L'induction : un saut à l'élastique

Mais avec quel élastique ?

A. Cornuéjols

AgroParisTech – INRA   MIA 518

Trame

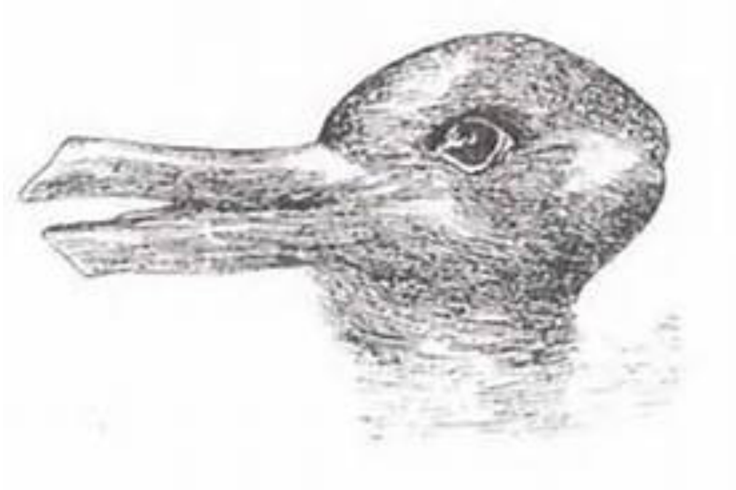
1. Diverses **illustrations** de l'induction
2. Le **no-free-lunch theorem**
3. Interprétation / **complétion** de percepts
4. Étude de l'**induction supervisée**
5. **Variantes** de l'induction supervisée
6. **Transfert, analogie, éducation** : quel principe inductif ?

Trame

1. Diverses **illustrations** de l'induction
2. Le **no-free-lunch theorem**
3. Interprétation / **complétion** de percepts
4. Étude de l'**induction supervisée**
5. **Variantes** de l'induction supervisée
6. **Transfert, analogie, éducation** : quel principe inductif ?

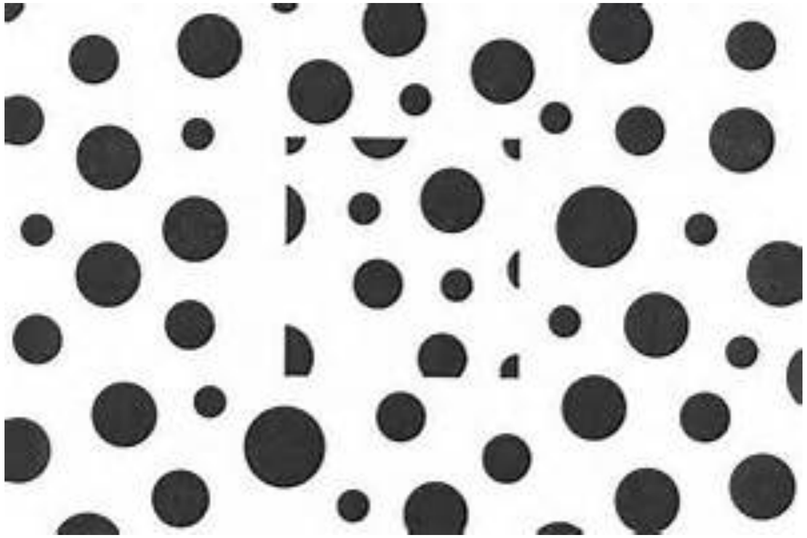
Induction(s) : Illustrations

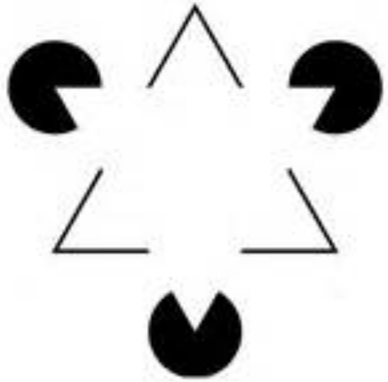
Interprétation – complétion de percepts

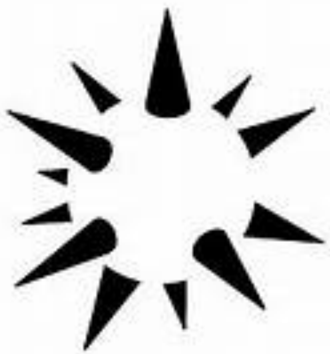


Interprétation – complétion de percepts

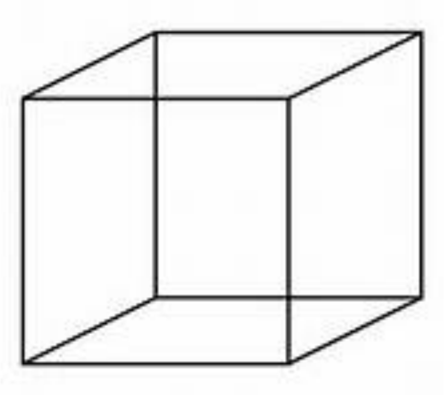
Interprétation – complétion de percepts

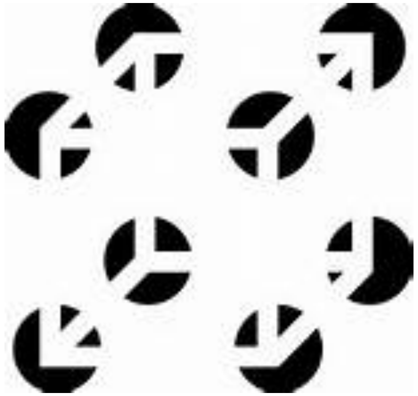




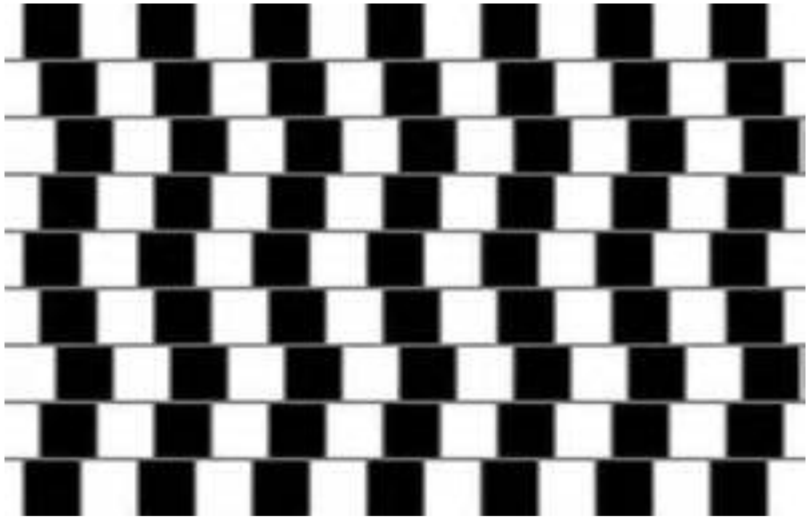


Interprétation – complétion de percepts



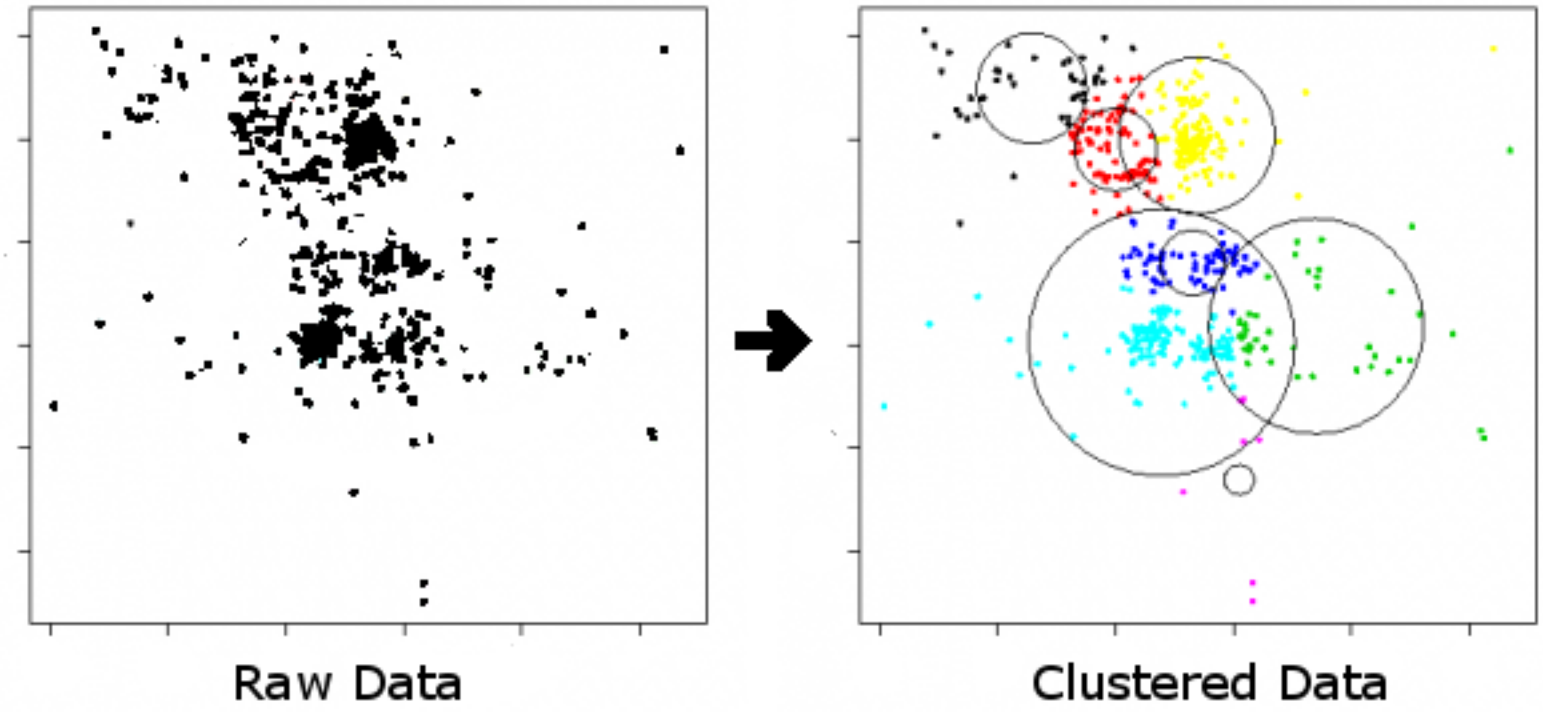


Illusions d'optique

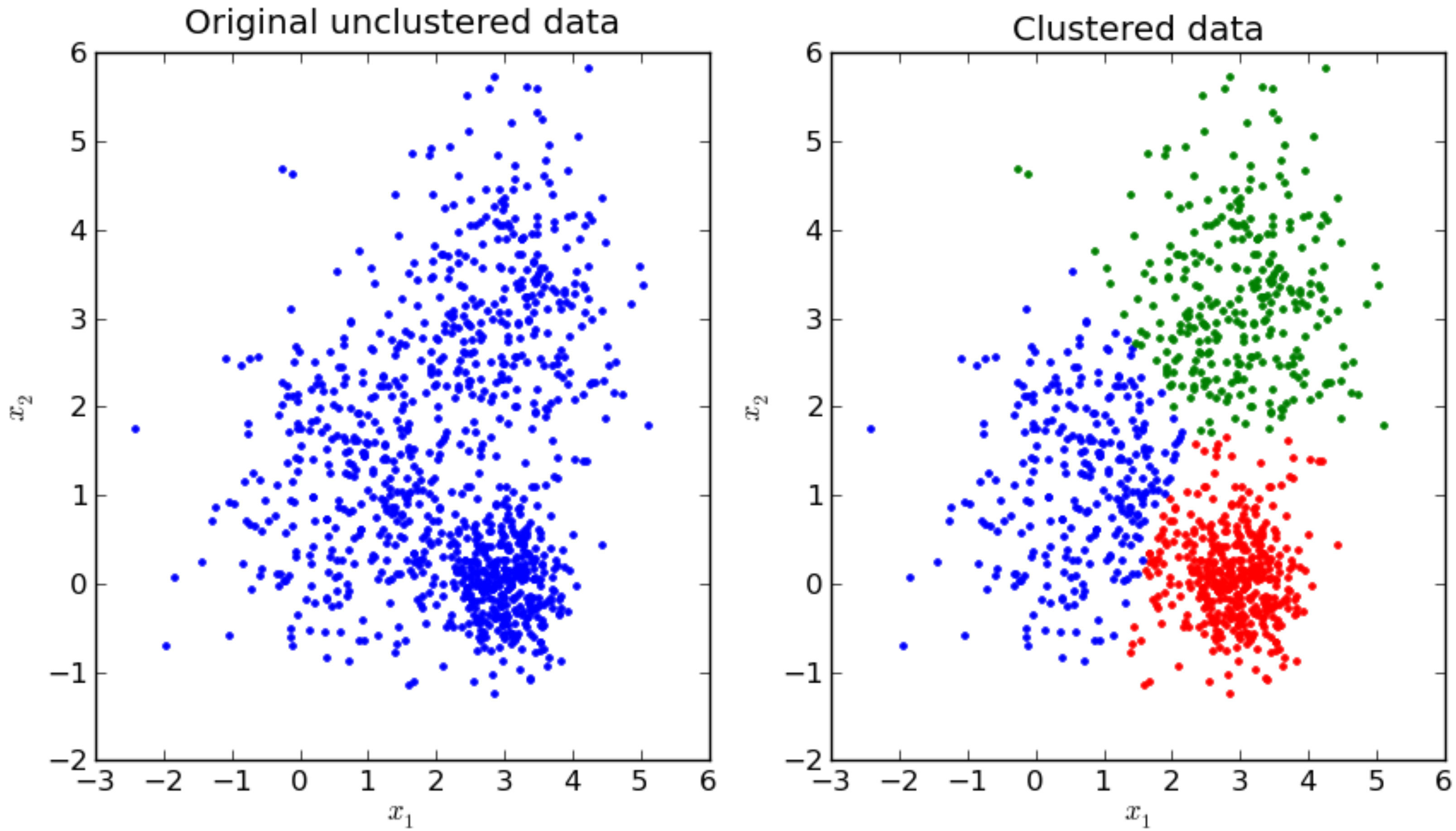


Illusion / interprétation

Clustering



Clustering


Original unclustered data
Clustered data
x_1
x_2
6
5
4
3
2
1
0
−1
−2
−3

Séquences

- 1 1 2 3 5 8 13 21 …
- 1 2 3 5 …
- 1 1 1 2 1 1 2 1 1 1 1 1 2 2 1 3 1 2 2 1 1 …

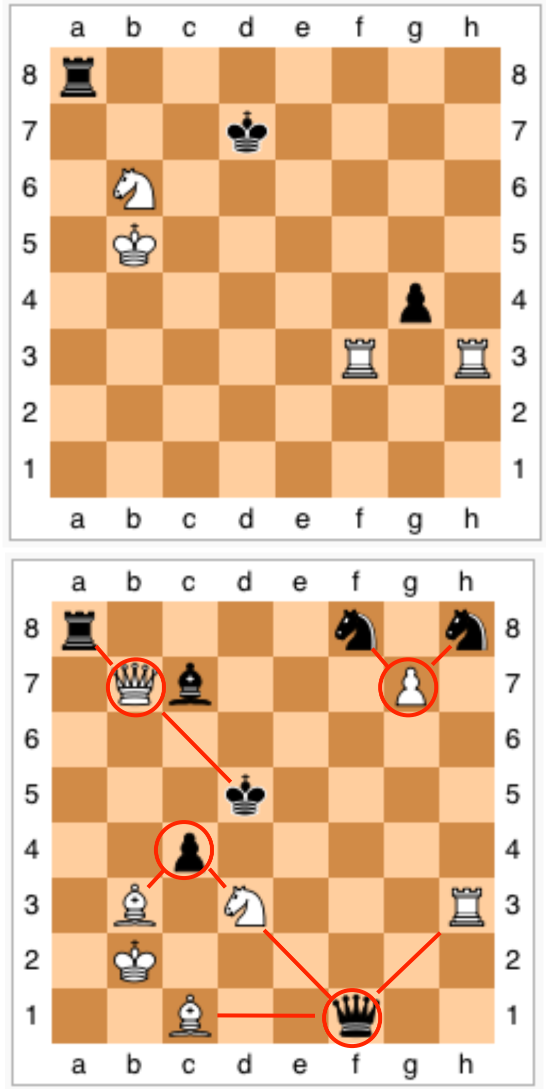
  – **Comment** ?
  – **Pourquoi** serait-il possible de faire de l'induction ?
  – Est-ce qu'**un exemple supplémentaire** doit augmenter la confiance dans la règle induite ?
  – **Combien faut-il d'exemples** ?

Induction et raisonnement : Explanation-Based Learning

1. Un **exemple unique**

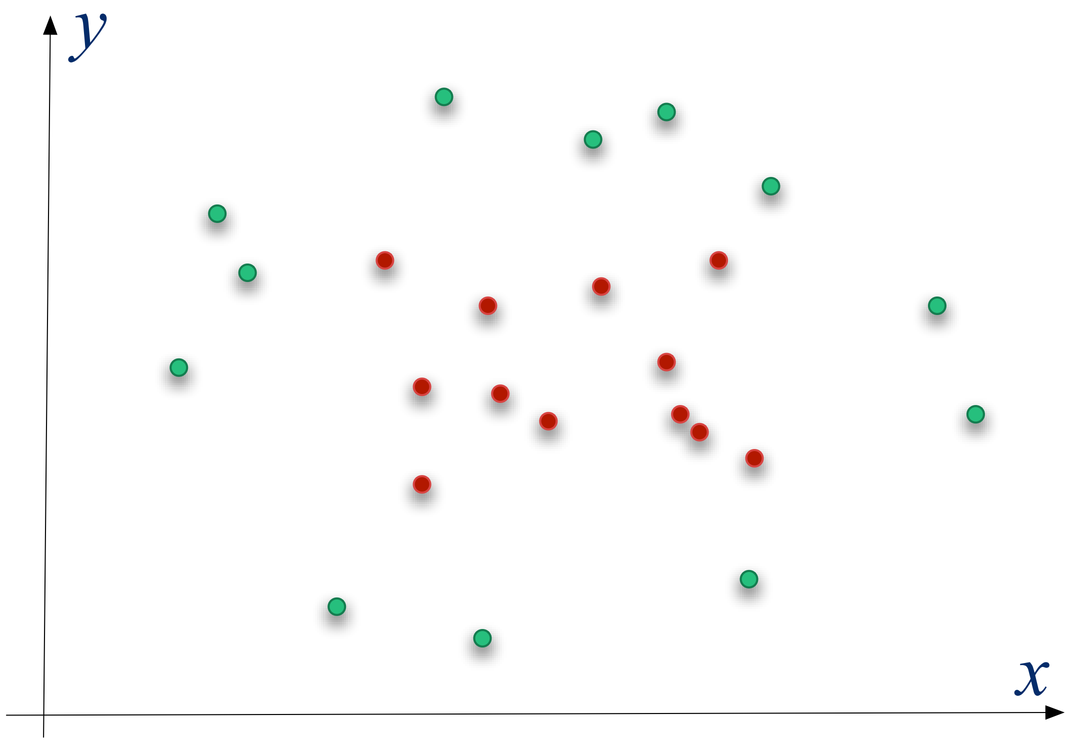
2. **Recherche de la preuve** de la « fourchette »

3. **Généralisation**



Induction supervisée

- Comment choisir la fonction de décision ?

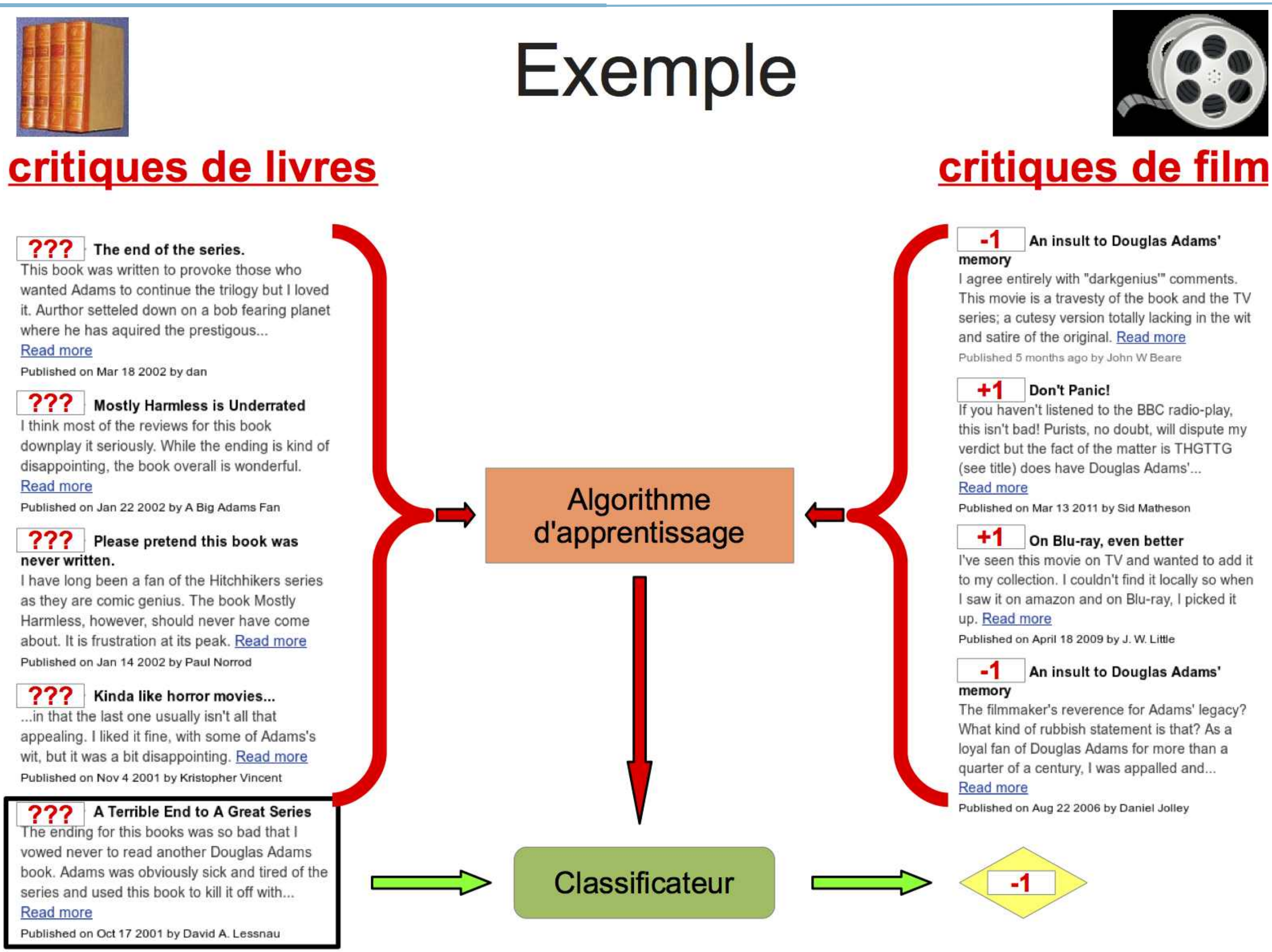


Domain adaptation

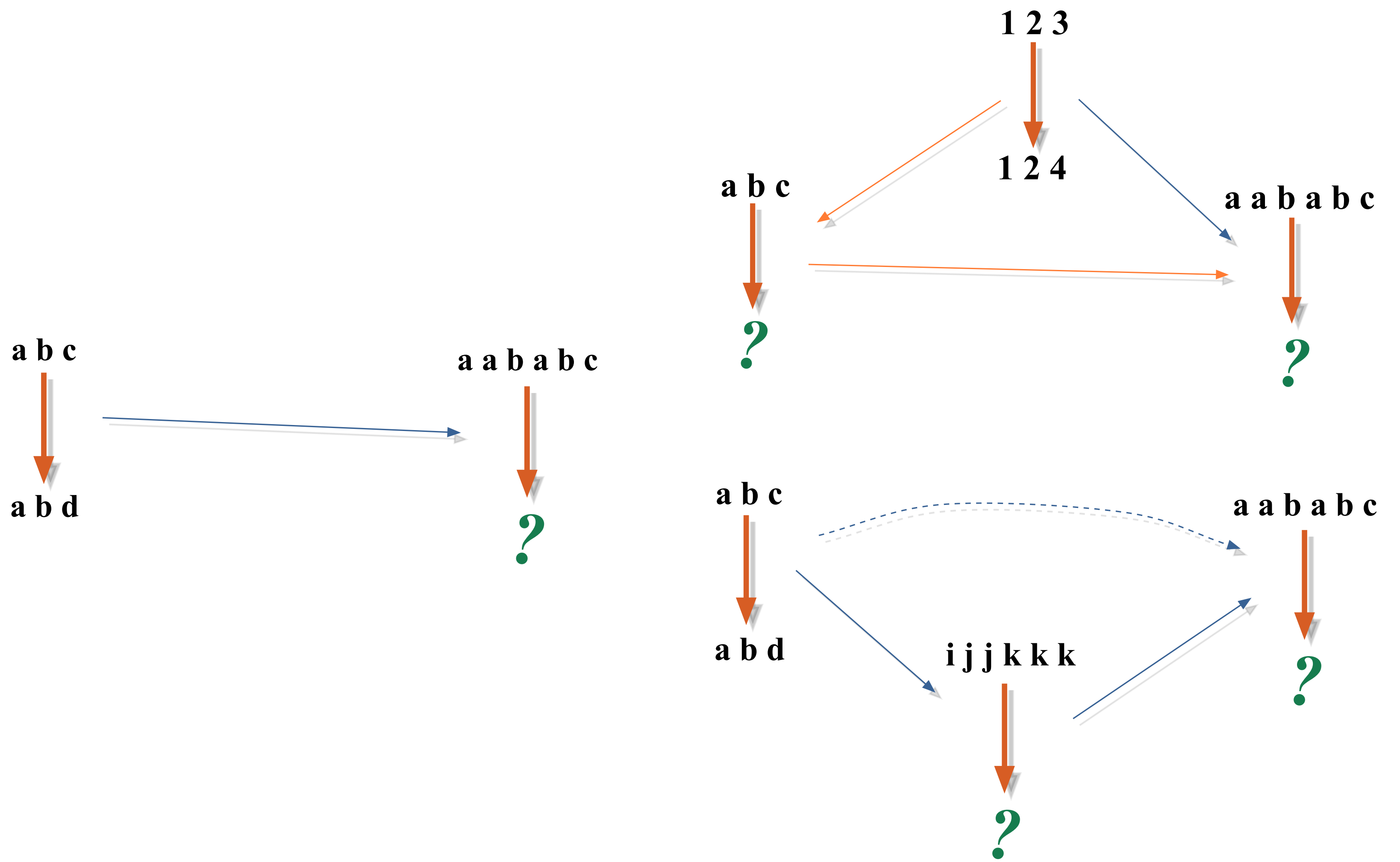
- **Definition** [Pan, TL-IJCAI'13 tutorial]
  - Ability of a system to **recognize** and **apply** knowledge and skills learned in **previous domains/tasks** to **novel domains/tasks**

- Example
  - We have **labeled images** from a **web corpus**
  - **Novel task: is there a person in unlabeled images from a video corpus?**

Domain adaptation for sentiment analysis [Pan, TL-IJCAI'13 tutorial]



Transfer and sequence effects



- toto

Interrogations

- À chaque fois :

  Cas particuliers => loi générale ou régularités

1. **Qu'est-ce qui autorise** ce passage ?

2. Est-ce que l'on peut **garantir quelque chose** ?

Trame

1. Diverses **illustrations** de l'induction

2. Le **no-free-lunch theorem**

3. Interprétation / **complétion** de percepts

4. Étude de l'**induction supervisée**

5. **Variantes** de l'induction supervisée

6. **Transfert, analogie, éducation** : quel principe inductif ?

Le no-free-lunch theorem

Trame

1. Diverses **illustrations** de l'induction
2. Le **no-free-lunch theorem**
3. Interprétation / **complétion** de percepts
4. Étude de l'**induction supervisée**
5. **Variantes** de l'induction supervisée
6. **Transfert, analogie, éducation** : quel principe inductif ?

Interprétation / complétion

de percepts

Interprétation, sciences cognitives et IA

Apprentissage de représentations parcimonieuses

Compressed sensing

- Une théorie [Candes, 2006] qui dit
  - Un signal parcimonieux
  - Peut être reconstruit
  - À partir de très peu de mesures linéaires
  - Sous certaines conditions

Acquisition du signal :

Reconstruction du signal :

If $Y = X\beta^* + \xi$, where X is an n by p matrix, $\xi \sim \mathcal{N}(0, \sigma^2 I)$, β^* has at most s non zero coefficients and $s \leq n/2$ then

$$\left\| \beta^\lambda - \beta^* \right\| \leq C s \sigma^2 \log p.$$

Bilan

- La notion de **consistance** de la méthode a remplacé la notion de **garantie**

- Notion de biais

- Biais
  - Favorable
  - Défavorable

Complétion et extrapolation de séquence ...

- … même chose ?

Extrapolation de séquence

- Peut-on confirmer une hypothèse ?

Encore un autre exemple

- Exemples décrits par :
  - *nombre* (1 ou 2); *taille* (petit ou grand); *forme* (cercle ou carré); *couleur* (rouge ou vert)
- Les objets appartiennent soit à la classe + soit à la classe -

Description	Votre réponse	Vraie réponse
1 grand carré rouge		
1 grand carré vert		
2 petits carrés rouges		
2 grands cercles rouges		
1 grand cercle vert		
1 petit cercle rouge		
1 petit carré vert		
1 petit carré rouge		
2 grands carrés verts		

L'induction : un jeu impossible ?

- **Nécessité** d'un biais

- **Types** de biais
  - Biais de **représentation** (déclaratif)
  - Biais de **recherche** (procédural)

Trame

1. Diverses **illustrations** de l'induction

2. Le **no-free-lunch theorem**

3. Interprétation / **complétion** de percepts

4. Étude de l'**induction supervisée**

5. **Variantes** de l'induction supervisée

6. **Transfert, analogie, éducation** : quel principe inductif ?

Étude de

l'induction supervisée

Trame

1. Diverses **illustrations** de l'induction

2. Le **no-free-lunch theorem**

3. Interprétation / **complétion** de percepts

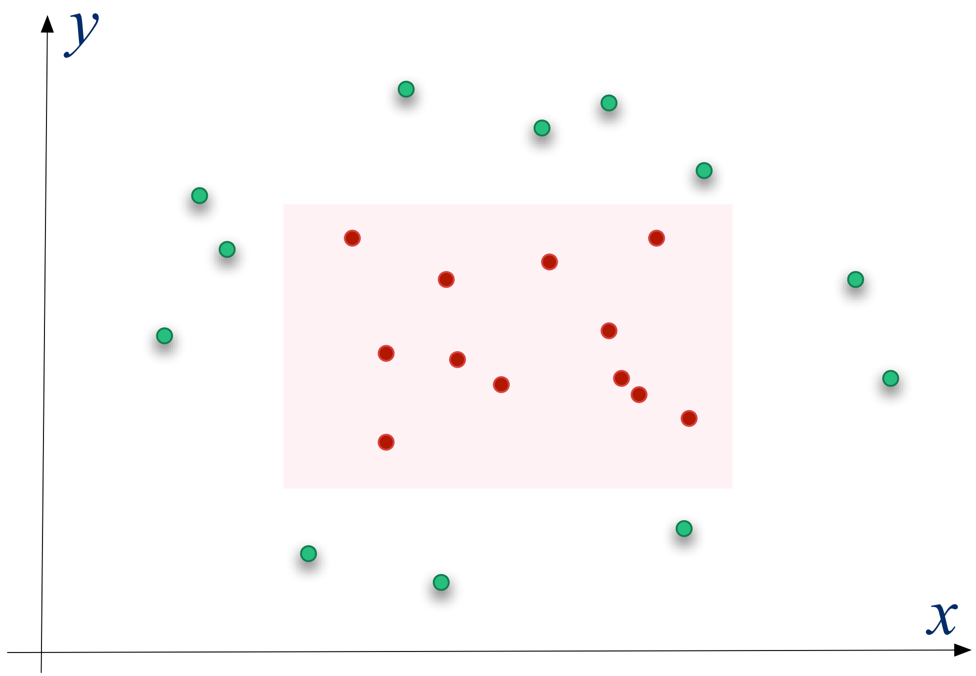
4. Étude de l'**induction supervisée**

5. **Variantes** de l'induction supervisée

6. **Transfert, analogie, éducation** : quel principe inductif ?

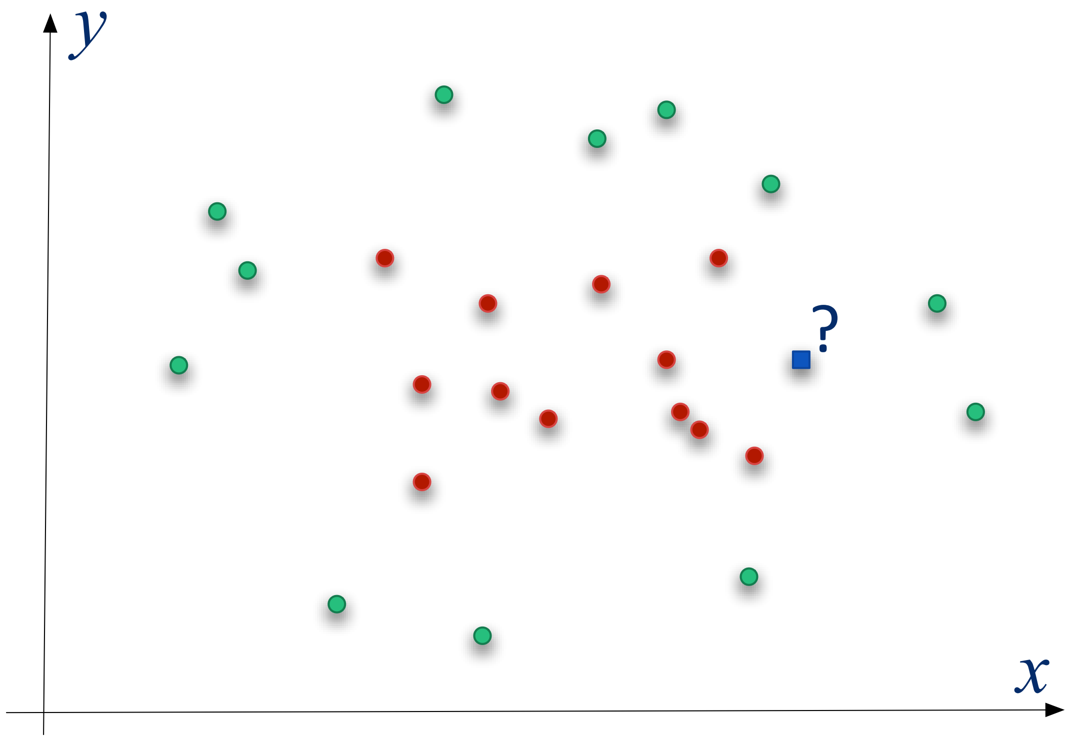
Apprentissage de rectangle

- Échantillon
  - D'exemples positifs $\mathbf{P}^{+}_{\mathcal{X}}$
  - D'exemples négatifs $\mathbf{P}^{-}_{\mathcal{X}}$



Apprentissage de rectangle

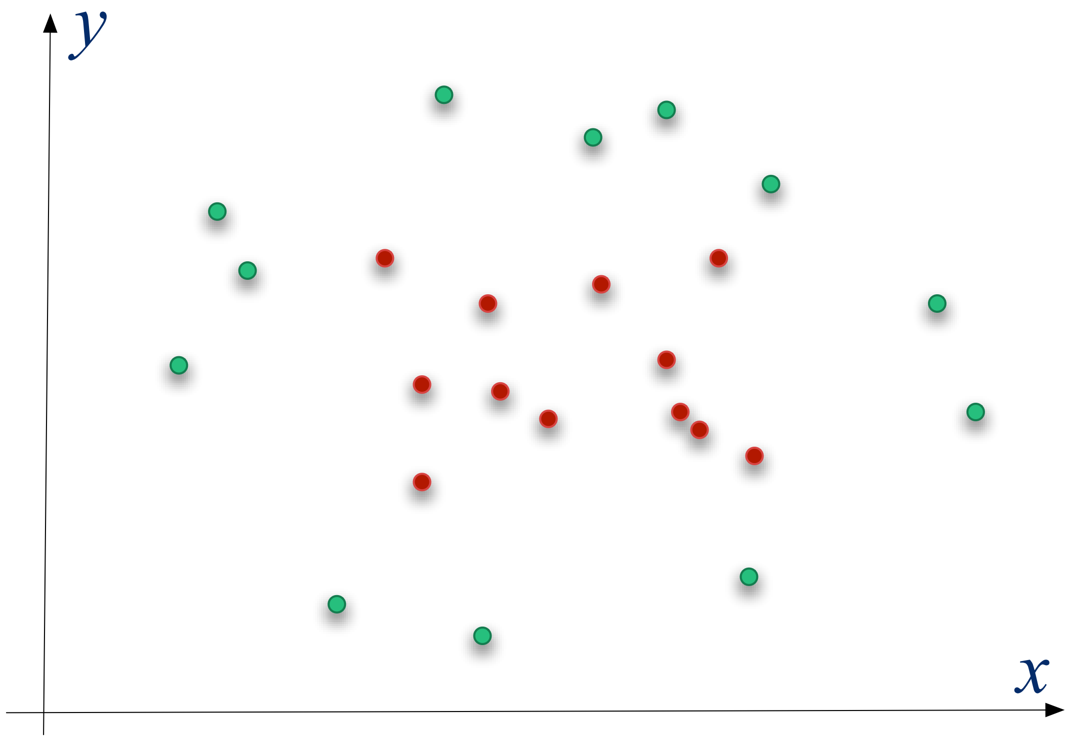
- Que cherche-t-on à apprendre ?



→ Une fonction de **décision** (de **prédiction**)

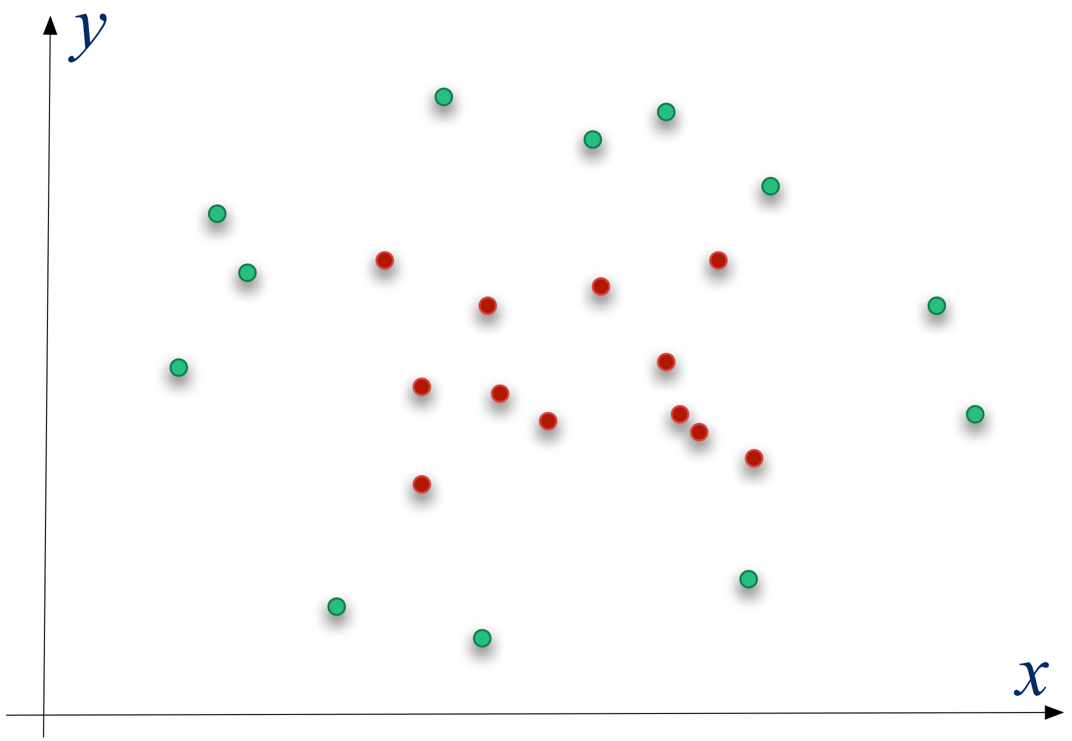
Apprentissage de rectangle

- Comment apprendre ?



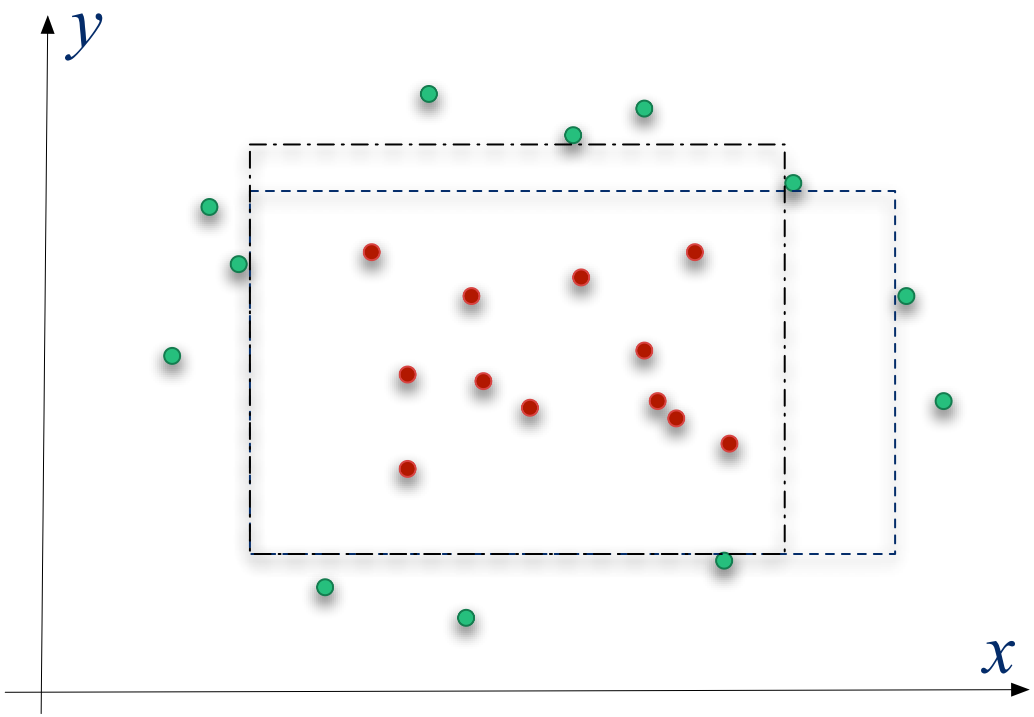
Apprentissage de rectangle

- Comment apprendre ?
  - Si **je sais que le concept cible est un rectangle**



Apprentissage de rectangle

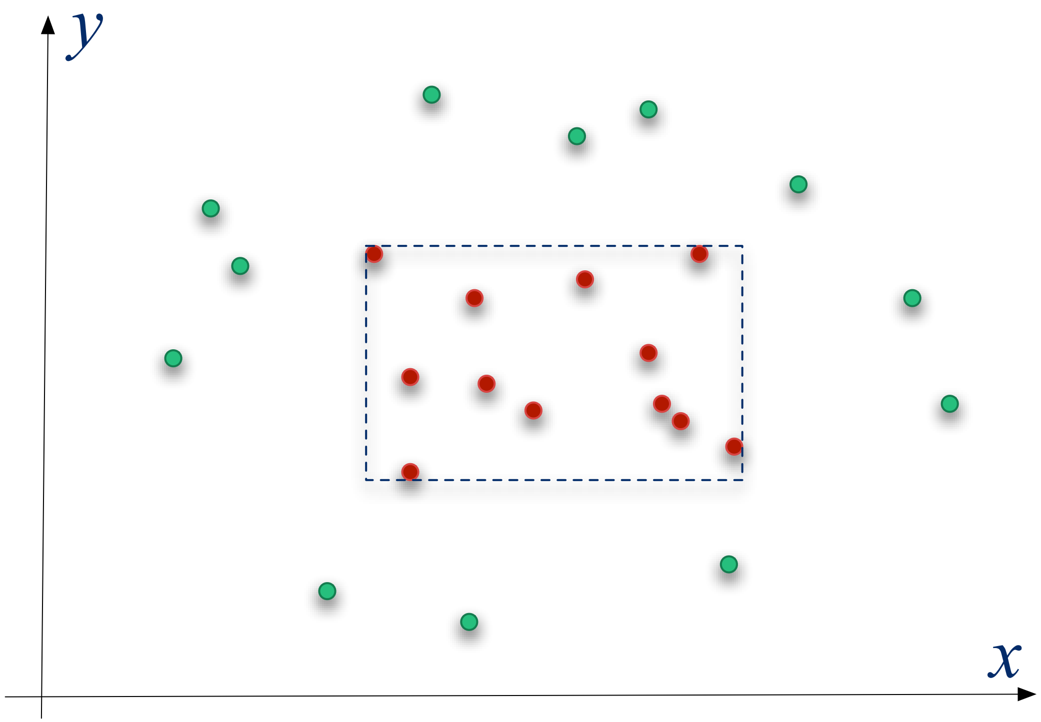
- Comment apprendre ?
  - Si **je sais que le concept cible est un rectangle**



Hypothèses les plus **générales**

Apprentissage de rectangle

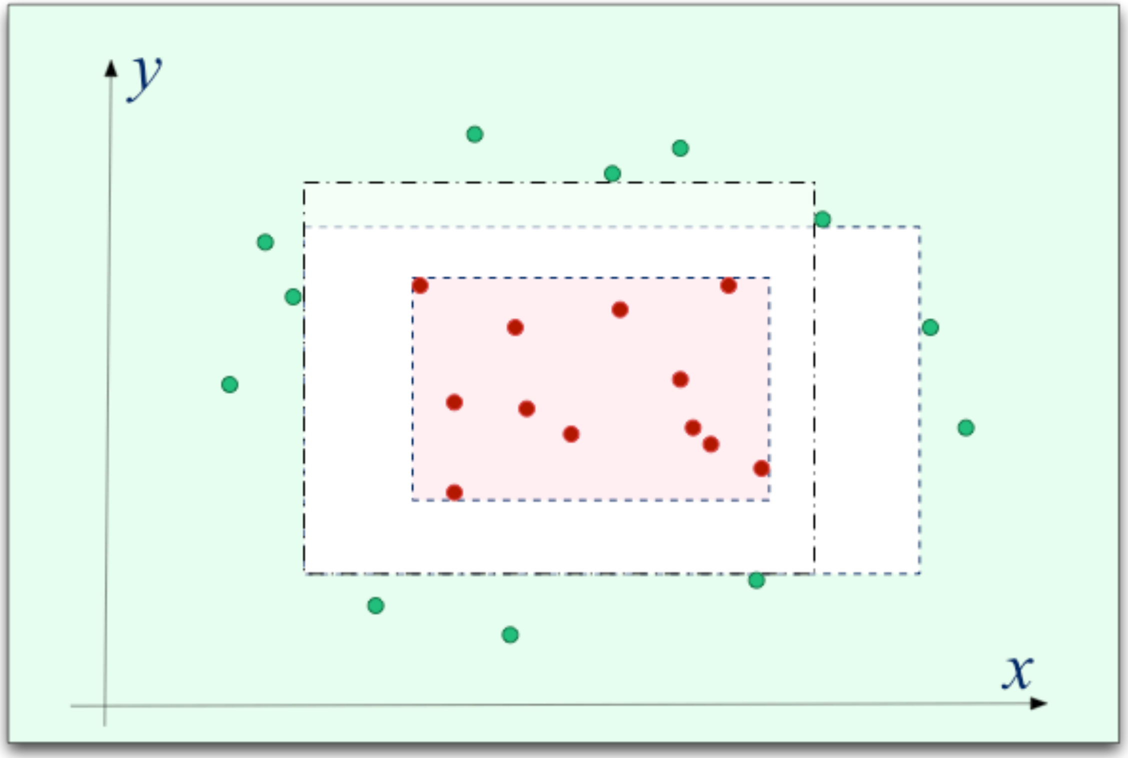
- Comment apprendre ?
  - Si **je sais que le concept cible est un rectangle**



Hypothèses les plus **spécifiques**

Apprentissage de rectangle

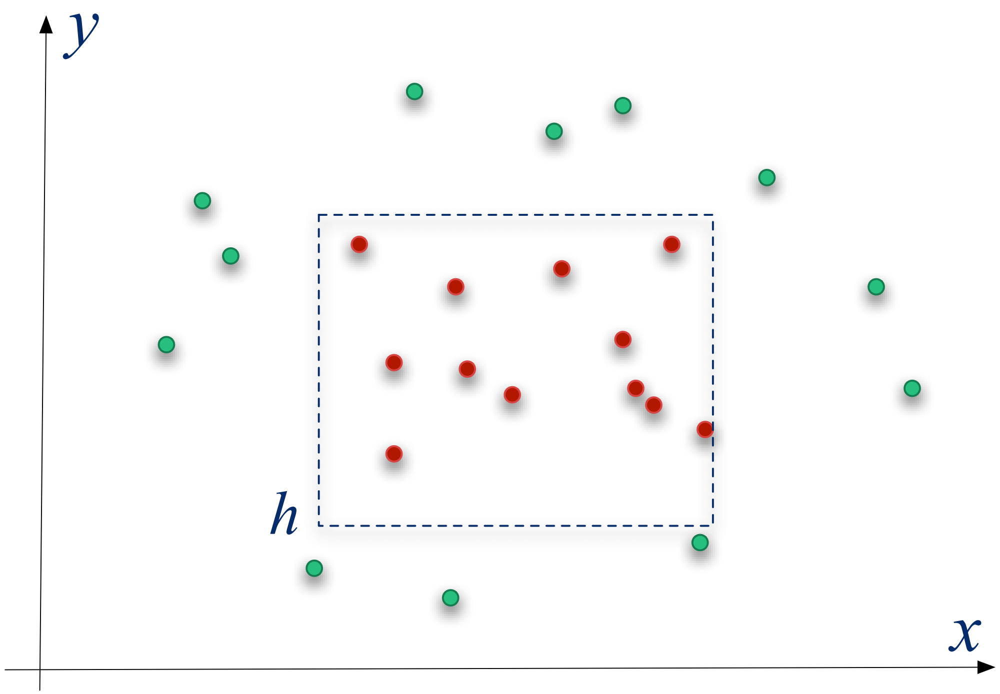
- Comment apprendre ?
  - Choix d'une hypothèse *h*



Espace des versions

Apprentissage de rectangle

- Apprentissage : choix de *h*
  - Quelle performance ?



1ère **étude statistique** de l'induction

Quelle performance ?

- Coût d'une erreur de prédiction
  - La ***fonction de perte***

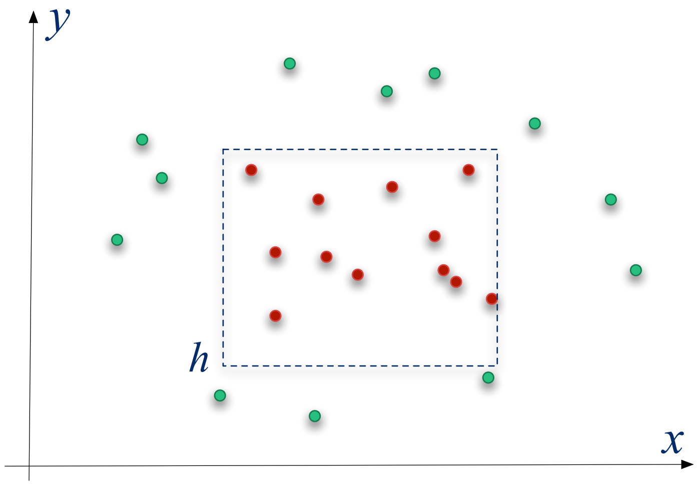
$$\ell\big(h(\mathbf{x}), y\big)$$

- Quel coût à venir si je choisis *h* ?
  - Espérance de coût : le « ***risque réel*** »

$$R(h) \;=\; \int_{\mathcal{X}\times\mathcal{Y}} \ell\big(h(\mathbf{x}), y\big)\; \mathbf{p}_{\mathcal{X}\mathcal{Y}}(\mathbf{x}, y)\; d\mathbf{x}\, dy$$

1ère **étude statistique** de l'induction

- Quelle performance attendue pour *h* ?
  - Pas d'erreur sur l'échantillon d'apprentissage *S*



Le « risque empirique »

$$\hat{R}(h) \;=\; \frac{1}{m} \sum_{i=1}^{m} \ell\big(h(\mathbf{x}_i), y_i\big)$$

Nouvelle théorisation

- Le principe de minimisation du risque empirique (ERM)

  … est-il sain ?

  – Si je choisis *h* telle que

$$\hat{h} \;=\; \underset{h \in \mathcal{H}}{\operatorname{ArgMin}} \; \hat{R}(h)$$

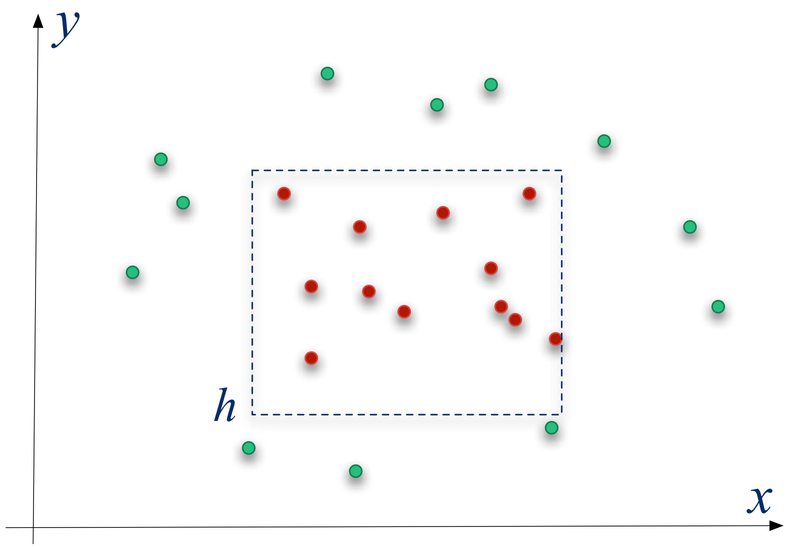
  – Est-ce que *h* est bonne relativement au risque réel ?

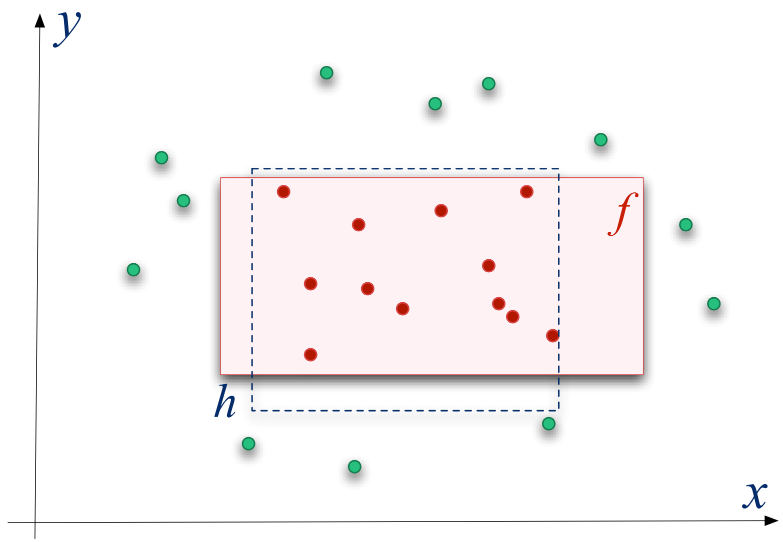
$$h^* \;=\; \underset{h \in \mathcal{H}}{\operatorname{ArgMin}} \; R(h)$$

$$R(h^*) \overset{?}{\longleftrightarrow} R(\hat{h})$$

1ère **étude statistique** de l'induction

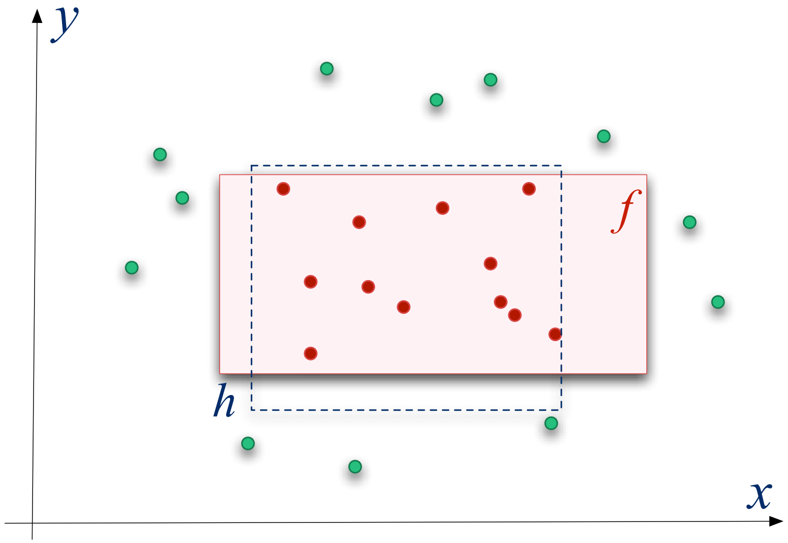
- Stratégie d'apprentissage :
  – choix d'une hypothèse de risque empirique nul (pas d'erreur sur l'échantillon d'apprentissage *S*)
  – Quelle performance attendue pour *h* ?

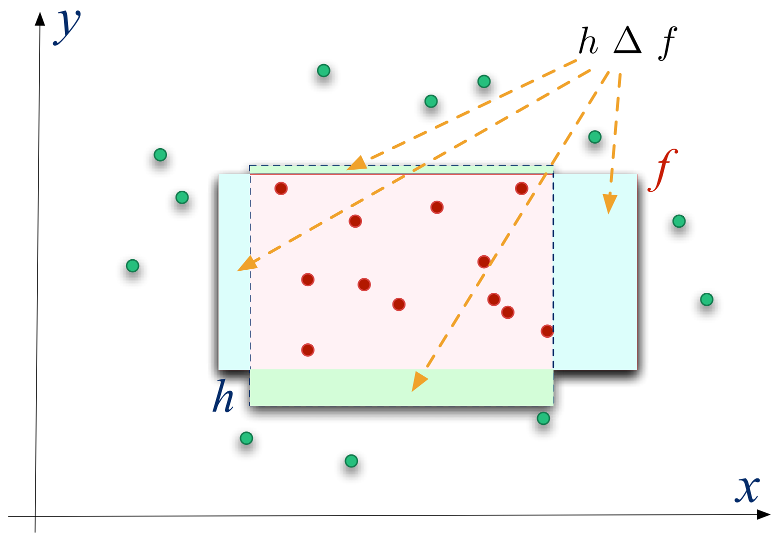




$1^{ère}$ **étude statistique** de l'induction

- choix d'une hypothèse de risque empirique nul (pas d'erreur sur l'échantillon d'apprentissage *S*)
- Quelle performance attendue pour *h* ?
- Quel est le risque d'avoir une erreur $R(h) > \varepsilon$?





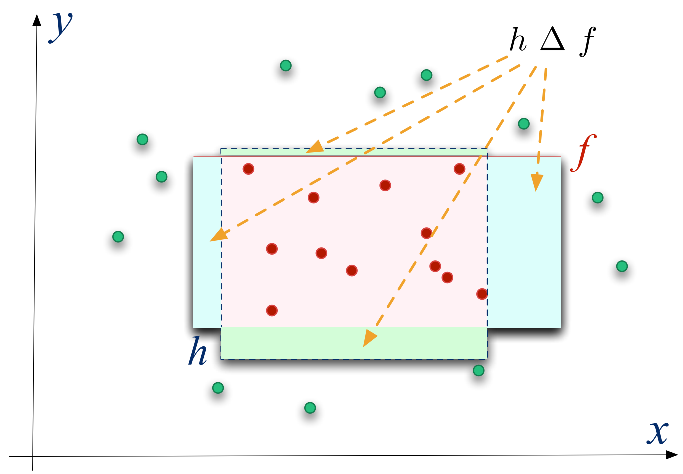
$1^{\text{ère}}$ **étude statistique** de l'induction

- Supposons *h* tq. $R(h) \geq \varepsilon$
- Quelle est la probabilité que pourtant *h* ait été sélectionnée ?

$$R(h) \; = \; \mathbf{p}_{\mathcal{X}}(h \,\Delta\, f)$$

Après un exemple : $p\big(\hat{R}(h) = 0\big) \;\leq\; 1 - \varepsilon$

« tombe » en dehors de $h \,\Delta\, f$



Après *m* exemple (i.i.d.) :

$$p^m\big(\hat{R}(h) = 0\big) \;\leq\; (1 - \varepsilon)^m$$

On veut : $\forall\, \varepsilon, \delta \in [0,1] : \quad p^m\big(R(h) \geq \varepsilon\big) \;\leq\; \delta$

1ère **étude statistique** de l'induction

- On cherche : $\forall\, \varepsilon, \delta \in [0,1] : \quad p^m\big(R(h) \geq \varepsilon\big) \;\leq\; \delta$

Soit :

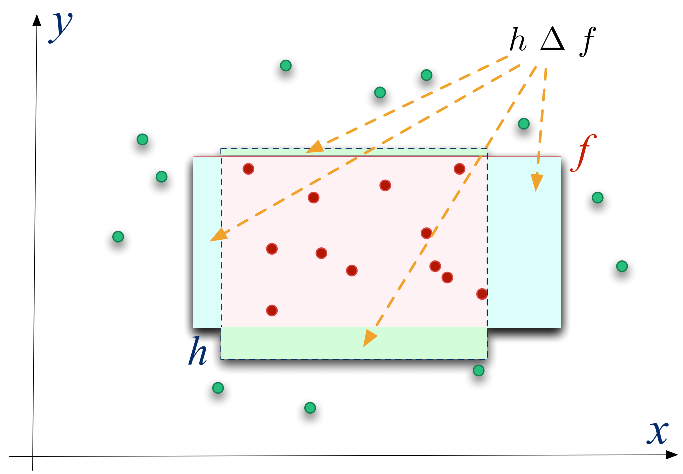
$$(1 - \varepsilon)^m \;\leq\; \delta$$

$$e^{-\varepsilon\, m} \;\leq\; \delta$$

$$-\varepsilon\, m \;\leq\; \ln(\delta)$$

D'où :

$$\boxed{m \;\geq\; \frac{\ln(1/\delta)}{\varepsilon}}$$



L'analyse « PAC learning »

- Quelle est la probabilité que je choisisse une hypothèse h_{err} de risque réel > ε **et que je ne m'en aperçoive pas** après l'observation de *m* exemples ?
- Probabilité de survie de ***h*$_{err}$ après 1 exemple** : $(1 - \varepsilon)$
- Probabilité de survie de ***h*$_{err}$ après *m* exemples** : $(1 - \varepsilon)^m$
- Probabilité de survie d'**au moins une hypothèse dans $\mathcal{H}$** : $|\mathcal{H}|\,(1 - \varepsilon)^m$
  - On utilise la probabilité de l'union $\mathbf{P}(A \cup B) \;\leq\; \mathbf{P}(A) + \mathbf{P}(B)$

- On veut que la probabilité qu'il reste au moins une hypothèse de risque réel > ε dans l'espace des versions soit bornée par δ :

$$|\mathcal{H}|\,(1 - \varepsilon)^m \;<\; |\mathcal{H}|e^{(-\varepsilon\, m)} \;<\; \delta$$

$$\log|\mathcal{H}| - \varepsilon\, m \;<\; \log\delta$$

$$\boxed{m \;>\; \frac{1}{\varepsilon}\,\log\frac{|\mathcal{H}|}{\delta}}$$

2ème étude statistique de l'induction

- L'hypothèse est choisie sur la base de *S*

« cas réalisable »

- On veut donc en fait :

$$\forall\, \varepsilon, \delta \in [0,1] : \quad p^m \big(\exists h \; : \; R(h) \geq \varepsilon\big) \;\leq\; \delta$$

- On suppose : $\quad |\mathcal{H}| \; < \; \infty$

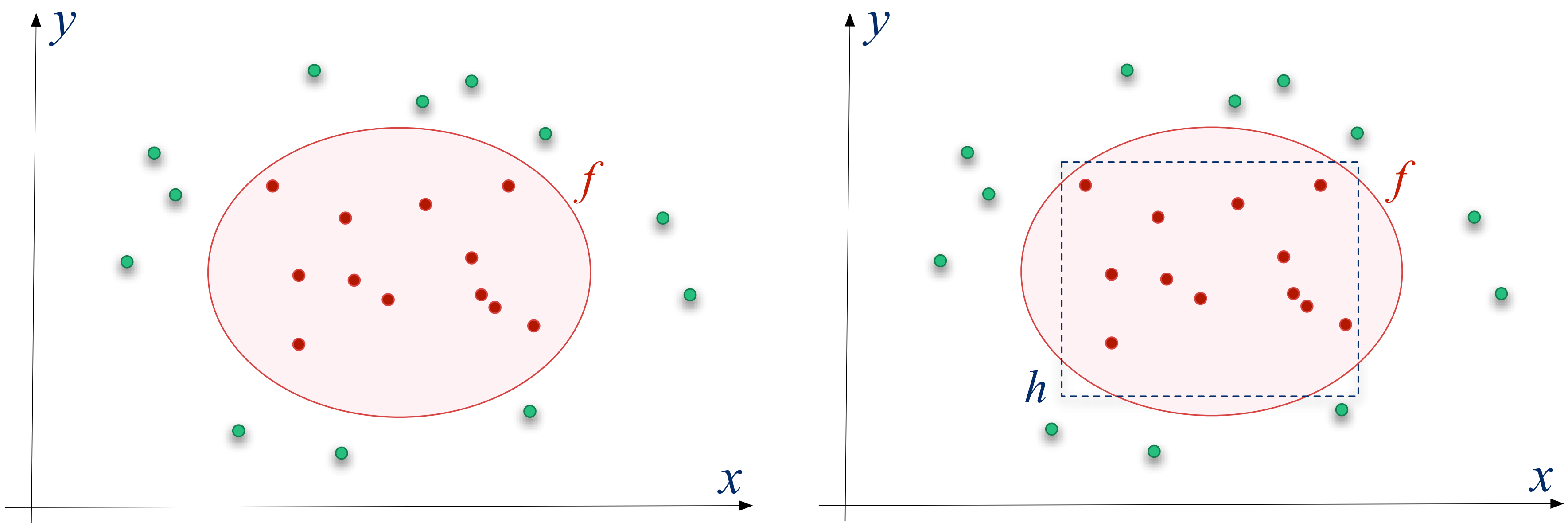
Alors : $\quad |\mathcal{H}|\,(1 - \varepsilon)^m \;\leq\; |\mathcal{H}|\, e^{-\varepsilon\, m} \;=\; \delta$

$$-\varepsilon\, m \;\leq\; \ln(\delta) \;-\; \ln(|\mathcal{H}|)$$

$$\boxed{m \;\geq\; \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}}$$

Cas « non réalisable »

- $\mathcal{H} \neq \mathcal{F}$



Il n'existe plus d'hypothèse de risque réel nul.

Pas de garantie non plus de trouver une hypothèse de risque empirique nul.

3ème étude statistique de l'induction

Théorème 1 (Inégalité de Hoeffding). *Si les ξ_i sont des variables aléatoires, tirées **indépendamment** et selon une **même distribution** et prenant leur valeur dans l'intervalle $[a, b]$, alors :*

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i - \mathbb{E}(\xi)\right| \geq \varepsilon\right) \;\leq\; 2\exp\left(-\frac{2\,m\,\varepsilon^2}{(b-a)^2}\right)$$

Appliquée au risque empirique et au risque réel, cette inégalité nous donne :

$$P(|R_{\text{Emp}}(h) - R_{\text{Réel}}(h)| \geq \varepsilon) \;\leq\; 2\exp(-\frac{2\,m\,\varepsilon^2}{(b-a)^2}) \tag{1}$$

si la fonction de perte ℓ est définie sur l'intervalle $[a, b]$.

« $\mathcal{H}$ fini »

$$\begin{aligned} P^m[\exists h \in \mathcal{H} : R_{\text{Réel}}(h) - R_{\text{Emp}}(h) > \varepsilon] \;&\leq\; \sum_{i=1}^{|\mathcal{H}|} P^m[R_{\text{Réel}}(h^i) - R_{\text{Emp}}(h^i) > \varepsilon] \\ &\leq\; |\mathcal{H}|\exp(-2\,m\,\varepsilon^2) \;=\; \delta \end{aligned}$$

en supposant ici que la fonction de perte ℓ prend ses valeurs dans l'intervalle $[0, 1]$.

3ème étude statistique de l'induction

- On en tire :

$$\varepsilon \;=\; \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2\,m}} \qquad \textbf{et} \qquad m \;\geq\; \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2\,\varepsilon^2}$$

- Au lieu de (« cas réalisable ») :

$$\varepsilon \;=\; \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{m} \qquad \textbf{et} \qquad m \;\geq\; \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{\varepsilon}$$

Quelques remarques

1. On n'a pas vraiment parlé d'***algorithme d'apprentissage*** !?
   - Où est l'apprentissage ?

2. Importance cruciale de l'hypothèse de ***stationnarité*** et d'***indépendance*** : tirage i.i.d.

3. Importance du **choix de** $\mathcal{H}$

Lien entre risque réel et risque empirique

- $\mathcal{H}$ fini, cas réalisable

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m\left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{m}\right] \;>\; 1 - \delta$$

- $\mathcal{H}$ fini, cas **non** réalisable

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m\left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2\,m}}\right] \;>\; 1 - \delta$$

4ème **étude statistique** de l'induction

- Cas non réalisable et $\mathcal{H}$ non finie
- L'analyse de Vapnik et Chervonenkis (puis d'autres)

4ème **étude statistique** de l'induction

- Cas non réalisable et $\mathcal{H}$ non finie

Comment faire ?

– Principe général :

1. **Réduire** l'étude du cas **infini** à celui de l'analyse d'un **ensemble fini d'hypothèses**

2. **Mesurer** à quel point, **pour n'importe quel échantillon** $\mathcal{S}$ de points étiquetés, on peut trouver une **hypothèse** de $\mathcal{H}$ **pouvant s'adapter** à $\mathcal{S}$

4ème étude statistique de l'induction

- La ***complexité de Rademacher***

  – Soit : $\mathcal{S} \; = \; \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} \; = \; \{z_1, \ldots, z_m\}$

  – Mesure de la corrélation entre les prédictions et les étiquettes :
  $$\frac{1}{m} \sum_{i=1}^{m} y_i \, h(\mathbf{x}_i)$$

  – L'hypothèse maximisant cette corrélation :
  $$h^* \; = \; \underset{h \in \mathcal{H}}{\text{ArgMax}} \, \frac{1}{m} \sum_{i=1}^{m} y_i \, h(\mathbf{x}_i)$$

  – Mesure caractérisant l'adéquation de $\mathcal{H}$ avec $\mathcal{S}$.

4ème étude statistique de l'induction

- La ***complexité de Rademacher*** (suite)
  - Supposons que les étiquettes soient choisies au hasard
    - Chaque y_i est remplacée par une variable aléatoire σ_i = -1 ou +1
  - On peut mesurer comment $\mathcal{H}$ peut s'ajuster à ce bruit par l'espérance :

$$R_{\mathcal{S}}(\mathcal{H}) \;=\; \mathbb{E}_{\sigma}\left[\underset{h \in \mathcal{H}}{\mathrm{Max}} \; \frac{1}{m} \sum_{i=1}^{m} \sigma_i \, h(\mathbf{x}_i)\right]$$

On en tire la borne :

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m\left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; R_{\mathcal{S}}(\mathcal{H}) \;+\; 3\sqrt{\frac{\log \frac{2}{\delta}}{2\,m}}\right] \;>\; 1 - \delta$$

4ème **étude statistique** de l'induction

- Mesure par la ***fonction de croissance***

  – **Critère** purement **combinatoire**, ne dépendant pas de la distribution $\mathbf{P}_{\mathcal{XY}}$

  – Nombre maximal de manière distinctes d'étiqueter *m* points de $\mathcal{X}$ en utilisant une hypothèse de $\mathcal{H}$

$$\forall m \in \mathbb{N}, \;\; \Pi_{\mathcal{H}}(m) \; = \max_{\{\mathbf{x}_1,\ldots,\mathbf{x}_m\} \subseteq \mathcal{X}} \Big| \big\{ \big(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)\big) \;:\; h \in \mathcal{H} \big\} \Big|$$

On en tire la borne :

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} \;+\; \sqrt{\frac{\log \frac{1}{\delta}}{2\, m}} \right] \;>\; 1 - \delta$$

4ème **étude statistique** de l'induction

- Mesure par la ***dimension de Vapnik-Chervonenkis***

  – **Critère** purement **combinatoire**, ne dépendant pas du nombre d'exemples

  – Taille du plus grand ensemble de points pouvant être étiquetés de n'importe quelle manière par les hypothèses tirées de H

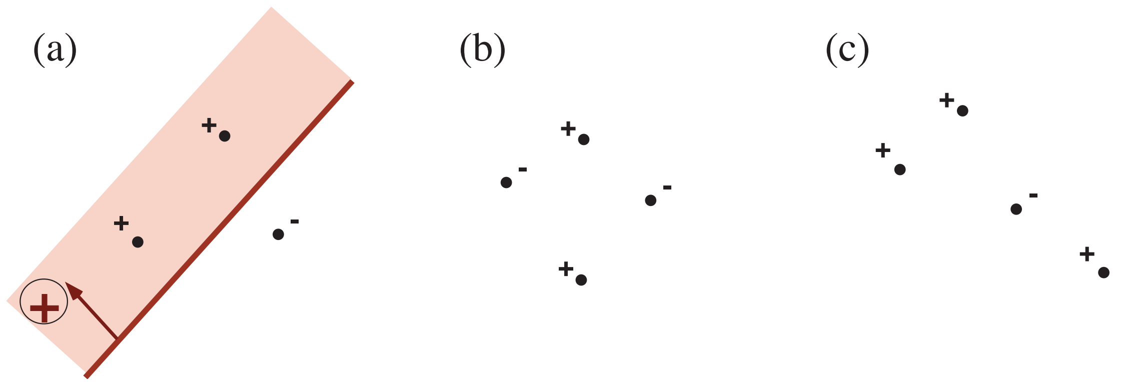
$$d_{VC}(\mathcal{H}) \;=\; \max\{m \;:\; \Pi_{\mathcal{H}}(m) \;=\; 2^m\}$$

On en tire la borne :

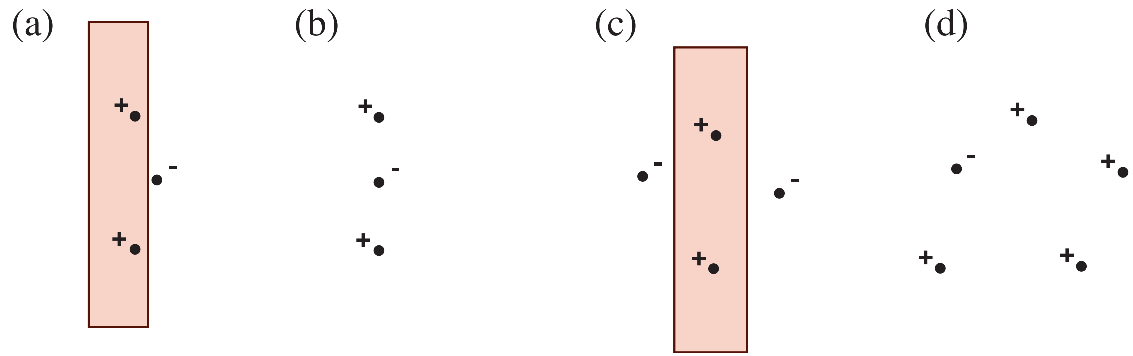
$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \sqrt{\frac{8\, d_{VC}(\mathcal{H}) \log \frac{2\, e\, m}{d_{VC}(\mathcal{H})} + 8 \log \frac{4}{\delta}}{m}} \;\right] \;>\; 1 - \delta$$

VC dim : illustration

- d_{VC}(séparateurs linéraires) = ?



- d_{VC}(rectangles) = ?



Étude statistique de l'induction

1. Ces mesures de capacité **ne dépendent pas de la dimension** de $\mathcal{X}$!!

2. La complexité de Rademacher **de l'enveloppe convexe** d'espaces $\mathcal{H}$ n'est **pas plus grande** que celle de $\mathcal{H}$!

   – Intéressant pour les méthodes d'ensemble et les méthodes collaboratives

L'analyse « PAC learning »

- On arrive à :

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad \mathbf{P}^m \left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \overbrace{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{m}}^{\varepsilon} \right] \;>\; 1 - \delta$$

Le principe de minimisation du risque empirique

n'est sain que si il y a des **contraintes sur l'espace des hypothèses**

L’analyse « PAC learning »

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad \mathbf{P}^m \left[R_{\text{Réel}}(h) \; \leq \; \underbrace{R_{\text{Emp}}(h) \; + \; \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m}}_{\text{Risque régularisé}} \right] \; > \; 1 - \delta$$

- *Nouveau critère inductif* :
  - Le **risque empirique régularisé**
    1. Satisfaire les contraintes posées par les **exemples**
    2. Choisir le meilleur **espace d’hypothèses** (capacité de *H*)

L'induction supervisée

en trois questions

Trois ingrédients essentiels

1. Choix de ***l'espace des hypothèses*** $\mathcal{H}$
   - Contrôler sa « capacité »
2. Choix du ***critère à optimiser*** $R(h)$
   - Risque empirique régularisé
3. Choix de la ***méthode d'exploration*** de $\mathcal{H}$
   - Plus facile si $R(.)$ convexe

Nouveautés séduisantes

- **Algorithme d'apprentissage**
  - Générique : ***minimisation du risque empirique régularisé***
  - Apprentissage = optimisation
- **Faible a priori** sur le monde
  - Suppose données (et questions) **i.i.d.**
  - $f \in H$ ou $f \notin H$
  - **Valable dans le pire cas** : contre toute distribution cible
- **Bornes** en généralisation
  - Formalisation mathématique supportant le bien-fondé

Un paradigme triomphant

Apprentissage = choix de normes + optimisation

(~ 1995 - ~20??)

Nouvelle perspective

- **Poser un problème d'apprentissage**, c'est :

1. L'exprimer sous forme d'**un critère inductif** à optimiser
   - **Risque empirique**
     - avec une **fonction d'erreur** adéquate
   - Un **terme de régularisation**
     - exprimant les contraintes
     - et connaissances a priori
     - Si possible conduisant à problème convexe

$$h^\star \; = \; \underset{h \in \mathcal{H}}{\mathrm{ArgMin}} \left[R_{\mathrm{Emp}}(h) \; + \lambda \; \mathrm{reg}(h) \right]$$

2. Trouver un **algorithme d'optimisation** adapté

Recherches actuelles : démarche générale

- Un **critère inductif** approprié

$$h^\star \;=\; \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[R_{\text{Emp}}(h) \;+ \lambda \, \text{reg}(h) \right]$$

$$h^\star \;=\; \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[\sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i) \;+ \lambda \, \text{reg}(h) \right]$$

- Éventuellement une ré-expression pour **faciliter l'optimisation**
  - Convexité
    - E.g. Fonction de perte surrogée

« Traduction » : sélection de descripteurs

- Recherche d'**hypothèse linéaire** parcimonieuse

$$h^\star \;=\; \underset{h \in \mathcal{H}}{\operatorname{ArgMin}} \left[\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i) \;+ \lambda \, \operatorname{reg}(h) \right]$$

$$\boxed{h^\star \;=\; \underset{h \in \mathcal{H}}{\operatorname{ArgMin}} \left[\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i) \;+ \lambda \, ||h||_1 \right]}$$

Norme l_1 : $\quad ||\mathbf{w}||_1 \;=\; \sum_{j=1}^{p} |w_j|$

- Méthodes de type LASSO

« Traduction » : classification semi-supervisée

- *l* données **étiquetées**, *u* données **non étiquetées**

$$\mathcal{S} \;=\; \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}\}$$

$$\mathbf{h} = \big[h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_{l+u})\big]$$

*Mesure de **régularité sur les données***

$$\mathbf{h}^\top \mathcal{L}\, \mathbf{h} \;=\; \frac{1}{2} \sum_{i,j=1}^{l+u} W_{ij} \left(h(\mathbf{x}_i) - h(\mathbf{x}_i)\right)^2$$

$$h^\star \;=\; \underset{h \in \mathcal{H}}{\text{Argmin}} \left\{ \frac{1}{l} \sum_{i+1}^{l} \left(y_i - h(\mathbf{x})\right)^2 + \lambda_1\, ||h||_2 + \lambda_2\, \mathbf{h}^\top \mathcal{L}\, \mathbf{h} \right\}$$

« Traduction » : apprentissage multi-tâches

- T tâches de classification binaire définies sur $\mathcal{X} \times \mathcal{Y}$

$$\mathcal{S} = \{\{(\mathbf{x}_{11}, y_{11}), (\mathbf{x}_{21}, y_{21}), \ldots, (\mathbf{x}_{m1}, y_{m1})\}, \ldots, \{(\mathbf{x}_{1T}, y_{1T}), (\mathbf{x}_{2T}, y_{2T}), \ldots, (\mathbf{x}_{mT}, y_{mT})\}\}$$

$$h_j(\mathbf{x}) = \mathbf{w}_j \cdot \mathbf{x}$$ Hypothèses linéaires

Partage entre tâches $$\mathbf{w}_j = \mathbf{w}_0 + \mathbf{v}_j$$

$$h_1^\star, \ldots, h_T^\star = \underset{\mathbf{w}_0, \mathbf{v}_j, \xi_{ij}}{\mathrm{Argmin}} \left\{ \sum_{j=1}^{T} \sum_{i=1}^{m} \xi_{ij} + \frac{\lambda_1}{T} \sum_{j=1}^{T} ||\mathbf{v}_j||^2 + \lambda_2 \, ||\mathbf{w}_0||^2 \right\}$$

3.3 du chapitre 3. Ainsi, étant donnés un échantillon source étiqueté $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ constitué de m exemples *i.i.d.* selon P_S et un échantillon cible non étiqueté $T = \{(\mathbf{x}_i^t)\}_{i=1}^m$ composé de m exemples *i.i.d.* selon D_T, en posant $S_u = \{\mathbf{x}_i^s\}_{i=1}^m$ l'échantillon S privé de ses étiquettes, on veut minimiser :

$$\min_{\mathbf{w}} \; c\,m\, \mathrm{R}_S(G_{\rho_\mathbf{w}}) + a\,m\, \mathrm{dis}_{\rho_{\mathbf{w}}'}(S_u, T_u) + \mathrm{KL}(\rho_\mathbf{w} \| \pi_0), \tag{7.5}$$

où $\mathrm{dis}_{\rho_{\mathbf{w}}'}(S_u, T_u) = \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathrm{R}_{S_u}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathrm{R}_{T_u}(h,h') \right|$ est le désaccord empirique entre S_u et T_u spécialisé à une distribution $\rho_\mathbf{w}$ sur l'espace $\mathcal{H}$ des classifieurs linéaires considéré. Les réels $a > 0$ et $c > 0$ sont des hyperparamètres de l'algorithme. Notons que les constantes A et C du théorème 7.7 peuvent être retrouvées à partir de n'importe quelle valeur de a et c. Étant donnée la fonction $\ell_{\mathrm{dis}}(x) = 2\,\ell_{\mathrm{Erf}}(x)\,\ell_{\mathrm{Erf}}(-x)$ (illustrée sur la figure 7.1) , pour toute distribution D sur X, on a :

$$\begin{aligned}
\mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathrm{R}_D(h,h') &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathbf{I}\left[h(\mathbf{x}) = 1\right] \mathbf{I}\left[h'(\mathbf{x}) = -1\right] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{h\sim\rho_\mathbf{w}} \mathbf{I}\left[h(\mathbf{x}) = 1\right] \mathop{\mathbf{E}}_{h'\sim\rho_\mathbf{w}} \mathbf{I}\left[h'(\mathbf{x}) = -1\right] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \ell_{\mathrm{Erf}}\left(\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right) \ell_{\mathrm{Erf}}\left(-\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right).
\end{aligned}$$

Ainsi, trouver la solution optimale de l'équation (7.5) revient à chercher le vecteur $\mathbf{w}$ qui minimise :

$$c \sum_{i=1}^{m} \ell_{\mathrm{Erf}}\left(y_i^s \frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) + a \left| \sum_{i=1}^{m} \left[\ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) - \ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^t\rangle}{\|\mathbf{x}_i^t\|}\right) \right] \right| + \frac{\|\mathbf{w}\|^2}{2}. \tag{7.6}$$

L'équation précédente est fortement non convexe. Afin de rendre sa résolution plus facilement contrôlable, nous remplaçons la fonction $\ell_{\mathrm{Erf}}(\cdot)$ par sa relaxation convexe $\ell_{\mathrm{Erf_{cvx}}}(\cdot)$ (comme pour PBGD3 et illustrée sur la figure 7.1). L'optimisation se réalise ensuite par une descente de gradient. Le gradient de l'équation 7.6 étant :

AgroParisTech

Industrie des bornes en généralisation

On peut étendre la démarche du PAC learning

Pour obtenir des bornes

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R(h) \;\leq\; \hat{R}(h) \;+\; \sqrt{\frac{\Omega(\text{satisfaction attentes})}{m}} \right] \;>\; 1 - \delta$$

- Si $\widehat{\text{err}(h)}$ = 0 (ou petite)
- Si attente sur le monde vérifiée (ou presque)
- Avec échantillon assez grand
- Alors err(*h*) < ε (en probabilité)

Bilan : l'empire des normes

- **Une démarche générique** et générale
  - Définition d'un **risque régularisé**
    - Traduisant des attentes sur les régularités d'intérêt
    - Assurant problème convexe
  - Algorithme d'**apprentissage** = algorithme d'**optimisation**

- Un **certificat d'excellence**
  - Bornes en généralisation, bien mathématiques

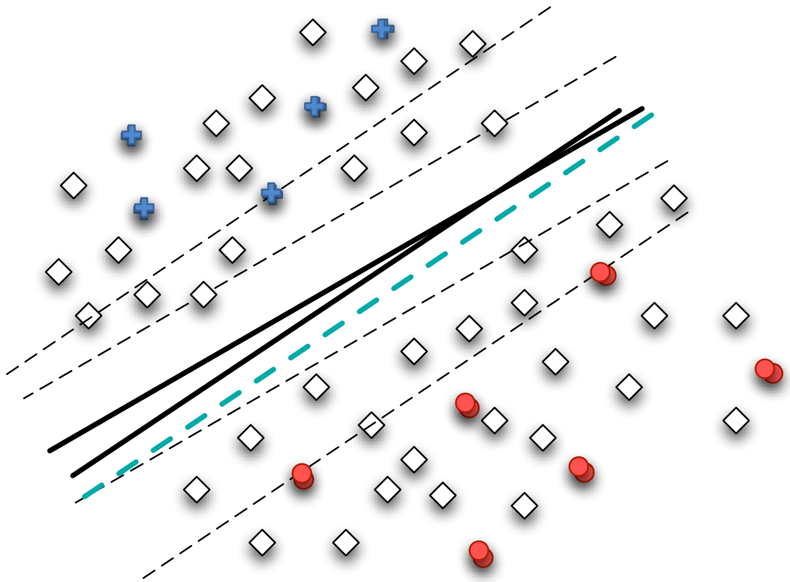
- Des **présupposés** supposés **modestes**
  - Et adaptés au « big data »

Données et questions **i.i.d.**

Variantes de

l'induction supervisée

L’apprentissage semi-supervisé



Une théorie PAC de l'apprentissage semi-supervisé

[M-F. Balcan & A. Blum (2006) « *An augmented PAC model for semi-supervised learning* » in Chapelle et al. (Eds) *Semi-supervised Learning*. MIT Press, 2006.]

The « luckiness framework »

Limites

- Apprentissage **passif** et **données et questions i.i.d.**
  - Agents situés : le monde n'est pas i.i.d.
- Requiert **beaucoup** d'exemples
  - Nous sommes beaucoup plus efficaces
  - « Producteurs de théories », théories que nous testons ensuite
- Pas adapté à la recherche de **causalités**
- Pas **intégré** avec un **raisonnement**

Les machines **apprenantes** ne sont pas des machines **pensantes**

Apprentissage en-ligne

kjkjkj

gfgfg

ghghgh

Trame

1. Diverses **illustrations** de l'induction

2. Le **no-free-lunch theorem**

3. Interprétation / **complétion** de percepts

4. Étude de l'**induction supervisée**

5. **Variantes** de l'induction supervisée

6. **Transfert, analogie, éducation** : quel principe inductif ?

Transfert, analogie, éducation

Quel principe inductif ?

Construire de nouveaux paradigmes : nouveaux et rigoureux

- Exemples non i.i.d.
- Et pas très nombreux

analogie ; transfert ; …

- Comment **fonder** une **nouvelle théorisation** adaptée ?
  - Quels présupposés ?
  - Quel critère de performance ?
  - Quels outils formels de prédiction et de preuve ?

Conclusion