Initiation à la Fouille de Données et à l'Apprentissage

Deuxième séance

M2 I2A
2011-2012

Valentin Emiya

Plan du cours

- Rappels de probabilités
- Modélisation de l'apprentissage supervisé (suite)
  - La régression
  - L'estimation de densité
- Principe ERM
- Validation d'un apprentissage

Plan du cours

- Rappels de probabilités
- Modélisation de l'apprentissage supervisé (suite)
  - La régression
  - L'estimation de densité
- Principe ERM
- Validation d'un apprentissage

Probabilités discrètes (1)

- Soit Ω un ensemble fini ou dénombrable qu'on appelle univers.
- Une probabilité sur Ω est une fonction p: $\Omega \rightarrow [0,1]$ telle que $\sum_{w \in \Omega} p(w) = 1$.
- Une partie A de Ω est un événement.

$$p(A) = \sum_{w \in A} p(w)$$

- Si des $(A_i)_{i \in I}$ sont incompatibles 2 à 2 :

$$p(\bigcup_{i \in I} A_i) = \sum_{i \in I} p(A_i)$$

Probabilités discrètes (2)

- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
- $p(\overline{A})=1-p(A)$ où $\overline{A} = \Omega \backslash A$
- $A \subseteq B$ implique $p(A) <= p(B)$
- Soit A un événement tel que $p(A) \neq 0$. Alors la fonction $p(.|A) : B \rightarrow p(B|A) = p(A \cap B)/_{p(A)}$ est une probabilité sur Ω.
- Loi de Bayes : $p(A|B) = p(B|A)*p(A)/_{p(B)}$

Probabilités discrètes (3)

- Si des $(A_i)_{i\in I}$ sont incompatibles 2 à 2 et vérifient $\bigcup_{i\in I} A_i=\Omega$ alors

  $p(B) = \sum_{i\in I} p(B,A_i) = \sum_{i\in I} p(B|A_i)\, p(A_i)$

- Des événements A et B sont indépendants ssi ils vérifient l'une des conditions suivantes :
  - $p(A|B) = p(A)$
  - $p(B|A) = p(B)$
  - $p(A\cap B) = p(A)p(B)$

Variables aléatoires (1)

- Une v.a. réelle est une application $X: \Omega \rightarrow R$
- La loi de probabilité de X est définie
  - par $P(X \in I) = P(X^{-1}(I))$ pour tout intervalle I de R
  - si $X(\Omega)$ est fini ou dénombrable, par les $P(X=x)$ pour $x \in X(\Omega)$. Notation abrégée : $P(x)$.

  Exemple : X=sommes des chiffres de deux dés
  $\Omega=\{1,...,6\}x\{1,...,6\}$, $X(\Omega)=\{1,...,12\}$, P(x)= ?
- X v.a. continue s'il existe une densité p t.q.

$$P(X \in I) = \int_I p(x)dx$$

Variables aléatoires (2)

- X et Y v.a. indépendantes si pour tous intervalles I, J,

$$P((X,Y) \in I \times J) = P(X \in I)\, P(Y \in J)$$

- Propr : si X et Y v.a. indépendantes continues,

$$p(x,y) = p(x)\, p(y)$$

- Loi de Bayes

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$$

Espérance

- Espérance E[X] de X :
  - si X fini ou dénombrable $E[X] = \sum_{x \in X(\Omega)} xP(X = x)$
  - si X continue $E[X] = \int xp(x)dx$

- Plus généralement,

$$E[g(X)] = \sum_{x \in X(\Omega)} g(x)P(X = x)$$

$$E[g(X)] = \int g(x)p(x)dx$$

Variance, moyenne

- Variance Var[X] de X :

$$Var[X] = E\left[(X - E[X])^2\right] = \begin{cases} \sum_{x \in X(\Omega)} (x - E[X])^2 P(x) \\ \int (x - E[X])^2 p(x) dx \end{cases}$$

- L'espérance E[X] est souvent estimée par la moyenne de n v.a. indépendantes identiquement distribuées (i.i.d) selon P(X), car

$$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = E[X] \text{ et } Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{Var[X]}{n}$$

Plan du cours

- Rappels de probabilités
- Modélisation de l'apprentissage supervisé (suite)
  - Correction de l'exercice 4 (TD1)
  - La régression
  - L'estimation de densité
- Principe ERM
- Validation d'un apprentissage

Plan du cours

- Rappels de probabilités
- Modélisation de l'apprentissage supervisé (suite)
  - La régression
  - L'estimation de densité
- Principe ERM
- Validation d'un apprentissage

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* :

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

- Th. : La fonction de régression de risque minimal est l'espérance des valeurs observable en x :

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

Estimation de densité

- On dispose de réalisations indépendantes $x_1, \ldots, x_n$ de X. On cherche à estimer P(x) pour tout x.
- On cherche une fonction $f : \mathbf{X} \rightarrow [0; 1]$ qui approche P (ou sa densité p dans le cas continu) au mieux.
- Fonction de perte : $L(x, y) = -\log(y)$
- Fonction de risque pour X discret et continu :

$$R(F) = \sum_{x \in \mathbf{X}} -\log(F(x))\, P(x) \text{ et } R(f) = \int_{\mathbf{X}} -\log(f(x))\, p(x) dx$$

- Th. : R(F) est minimal pour F=P (cas discret).
  R(f) est minimal pour f=p (cas continu).

Plan du cours

- Rappels de probabilités
- Modélisation de l'apprentissage supervisé (suite)
  - La régression
  - L'estimation de densité
- Principe ERM
- Validation d'un apprentissage

L'apprentissage en pratique : intuition

Intuition 1 : on veut utiliser une méthode telle que

- les arbres de décision,
- les réseaux de neurones,
- les fonctions linéaires, 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.

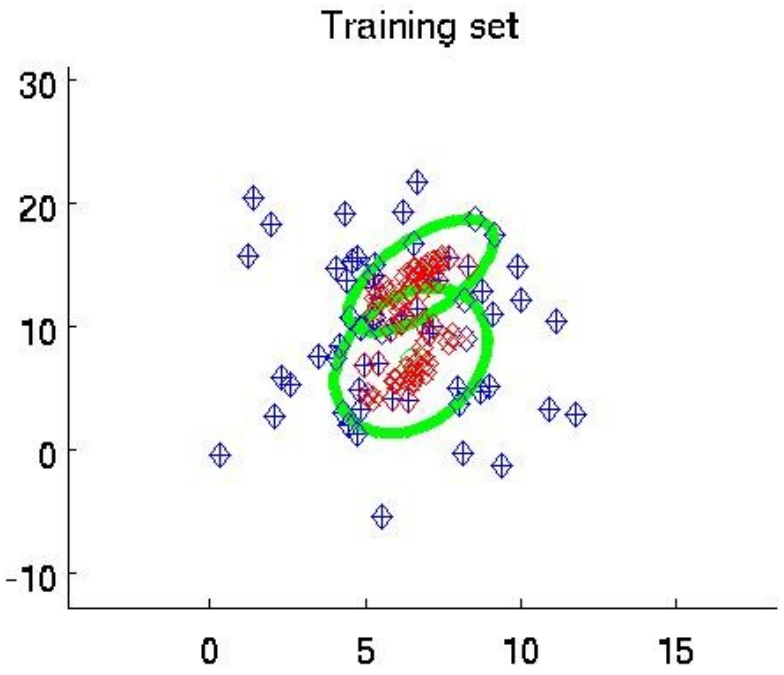
R_S s'appelle le ***risque empirique***

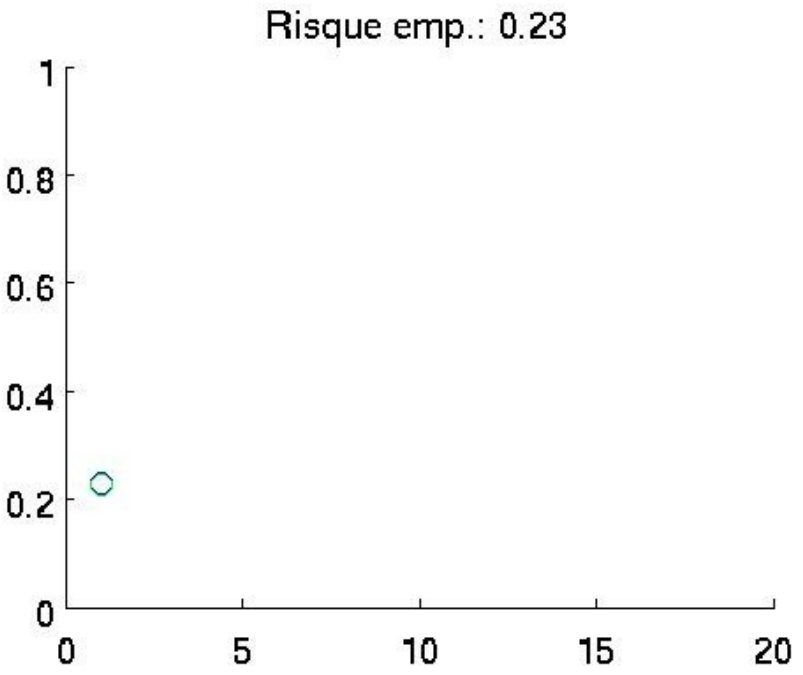
Le principe d'optimisation ci-dessus s'appelle

minimisation du risque empirique**.**

ERM : illustration

Initialisation

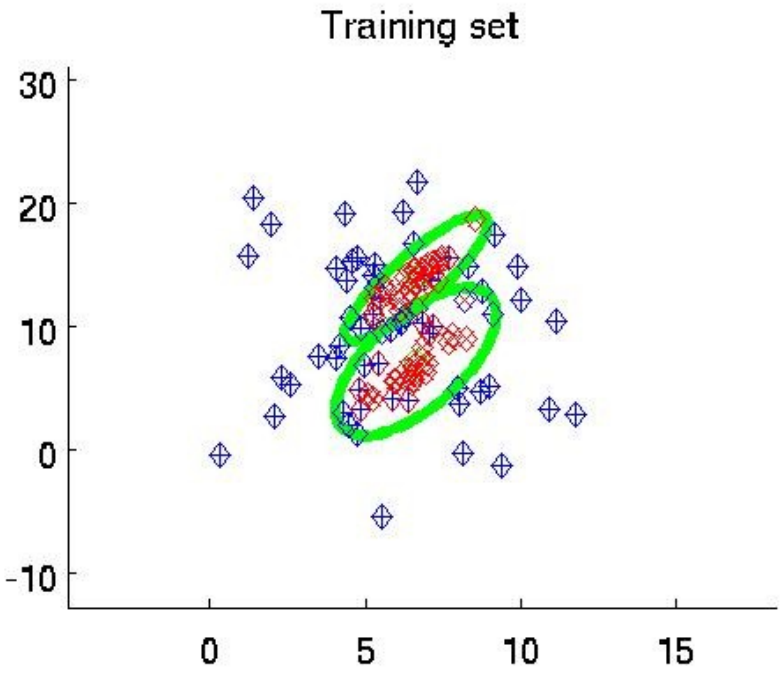


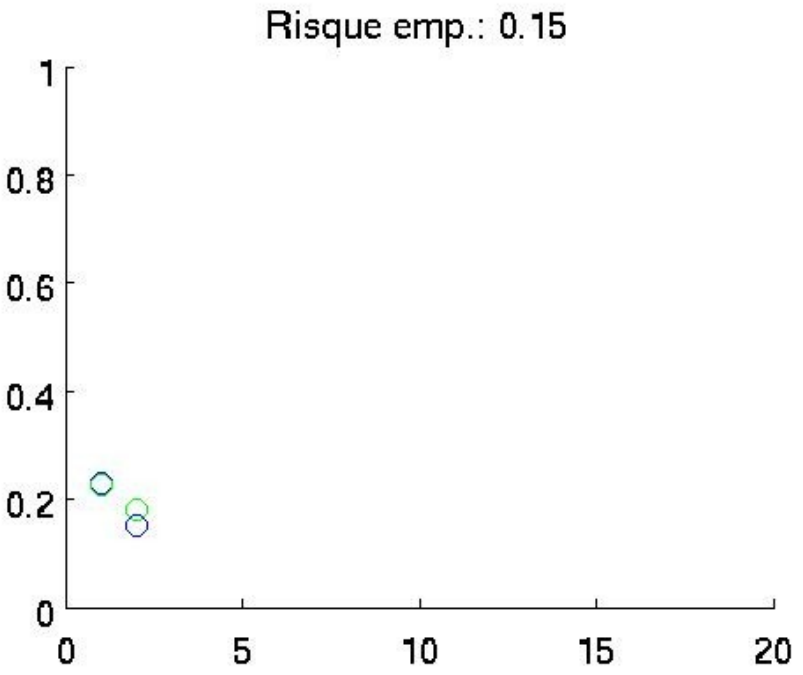


Itérations

ERM : illustration

Itération 1

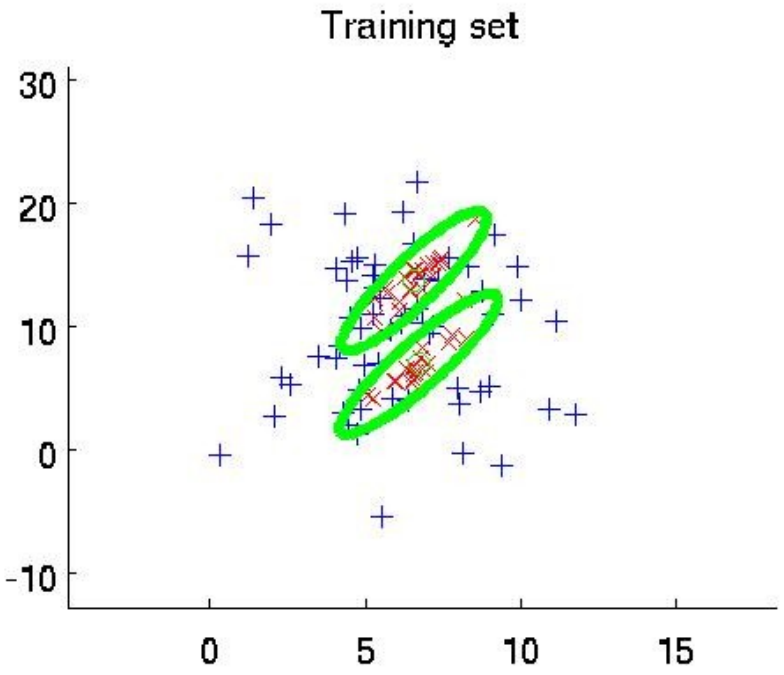


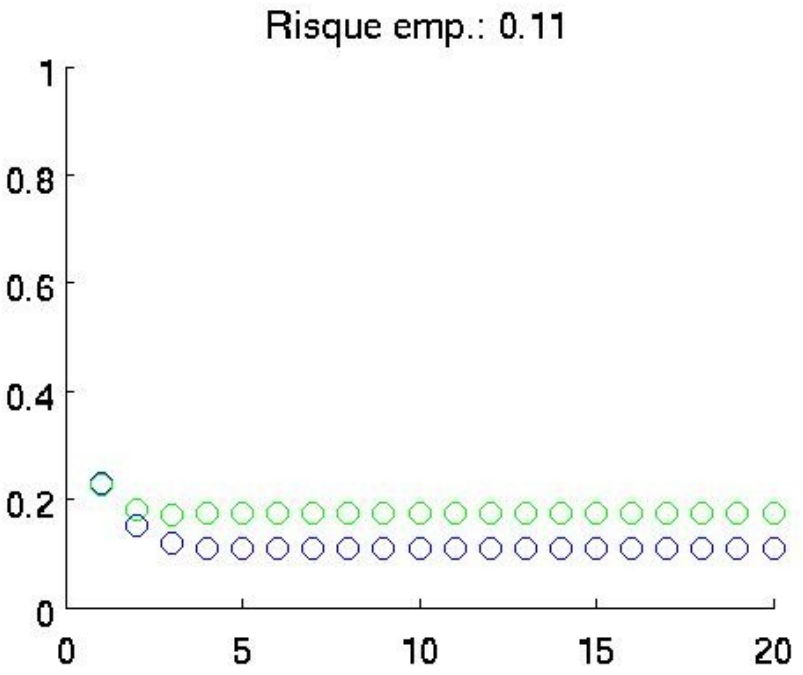


Itérations

ERM : illustration

Itération 20





Itérations

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 :

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

- En estimation de densité, $R_{emp}(h)$ est l'opposé de la log-vraisemblance de S :

$$R_{emp}(h) = 1/_{||S||} \sum_{x_i \in S} -\log (h(xi)) = -1/_{||S||} \log \prod_{x_i \in S} h(xi)$$

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.
- En estimation de densité = maximum de vraisemblance.

ERM : deux questions importantes

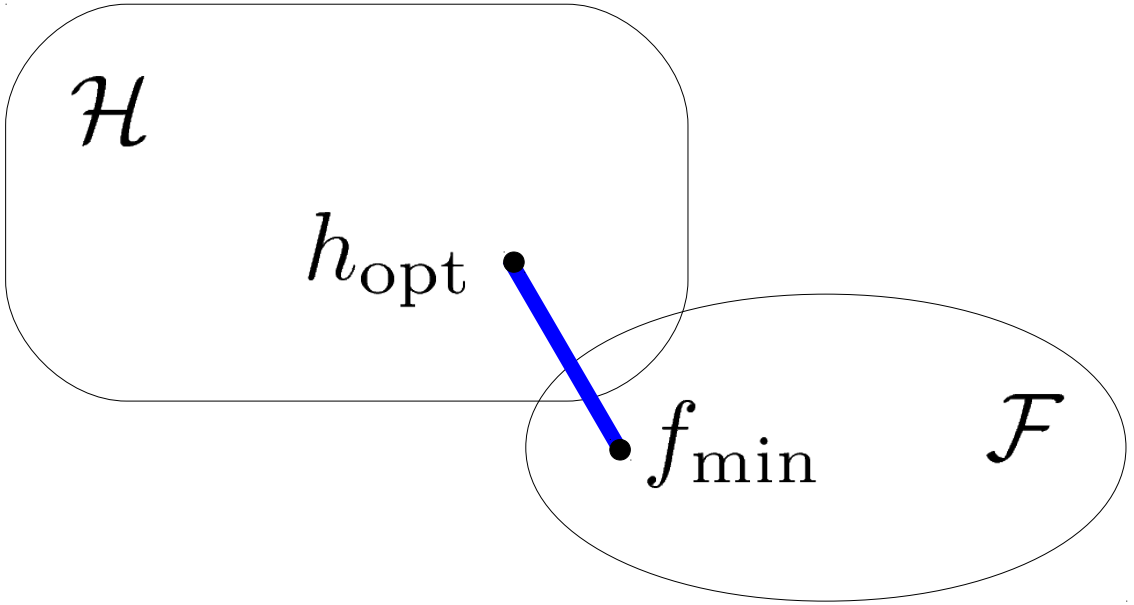
$$\text{Idéalement : } f_{\min} \triangleq \arg\min_{f} R(f)$$

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

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



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

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

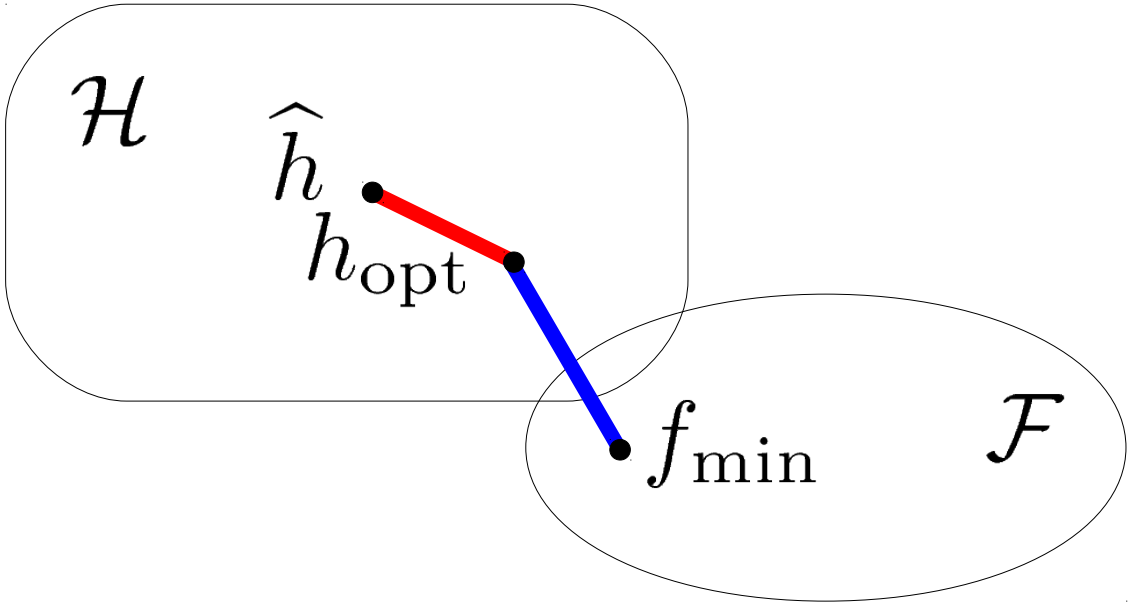
$\mathcal{H}$: espace des hypothèses possibles de l'apprenant.

$\mathcal{F}$: 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



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

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

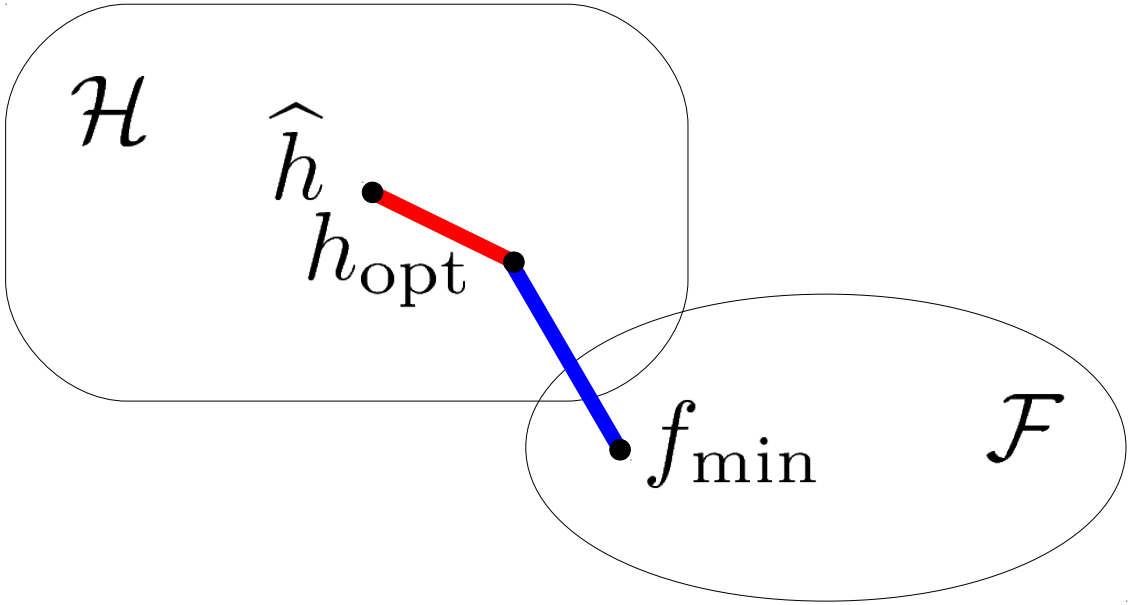
$$\widehat{h} \triangleq \arg\min_{h \in \mathcal{H}} R_S(h)$$

$R(h) \neq R_S(h)$ 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

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

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

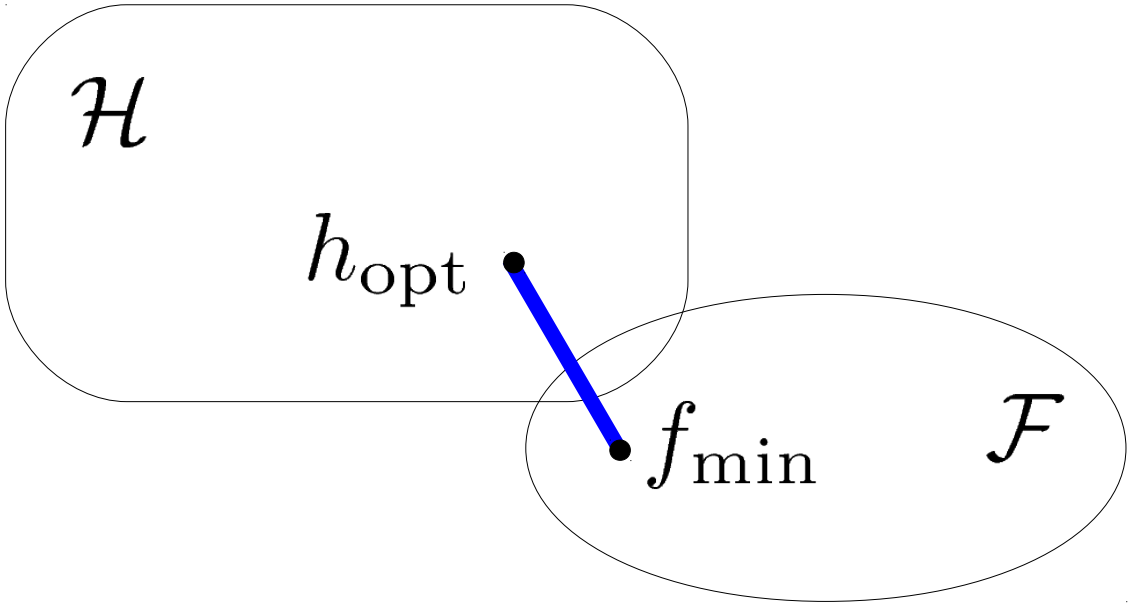
Enjeux relatifs

- 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



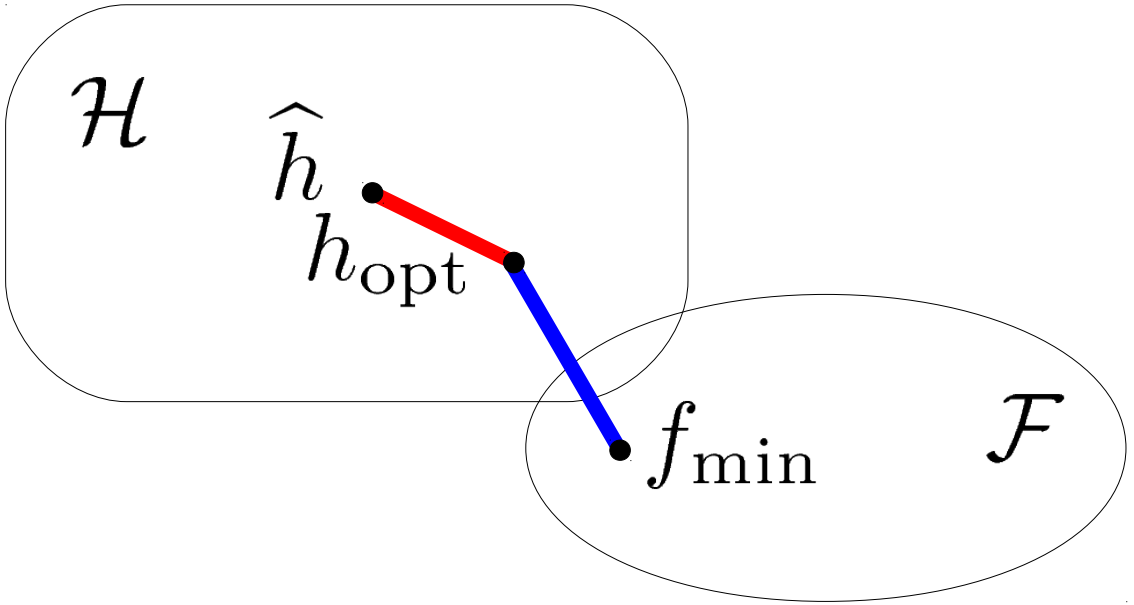
$$f_{\min} \triangleq \arg\min_{f} R(f)$$

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

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



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

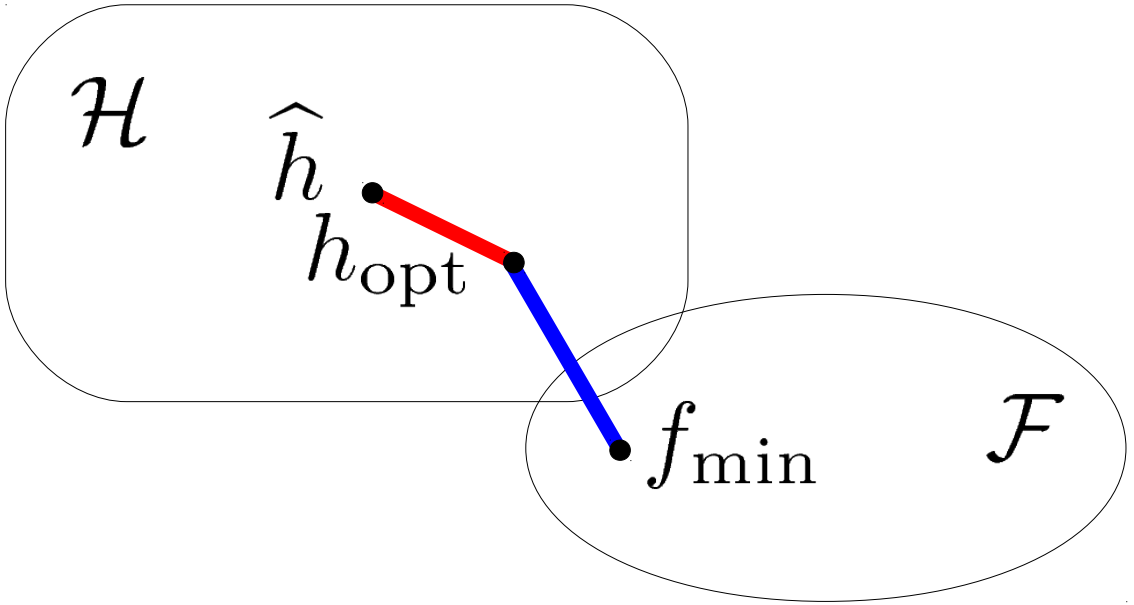
$$h_{\text{opt}} \triangleq \arg\min_{h \in \mathcal{H}} R(h)$$

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

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



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

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

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

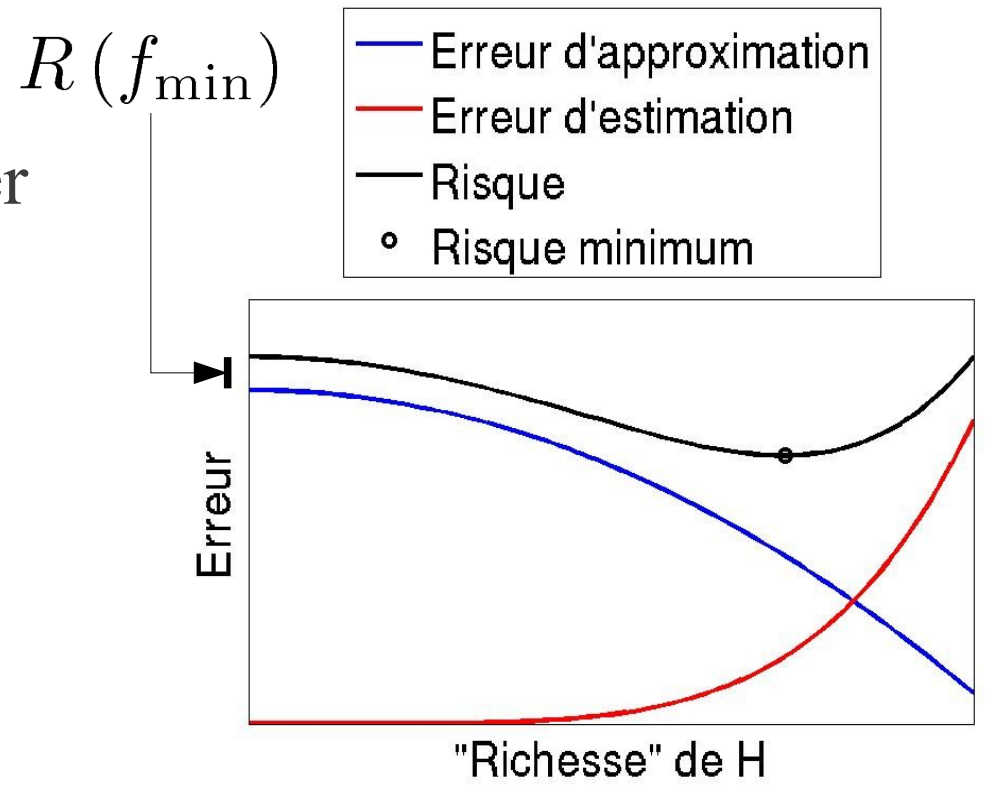
Solutions :

- 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.

Sur la difficulté d'un apprentissage

- Il existe au moins 4 raisons principales rendant un apprentissage difficile :
  - La nature non déterministe du problème.
  - La trop faible expressivité de H.
  - La non-consistance du principe ERM (ou plus généralement du principe inductif choisi) pour approcher une fonction optimal dans H.
  - La difficulté à minimiser le risque empirique (ou plus généralement à mettre en application le principe choisi).

Plan du cours

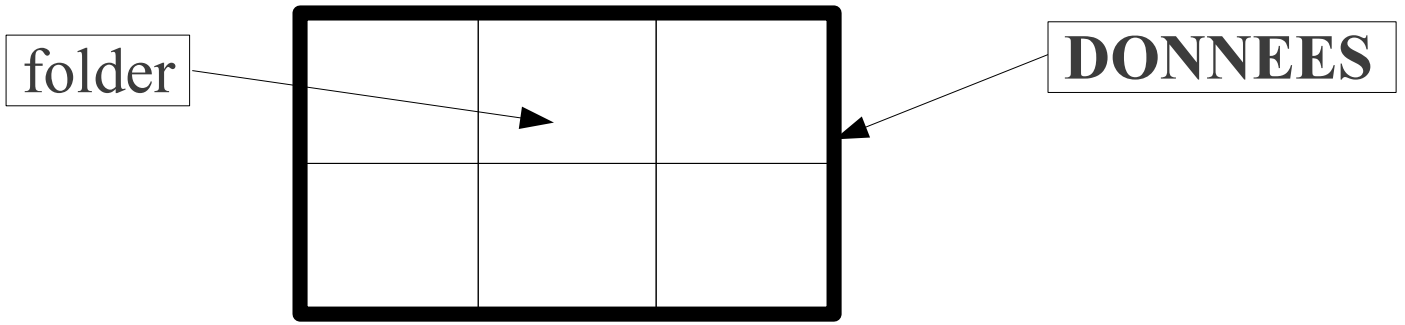
- Rappels de probabilités
- Modélisation de l'apprentissage supervisé (suite)
  - Correction de l'exercice 4 (TD1)
  - La régression
  - L'estimation de densité
- Principe ERM
- Validation d'un apprentissage

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.)

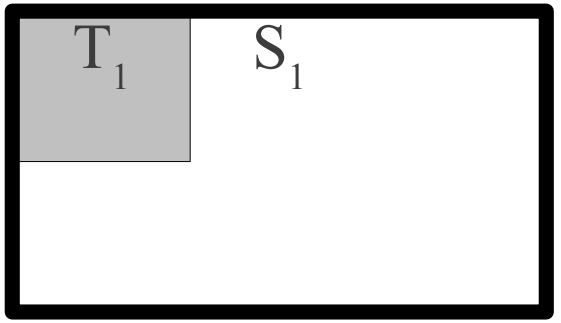
Validation croisée

- La cross-validation 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.



Validation croisée

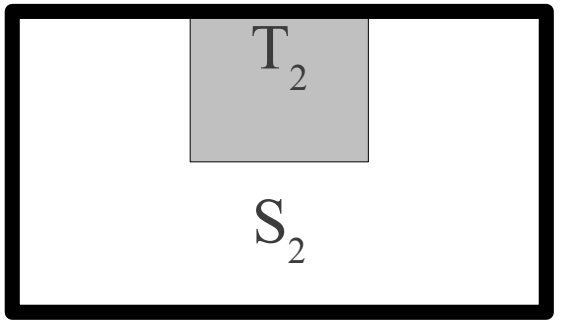
- Itération 1 :



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

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

- Itération 2 :



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

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

...

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

Application : 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 du cours

- Rappels de probabilités
- Modélisation de l'apprentissage supervisé (suite)
  - La régression
  - L'estimation de densité
- Principe ERM : risque empirique, composantes d'erreur, compromis biais-variance
- Validation d'un apprentissage
  - pour évaluer la performance d'un apprentissage
  - pour trouver un compromis biais-variance