

Principe de l'apprentissage supervisé

- **À partir des données**

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$$

trouver une fonction $h : \mathbf{X} \rightarrow \mathbf{Y}$

qui prédit y à partir de $\mathbf{x}$.

- **Défi : généraliser** pour tout $\mathbf{x} \in \mathbf{X}$

[Si $\mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$, facile : apprentissage par coeur.]

h est aussi appelée aussi hypothèse, règle, classifieur, etc.

Principe de l'apprentissage supervisé

- **À partir des données**

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$$

trouver une fonction $h : \mathbf{X} \rightarrow \mathbf{Y}$

qui prédit y à partir de $\mathbf{x}$.

Exemple

$$h_1(\mathbf{x}) = \begin{cases} \text{a\_risque} & \text{si } x_2 = \text{fumeur et } x_1 > 60 \\ \text{sans\_risque} & \text{sinon} \end{cases}$$

Performance : notion de risque

- **Fonction de perte** / *loss function* :

$$\text{Classification: } L(y, h(\mathbf{x})) = \begin{cases} 1 & \text{si } y \neq h(\mathbf{x}) \\ 0 & \text{sinon.} \end{cases}$$

$$\text{Régression: } L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$$

- **Fonction Risque** ou erreur = espérance mathématique de la fonction de perte :

$$R(h) = \int L(y, h(\mathbf{x}))dP(\mathbf{x}, y)$$

Problème général de l'apprentissage supervisé :
à partir de l'échantillon $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$
trouver $h : \mathbf{X} \rightarrow \mathbf{Y}$ qui minimise $R(h)$

Synthèse

Problème général de l'apprentissage supervisé :

- à partir de l'échantillon $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$ supposé i.i.d.
- trouver $h : \mathbf{X} \to \mathbf{Y}$, $h \in H$ (classe de fonctions)
- qui minimise $R(h)$

Quelques règles de classification

- Règle majoritaire :

$$\forall \mathbf{x} \in \mathbf{X}, f_{\text{maj}}(\mathbf{x}) = \text{argmax}_y P(y) = y_{\text{maj}}$$

- Règle du maximum de vraisemblance *(maximum likelihood)* :

$$\forall \mathbf{x} \in \mathbf{X}, f_{\text{mv}}(\mathbf{x}) = \text{argmax}_y P(\mathbf{x}|y)$$

- Règle de Bayes

$$\forall \mathbf{x} \in \mathbf{X}, f_{\text{B}}(\mathbf{x}) = \text{argmax}_y P(y|\mathbf{x})$$

La classification

- Nous cherchons une fonction $f : \mathbf{X} \rightarrow \mathbf{Y}$ qui prend des valeurs discrètes.
- Fonction de perte : *perte 0-1*

$$L\left(y, f(x)\right) = \delta\left(y \neq f(x)\right)$$

- Risque ou erreur de la fonction f : *erreur quadratique* :

$$R(h) = \int L(y, h(\mathbf{x}))dP(\mathbf{x}, y) = P(Y \neq h(X))$$

- Théorème : la fonction de classification de risque minimal est la règle de Bayes :

$$\forall \mathbf{x} \in \mathbf{X}, f_{\mathrm{B}}(\mathbf{x}) = \mathrm{argmax}_y P(y|\mathbf{x})$$

La régression

- Nous cherchons une fonction $f : \mathbf{X} \rightarrow \mathbf{Y}$ qui prend des valeurs continues.
- Fonction de perte : *écart quadratique :*

$$L\left(y, f(x)\right) = \left(y - f(x)\right)^2$$

- Risque ou erreur de la fonction f : *erreur quadratique* :

$$R\left(f\right) = \int_{\mathbf{X}\times\mathbf{Y}} \left(y - f(x)\right)^2 dP(x, y)$$

- Théorème : la fonction de régression de risque minimal est l'espérance des valeurs observables en x :

$$f^*\left(x\right) = \int_{\mathbf{Y}} y dP(y|x)$$

Plan

- Présentation du module et des sujets
- Apprentissage supervisé :
  - Principe, formalisation, concepts théoriques de base
  - Aperçu de quelques techniques : kNN, SVM
  - Validation d'un apprentissage
  - Principe ERM, compromis biais-variance
- Thématique de l'année : « sélection de modèle »

Plan

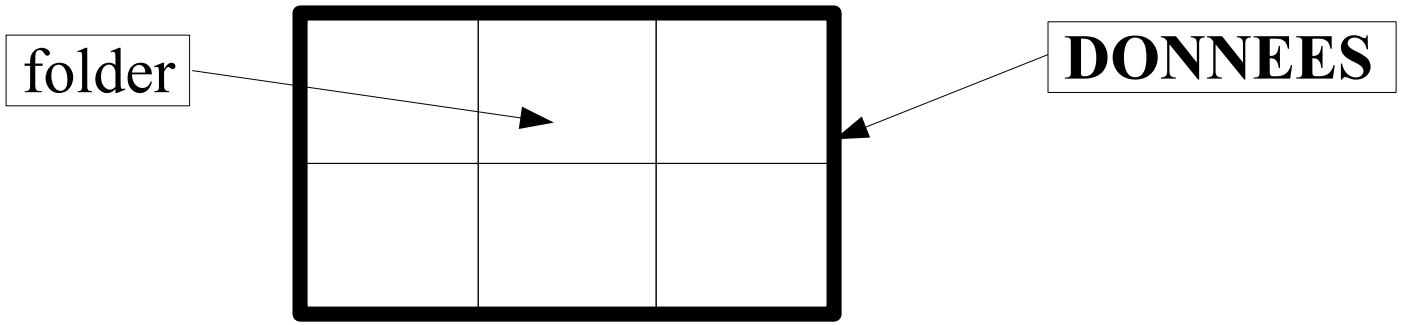
- Présentation du module et des sujets
- Apprentissage supervisé :
  - Principe, formalisation, concepts théoriques de base
  - Aperçu de quelques techniques : kNN, SVM
  - Validation d'un apprentissage
  - Principe ERM, compromis biais-variance
- Thématique de l'année : « sélection de modèle »

Validation empirique d'un apprentissage

- Plusieurs méthodes permettent de valider (ou d'infirmer) la valeur d'un processus d'apprentissage.
- Une des approches consiste à n'utiliser qu'une partie des données pour apprendre et à se servir des autres données pour tester le résultat.
- Différentes mesures permettent alors de comparer des processus (erreur, F-score, etc.)

Validation croisée *(cross-validation)*

- La validation croisée est une généralisation de la méthode précédente.
- Elle consiste à diviser les données en K *folders*, à en enlever un pour l'apprentissage puis à l'utiliser pour la phase de test. Le processus est ensuite réitéré.
- L'erreur moyenne tend alors vers l'erreur en généralisation.



Plan

- Présentation du module et des sujets
- Apprentissage supervisé :
  - Principe, formalisation, concepts théoriques de base
  - Aperçu de quelques techniques : kNN, SVM
  - Validation d'un apprentissage
  - Principe ERM, compromis biais-variance
- Thématique de l'année : « sélection de modèle »

L'apprentissage en pratique : intuition

Intuition 1 : on veut utiliser une méthode telle que

- les arbres de décision,
- les réseaux de neurones,
- les kNN, les SVM, etc.

= choix d'une classe H d'hypothèses $h \in H$

Intuition 2 : pour trouver le « meilleur » $h \in H$,

- on ne peut pas calculer le risque R(h)
- on ne dispose que d'un échantillon S.

= parmi les $h \in H$, on prend celui qui décrit le mieux S

L'apprentissage en pratique : intuition

L'apprentissage consisterait donc à faire

$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

où R_s une mesure de performance sur les exemples connus S : $\hat{h}$ est optimal sur S.

*R_S s'appelle le **risque empirique***

Le principe d'optimisation ci-dessus s'appelle

***minimisation du risque empirique*.**

Risque empirique

- En classification, $R_{emp}(h)$ (ou $R_S(h)$) est la moyenne du nombre d'erreurs sur l'échantillon S :

  $R_{emp}(h) = ||\{i : h(x_i) \neq y_i\}|| \,/\, ||S||$

- En régression, $R_{emp}(h)$ est la moyenne des carrés des écarts à la moyenne de h sur S :

  $R_{emp}(h) = 1/_{||S||} \sum_{x_i \in S} (y_i - h(x_i))^2$

Risque (réel) vs. risque empirique

- Risque (réel)

  = **espérance** de la fonction de perte

$$R(h) = E[L\,(Y, h(X))] = \int L\,(y, h(x))\,dP(x, y) \;\; \text{(classif./régr.)}$$

- Risque empirique :

  = **moyenne sur S** de la fonction de perte

$$R_{\text{emp}}(h) = \frac{1}{l} \sum_{(x,y)\in S} L\,(y, h(x)) \;\; \text{(classif./régr)}$$

- Donc $R_{emp}(h)$ est une estimation de R(h)

Synthèse (provisoire)

- Intuitivement, le principe de minimisation du risque empirique (ERM) recommande de

rechercher une fonction h de H minimisant $R_{emp}(h)$

- En classification = minimiser le nombre d'erreurs de h sur l'échantillon.
- En régression = méthode des moindres carrés.

ERM : deux questions importantes

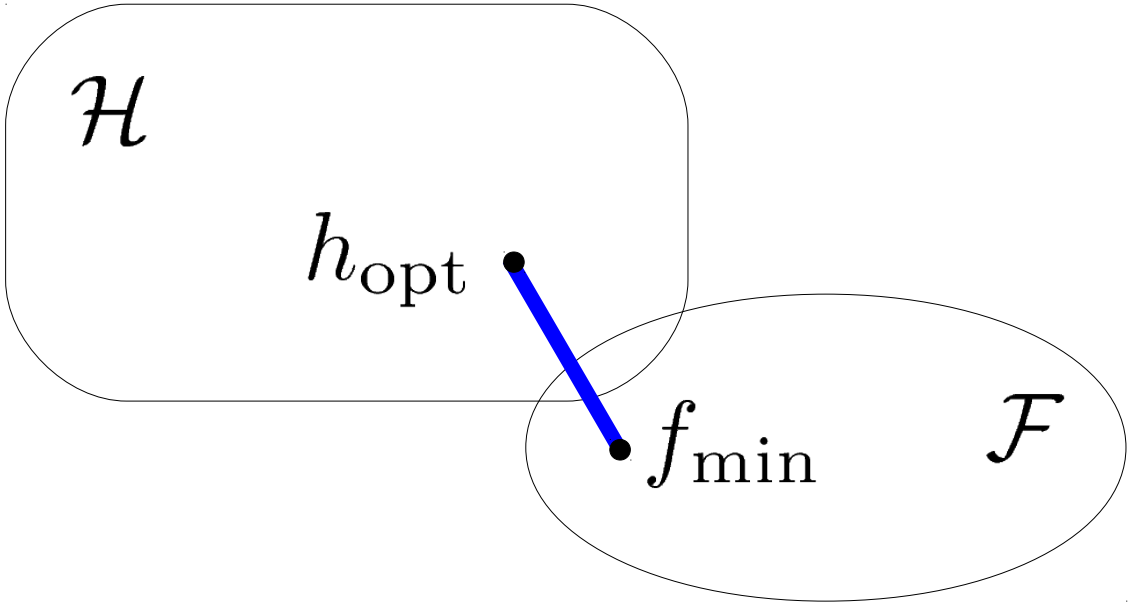
$$\text{Idéalement : } f_{\min} \triangleq \arg\min_{f} R(f)$$

$$\text{Principe ERM : } \widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

Quelles sont les enjeux et conséquences relatifs

- au choix de *H* ?
- au « remplacement » du risque réel *R(h)* par le risque empirique $R_S(h)$?

Conséquence du choix de H : le biais



$$f_{\text{min}} \triangleq \arg\min_{f} R(f)$$

$$h_{\text{opt}} \triangleq \arg\min_{h \in \mathcal{H}} R(h)$$

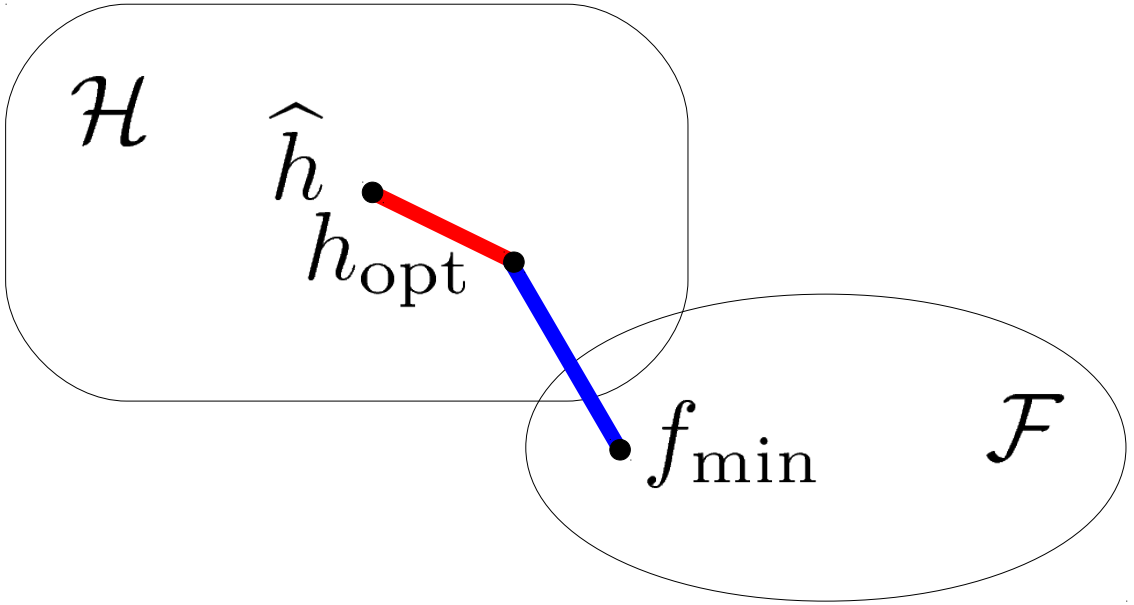
$\mathcal{H}$: espace des hypothèses possibles de l'apprenant.

$\mathcal{F}$: espace des fonctions cible de la nature.

En général, $f_{\text{min}} \notin \mathcal{H}$, et $R(h_{\text{opt}}) \geq R(f_{\text{min}})$.

$R(h_{\text{opt}}) - R(f_{\text{min}})$ est l'**erreur d'approximation** ou biais.

Minimisation de R_S : la variance



$$f_{\min} \triangleq \arg\min_{f} R(f)$$

$$h_{\text{opt}} \triangleq \arg\min_{h \in \mathcal{H}} R(h)$$

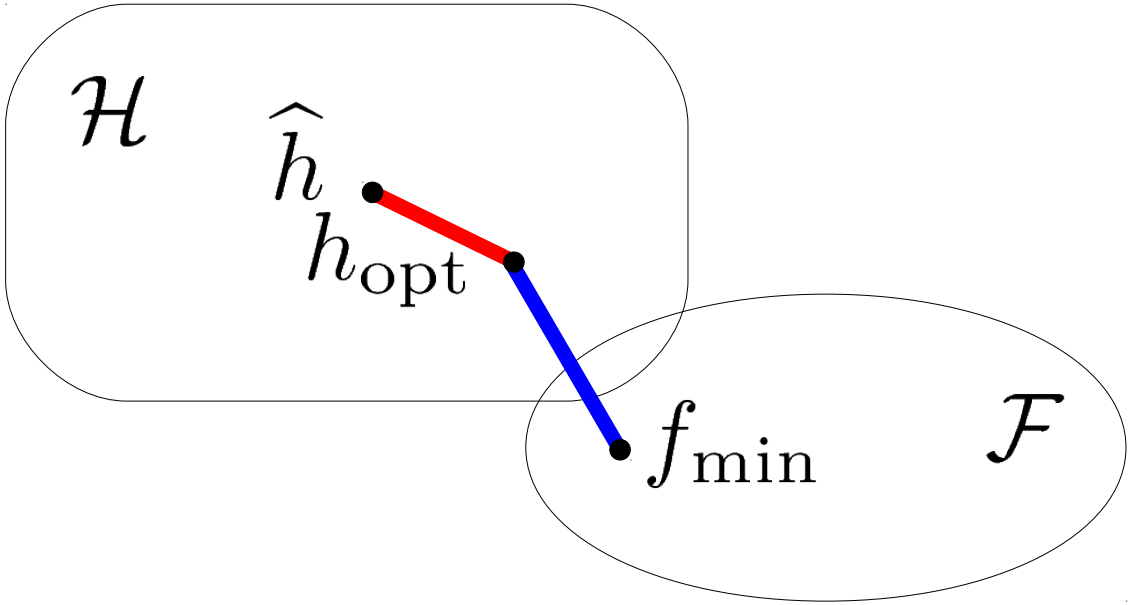
$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

$R(h) \neq R_S(h)$ donc on peut avoir $\widehat{h} \neq h_{\text{opt}}$ et $R\left(\widehat{h}\right) \geq R(h_{\text{opt}})$

$R\left(\widehat{h}\right) - R(h_{\text{opt}})$ est l'**erreur d'estimation** ou variance.

En particulier, problème du surapprentissage.

Décomposition de la performance



$$f_{\min} \triangleq \arg\min_{f} R(f)$$

$$h_{\text{opt}} \triangleq \arg\min_{h \in \mathcal{H}} R(h)$$

$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

$$R\left(\widehat{h}\right) = R(f_{\min}) + R(h_{\text{opt}}) - R(f_{\min}) + R\left(\widehat{h}\right) - R(h_{\text{opt}})$$

$R(f_{\min})$ est l'**erreur minimale, intrinsèque**.

$R(h_{\text{opt}}) - R(f_{\min})$ est l'**erreur d'approximation**.

$R\left(\widehat{h}\right) - R(h_{\text{opt}})$ est l'**erreur d'estimation**.

ERM : deux questions importantes

$$\text{Idéallement : } f_{\min} \triangleq \arg\min_{f} R(f)$$

$$\text{Principe ERM : } \widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

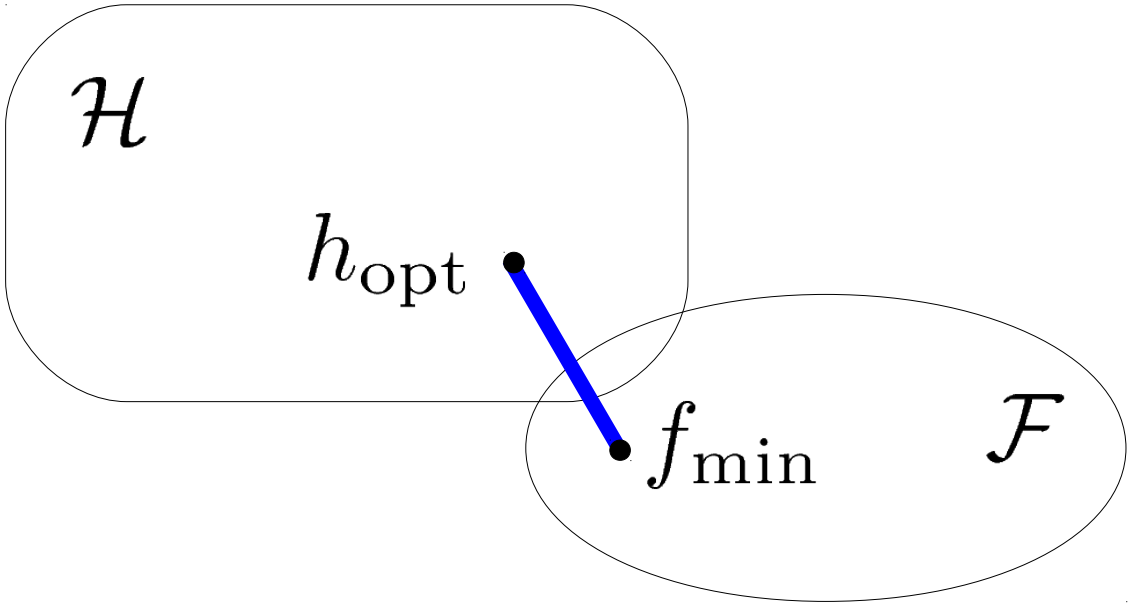
Enjeux relatifs

- au choix de *H* : erreur d'approximation
- au « remplacement » du risque réel *R(h)* par le risque empirique $R_S(h)$: erreur d'estimation

$$R\left(\widehat{h}\right) = R_{\min} + R_{\text{approx.}} + R_{\text{estim.}}$$

Comment minimiser ces erreurs ?

Réduire l'erreur d'approximation/biais



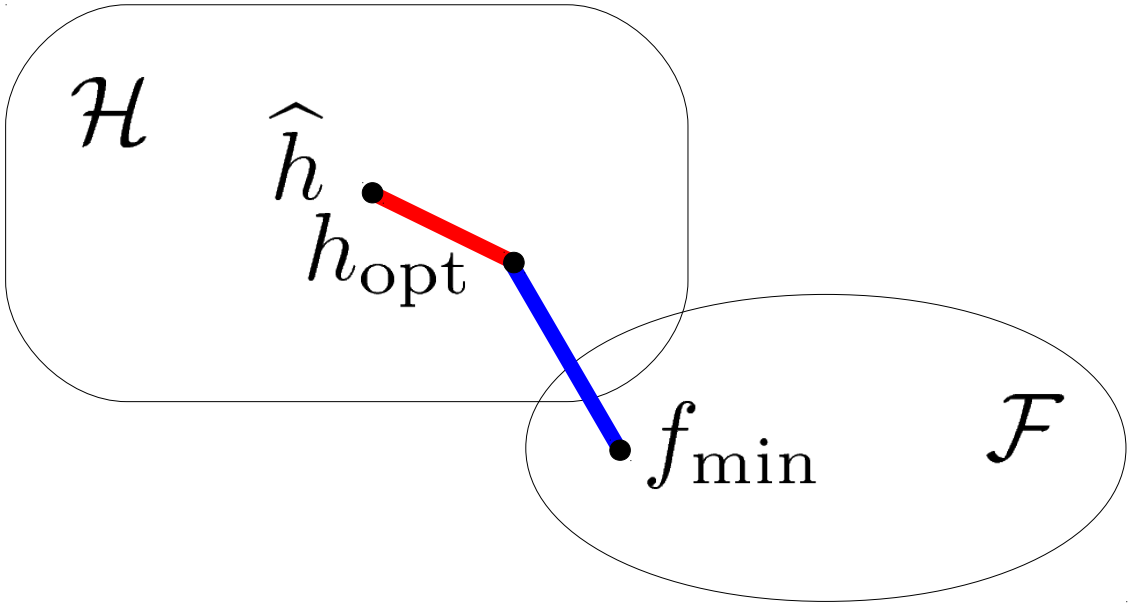
$$f_{\min} \triangleq \arg\min_{f} R(f)$$

$$h_{\text{opt}} \triangleq \arg\min_{h \in \mathcal{H}} R(h)$$

Solutions :

- Choisir pour *H* une famille de fonctions adaptée au problème
- Augmenter la « richesse » de H : ordre, nombre de degrés de liberté, etc.

Réduire l'erreur d'estimation/variance



$$f_{\min} \triangleq \arg\min_{f} R(f)$$

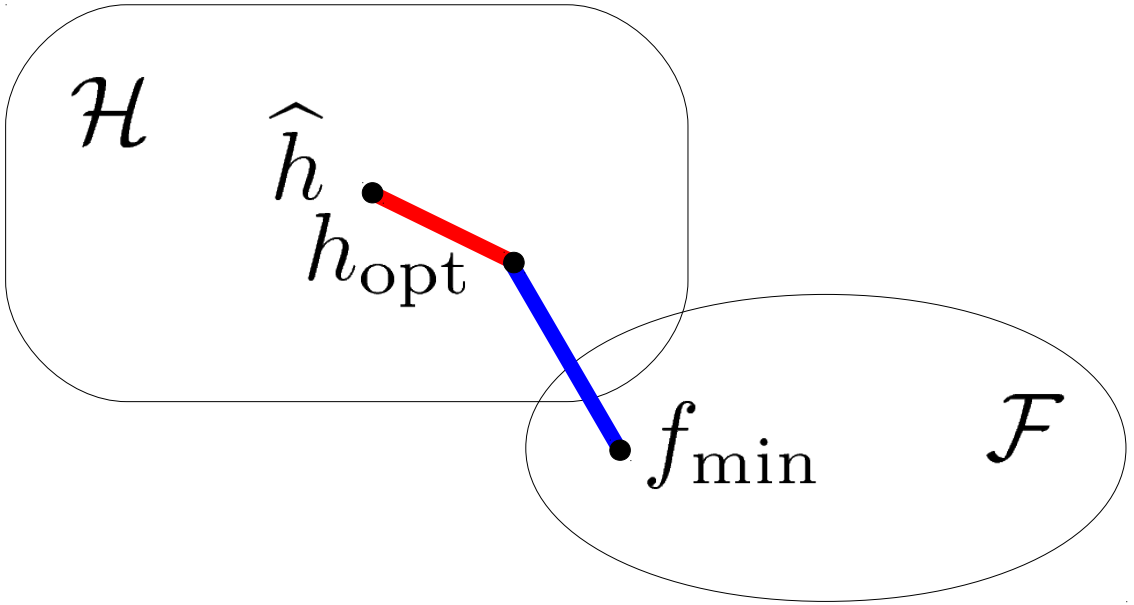
$$h_{\text{opt}} \triangleq \arg\min_{h \in \mathcal{H}} R(h)$$

$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

Solutions :

- Augmenter la taille de S (coûteux) : le principe ERM est *consistant* dans la classe H si $\hat{h}$ tend vers h_{opt} quand le nombre d'exemples tend vers l'infini.
-

Réduire l'erreur d'estimation/variance



$$f_{\min} \triangleq \arg\min_{f} R(f)$$

$$h_{\text{opt}} \triangleq \arg\min_{h \in \mathcal{H}} R(h)$$

$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

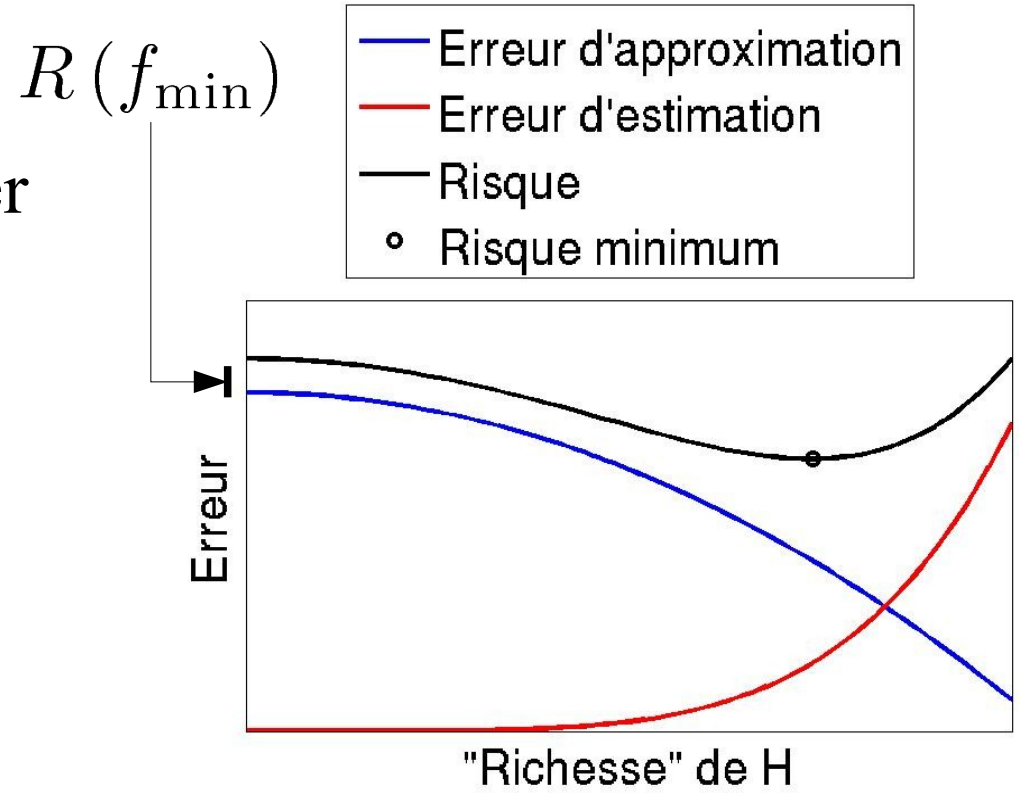
Solutions :

- Augmenter la taille de S (coûteux) : le principe ERM est *consistant* dans la classe H si $\hat{h}$ tend vers h_{opt} quand le nombre d'exemples tend vers l'infini.
- Réduire la « richesse » de H : ordre, nombre de degrés de liberté, etc.

Le compromis biais-variance

$$R\left(\widehat{h}\right) = R\left(f_{\min}\right) + R\left(h_{\text{opt}}\right) - R\left(f_{\min}\right) + R\left(\widehat{h}\right) - R\left(h_{\text{opt}}\right)$$

= compromis entre erreur d'approximation et erreur d'estimation pour minimiser l'erreur totale



Exemples :

- Ordre d'un polynôme
- Nombre de gaussiennes
- Etc.

Plan

- Présentation du module et des sujets
- Apprentissage supervisé :
  - Principe, formalisation, concepts théoriques de base
  - Aperçu de quelques techniques : kNN, SVM
  - Validation d'un apprentissage
  - Principe ERM, compromis biais-variance
- Thématique de l'année : « sélection de modèle »

Sélection de modèle : 2 questions

A partir d'un échantillon d'apprentissage S,

- qualitativement, quel modèle H choisir ?
- quantitativement, pour une complexité croissante

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_p$$

quel ensemble choisir pour obtenir un bon compromis biais-variance ?

Solutions : plusieurs méthodes de sélection de modèle sont possibles.

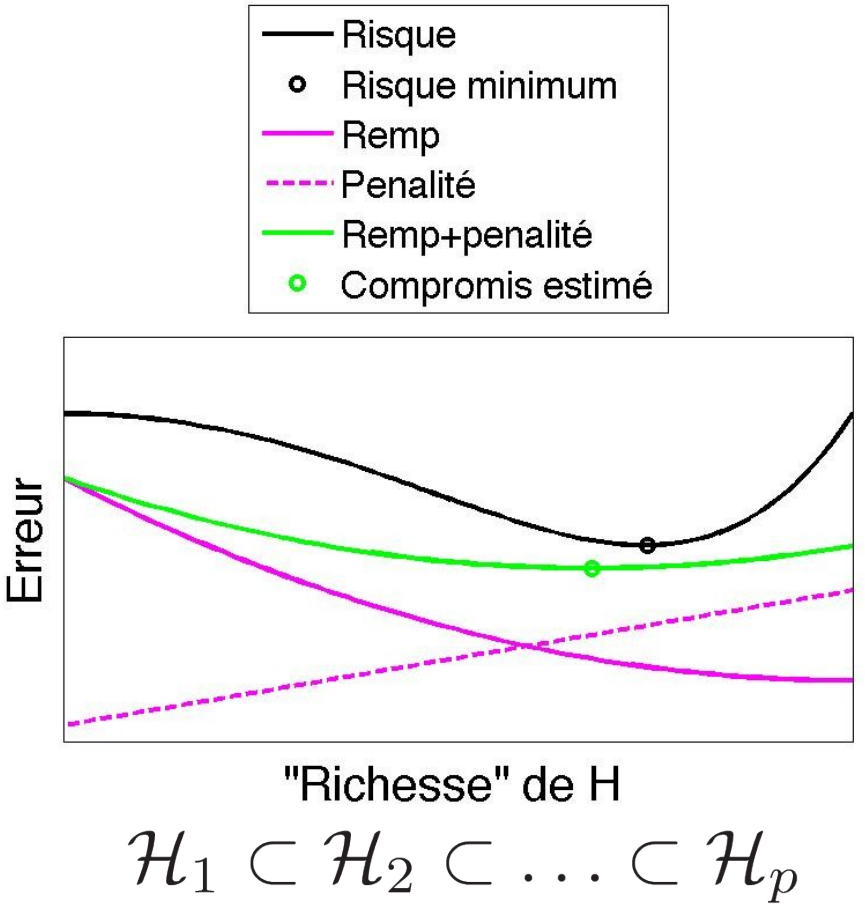
Solution 1 : pénaliser le risque empirique

En utilisant une pénalité q dépendant de la complexité du modèle, on remplace

$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}_m} R_S(h)$$

par

$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}_m, 1 \leq m \leq p} R_S(h) + \mathrm{q}(m)$$

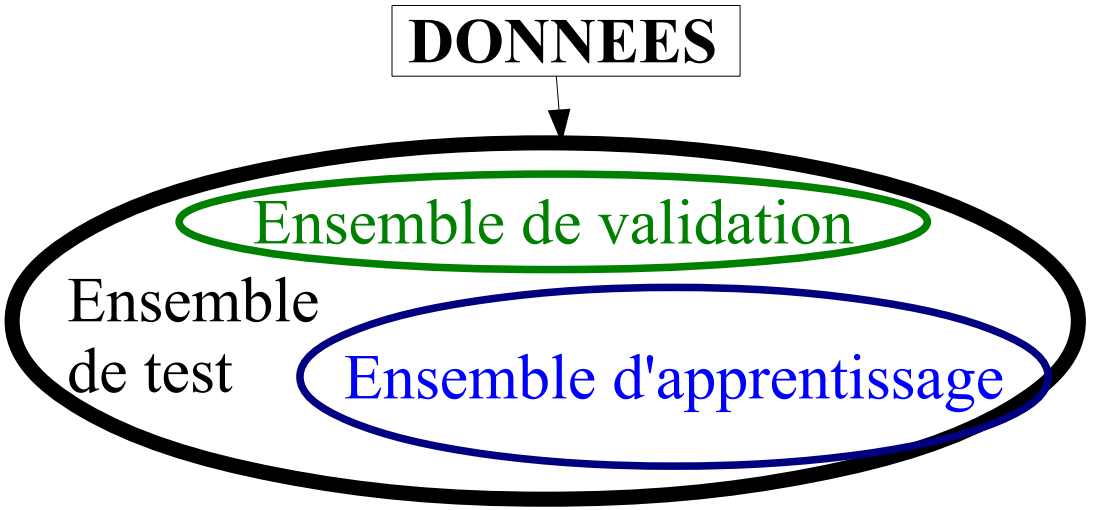


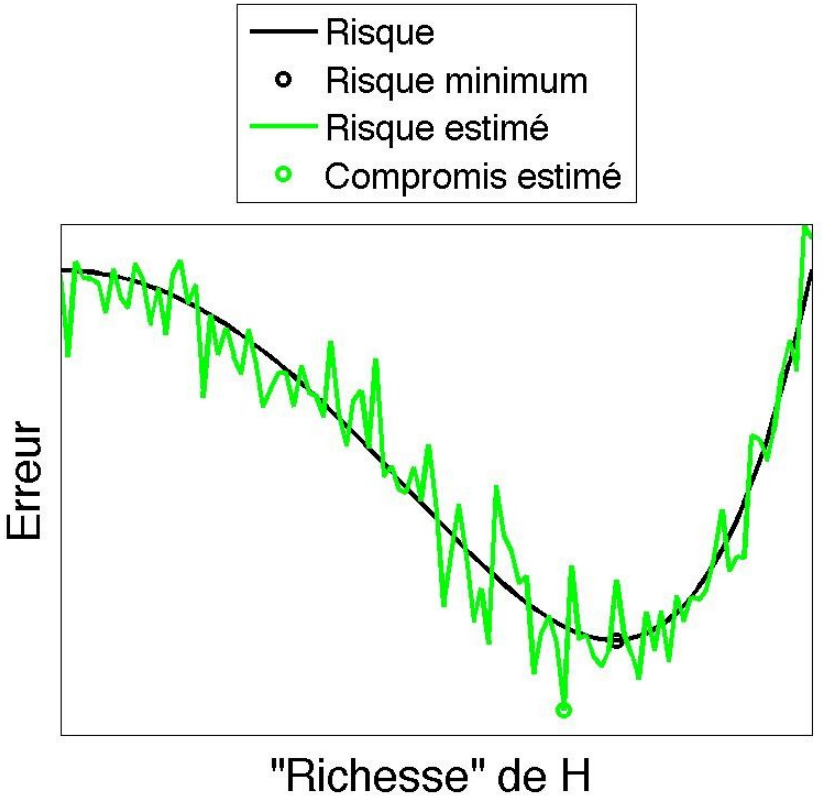
$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_p$$

Solution 2 : test d'hypothèses

Idée générale : $\widehat{\mathcal{H}} \triangleq \arg\min_{\mathcal{H}} p\left(\mathcal{H}\,|S\right)$

Solution 3 : estimer le risque réel à partir d'un échantillon de validation



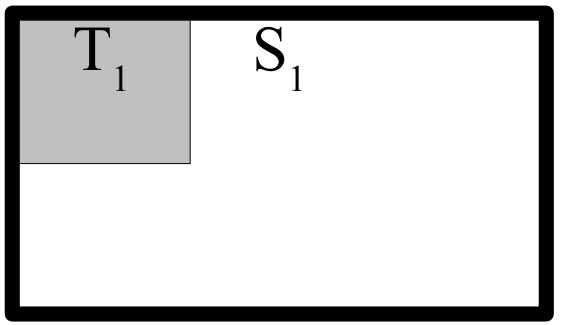


$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_p$$

En pratique, on utilise plutôt une validation croisée.

Validation croisée

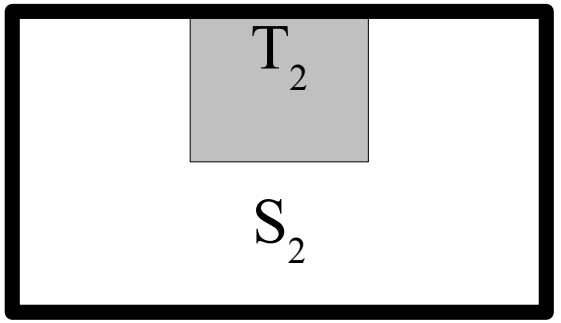
- Itération 1 :



$$\widehat{h}_1 \triangleq \arg\min_{h \in \mathcal{H}} R_{S_1}(h)$$

$$\rightarrow R_{T_1}\left(\widehat{h}_1\right)$$

- Itération 2 :



$$\widehat{h}_2 \triangleq \arg\min_{h \in \mathcal{H}} R_{S_2}(h)$$

$$\rightarrow R_{T_2}\left(\widehat{h}_2\right)$$

...

$\frac{1}{K}\sum_{k=1}^{K} R_{T_k}\left(\widehat{h}_k\right)$ est une bonne estimation de $R\left(\widehat{h}\right)$

Validation croisée pour trouver un compromis biais-variance

- On dispose d'un échantillon S
- Objectif : choisir entre $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_M$
- Pour chaque $\mathcal{H}_m$, on fait une validation croisée :

$$R^{(m)} = \frac{1}{K} \sum_{k=1}^{K} R_{T_k}\left(\widehat{h}_k^{(m)}\right) \qquad \left(\widehat{h}_k^{(m)} \triangleq \arg\min_{h \in \mathcal{H}_m} R_{S_k}(h)\right)$$

- On sélectionne finalement $\mathcal{H}_{\widehat{m}}$ tel que

$$\widehat{m} \triangleq \arg\min_{m} R^{(m)}$$

Plan

- Présentation du module et des sujets
- Apprentissage supervisé :
  - Principe, formalisation, concepts théoriques de base
  - Aperçu de quelques techniques : kNN, SVM
  - Validation d'un apprentissage
  - Principe ERM, compromis biais-variance
- Thématique de l'année : « sélection de modèle »