Théorie Appr. Autom. et Appr. Symbolique

Février - Mars 2021

Théorie «*computationnelle*» de l'apprentissage automatique

Théorie «*computationnelle*» de l'apprentissage automatique

Introduction : Questions théoriques sur l'apprentissage

C'est la version «*machine learning*» de la théorie de la complexité.

- Combien d'exemples faut-il pour apprendre un concept ?

Introduction : Questions théoriques sur l'apprentissage

C'est la version «*machine learning*» de la théorie de la complexité.

- Combien d'exemples faut-il pour apprendre un concept ?
- Peut-on apprendre un concept en temps polynomial ?

Introduction : Questions théoriques sur l'apprentissage

C'est la version «*machine learning*» de la théorie de la complexité.

- Combien d'exemples faut-il pour apprendre un concept ?
- Peut-on apprendre un concept en temps polynomial ?
- Quelles familles de concepts sont apprenables en pratique ?

Introduction : Questions théoriques sur l'apprentissage

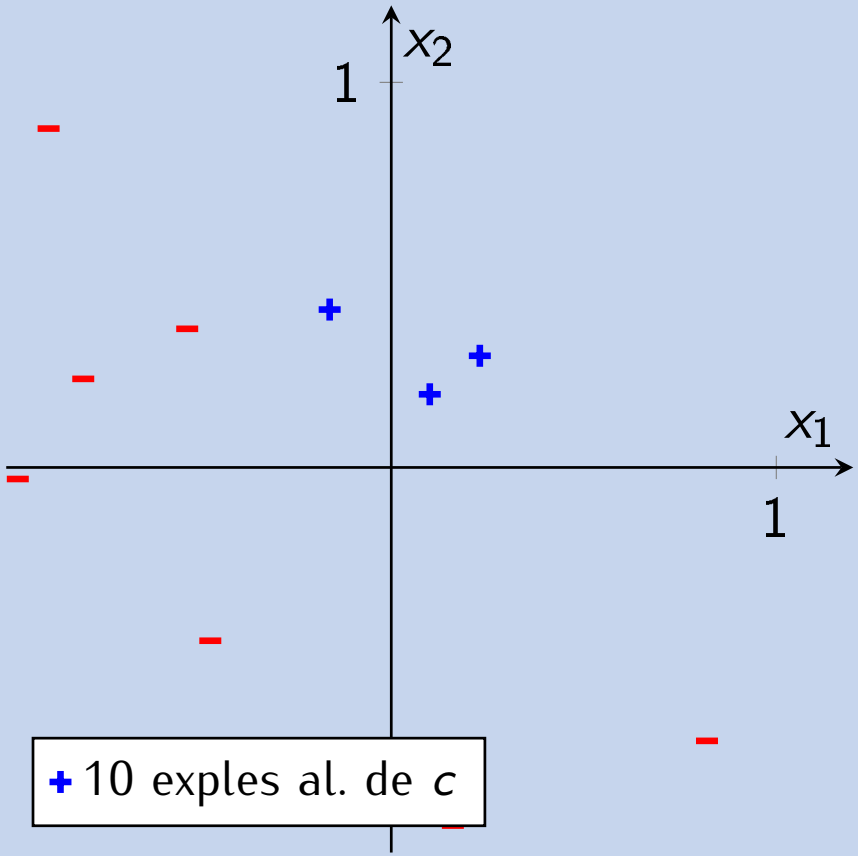
C'est la version «*machine learning*» de la théorie de la complexité.

- Combien d'exemples faut-il pour apprendre un concept ?
- Peut-on apprendre un concept en temps polynomial ?
- Quelles familles de concepts sont apprenables en pratique ?
- Combien d'erreurs risque-t-on de faire avant d'avoir bien appris ?

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

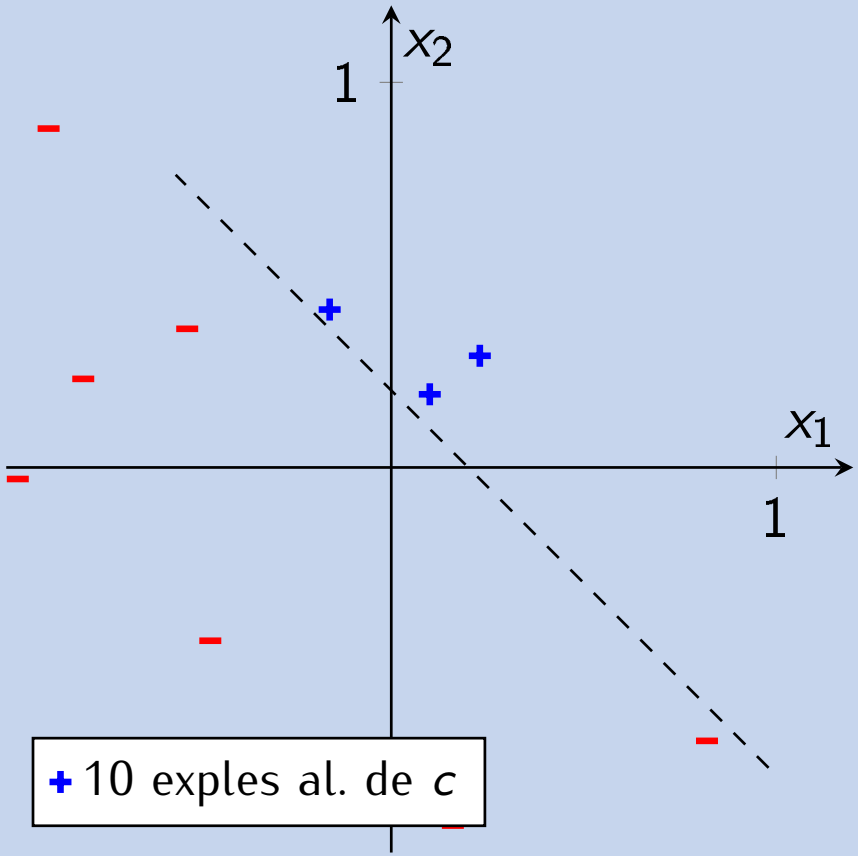
Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

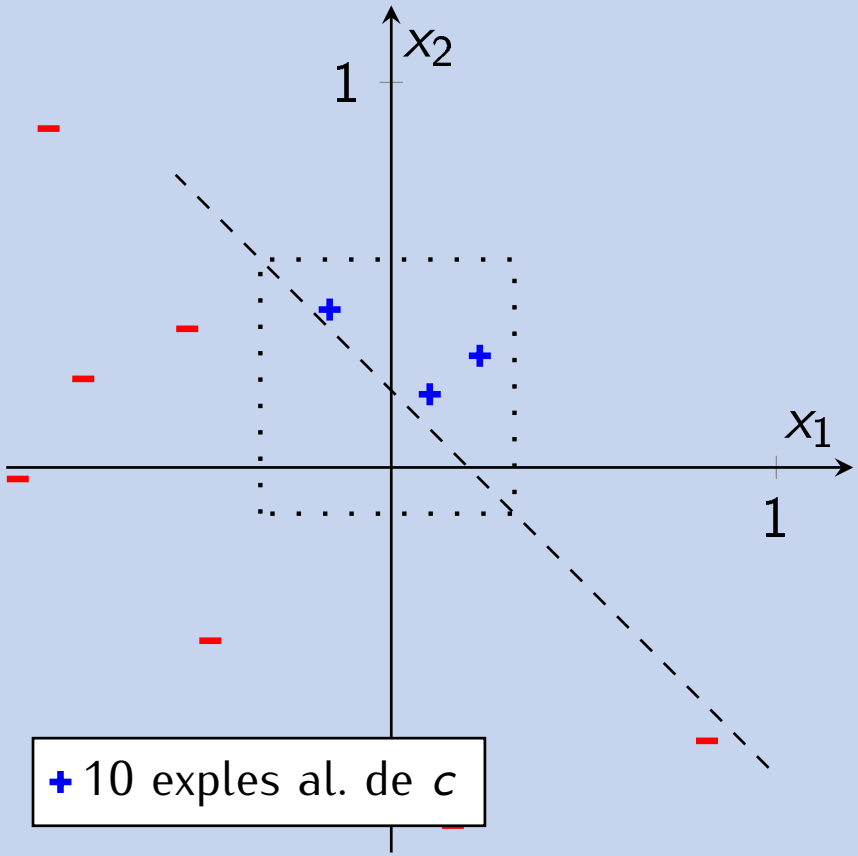
Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Introduction : Un exemple

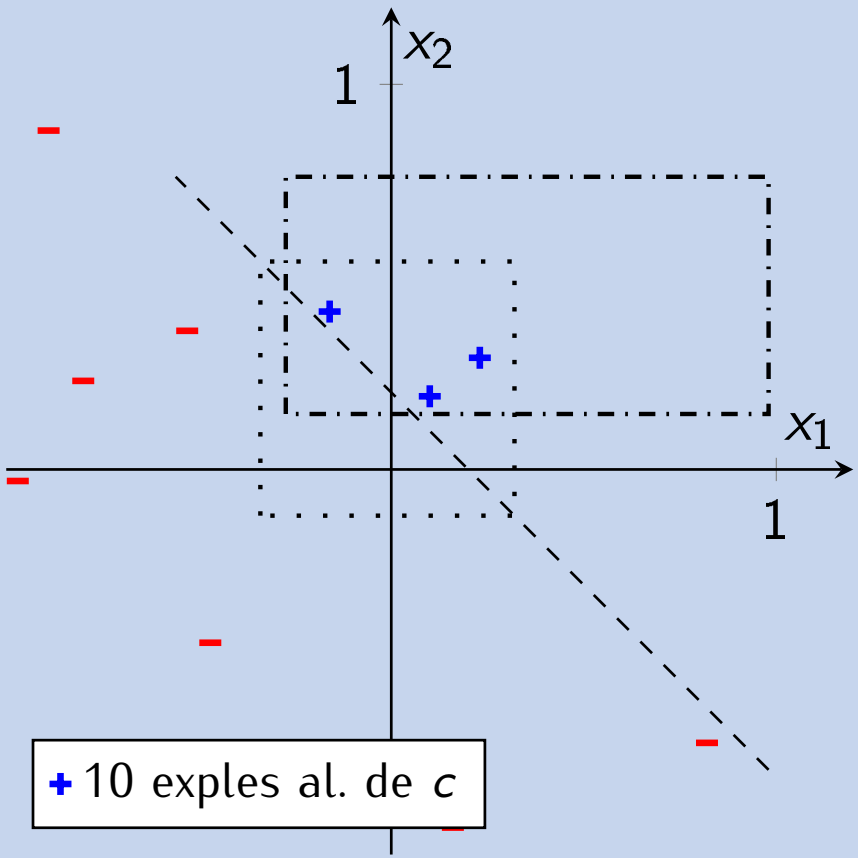
Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

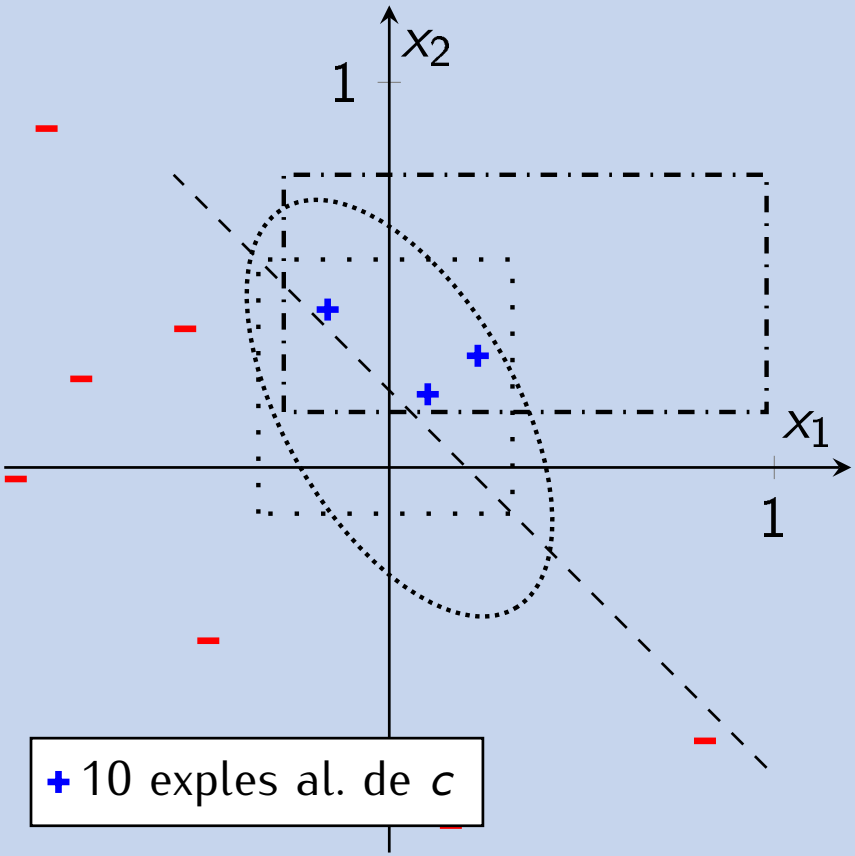
Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu

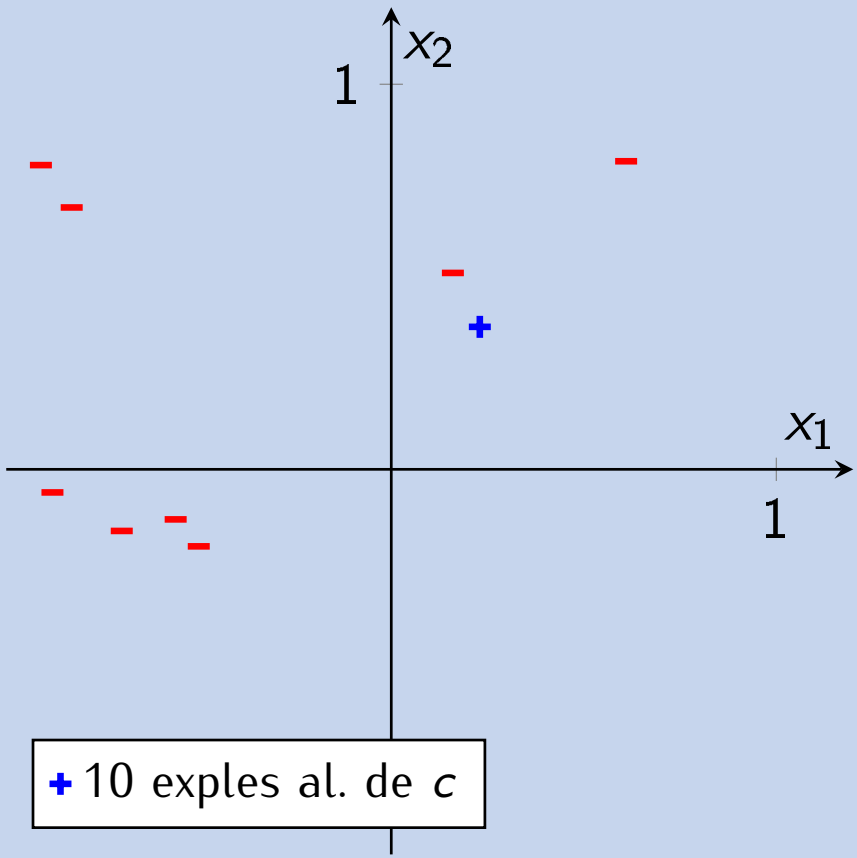


Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu

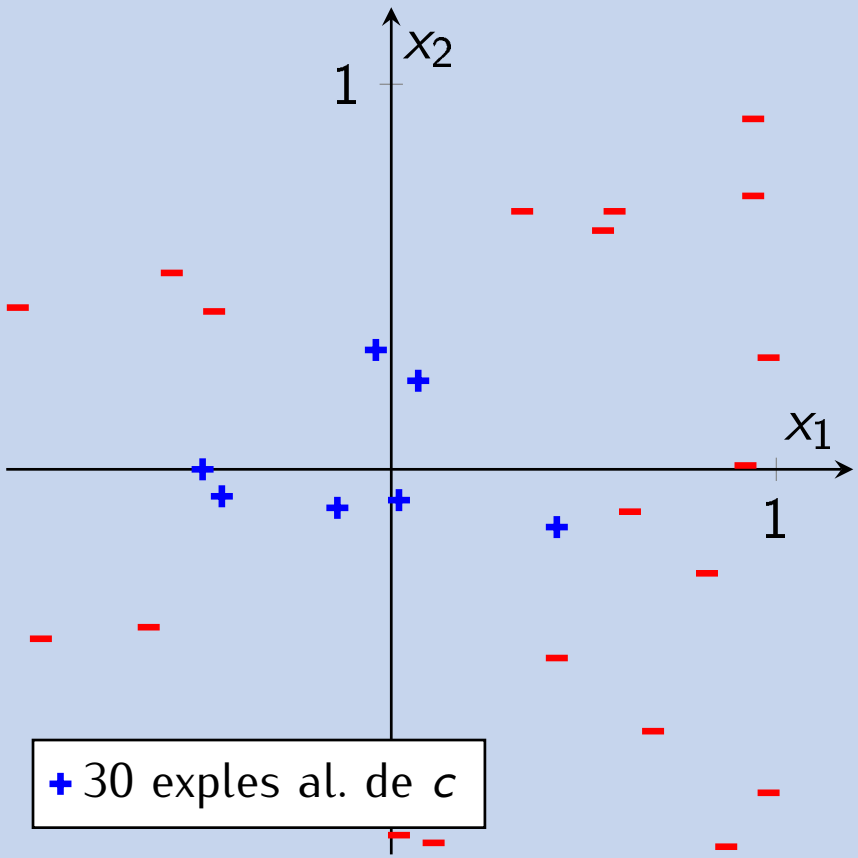
Tous ces exples générés
avec le m̂ concept inconnu c.



Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Tous ces exples générés
avec le m̂ concept inconnu c.

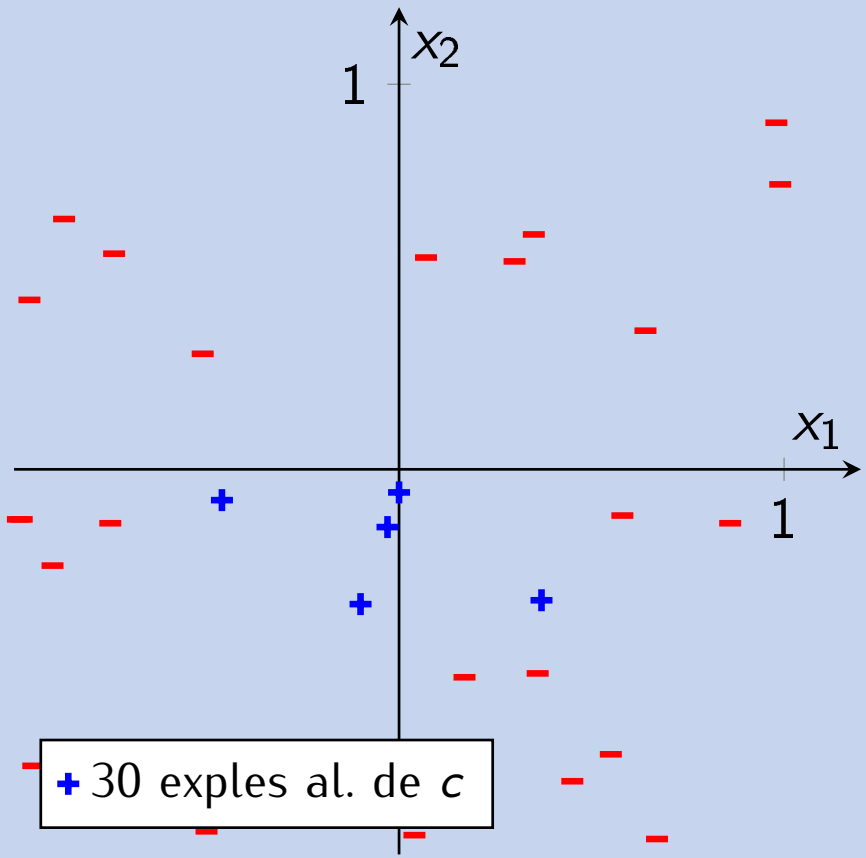
Intuitivement, on apprend mieux :

- avec plus d'exemples

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Tous ces exples générés
avec le m̂ concept inconnu c.

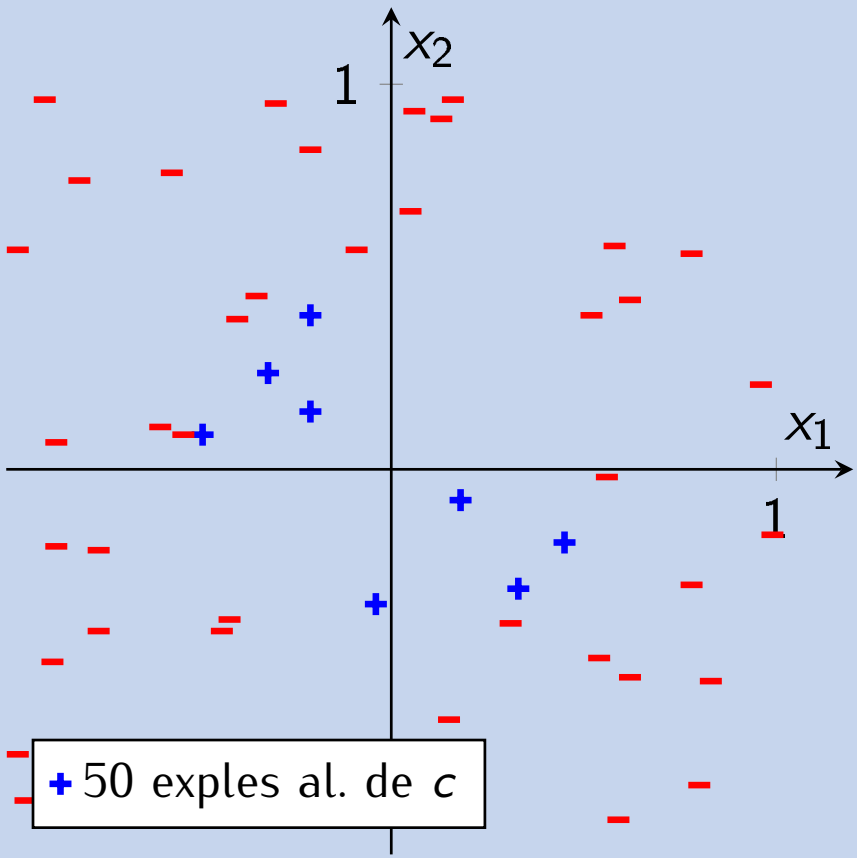
Intuitivement, on apprend mieux :

- avec plus d'exemples

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Tous ces exples générés
avec le m̂ concept inconnu c.

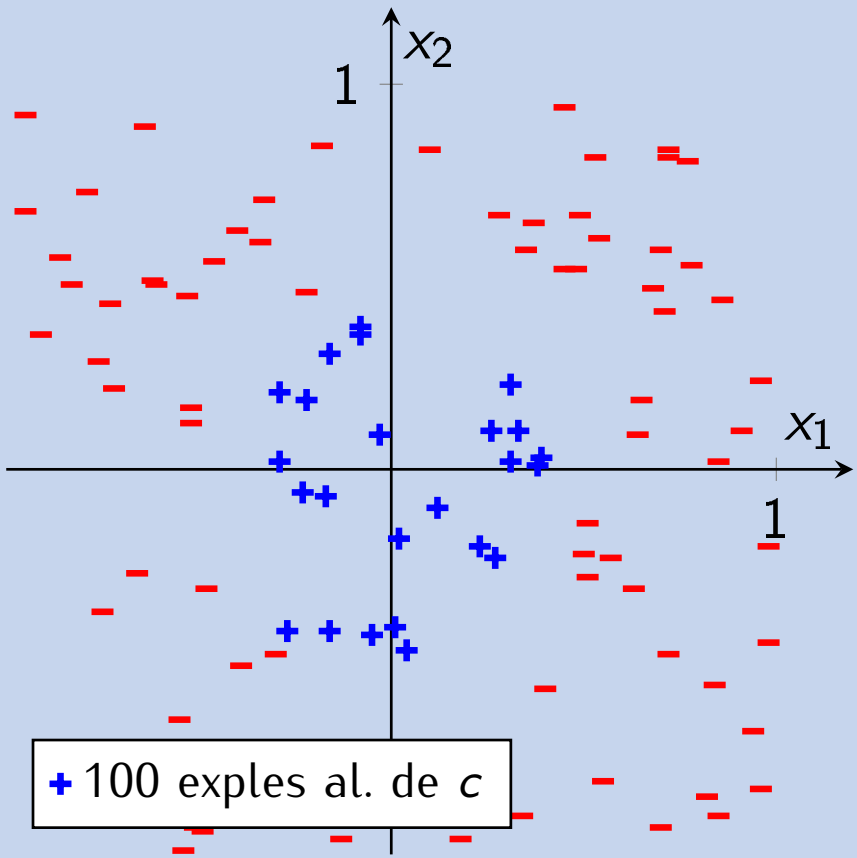
Intuitivement, on apprend mieux :

- avec plus d'exemples

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



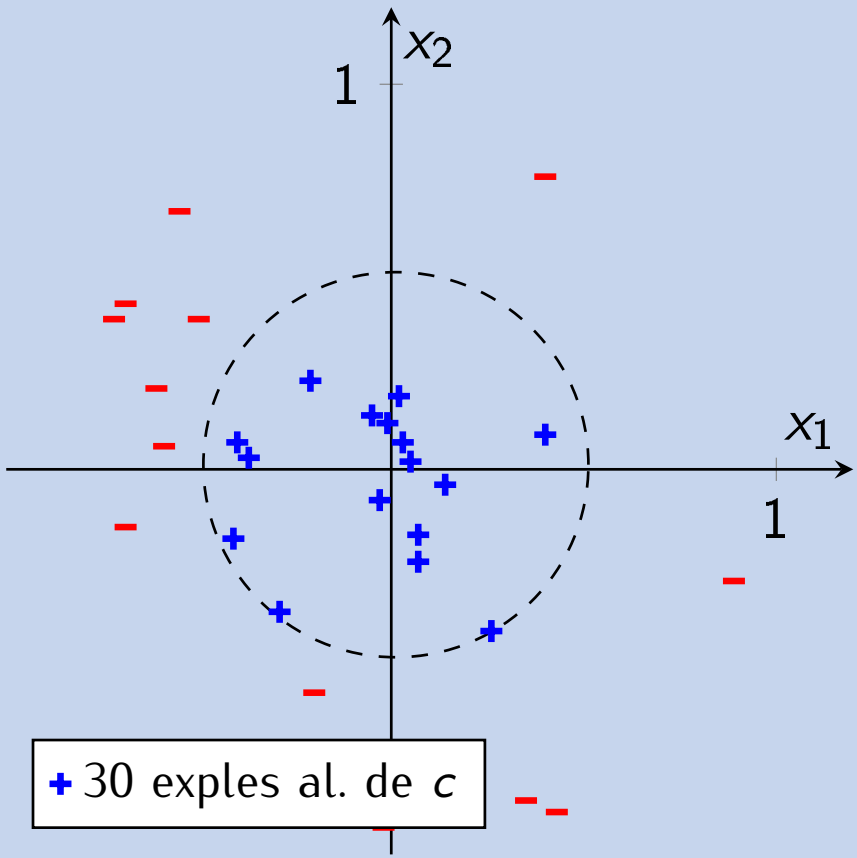
Tous ces exples générés
avec le m̂ concept inconnu c.

Intuitivement, on apprend mieux :

- avec plus d'exemples

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Tous ces exples générés
avec le m̂ concept inconnu c.

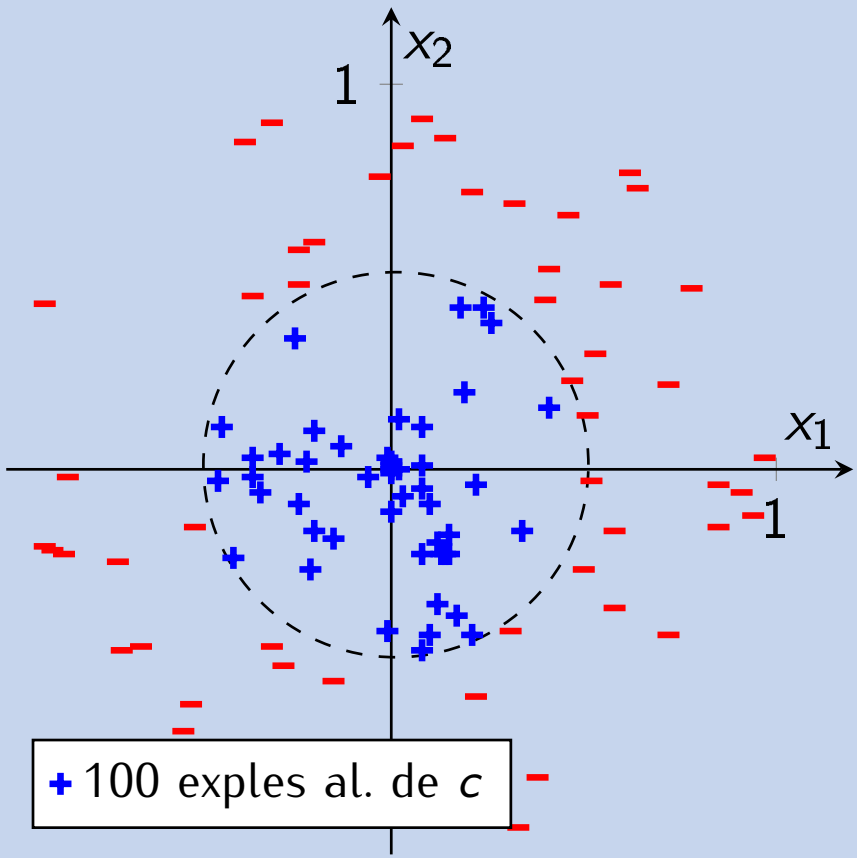
Intuitivement, on apprend mieux :

- avec plus d'exemples
- avec un espace d'hypothèses plus contraint, par exple :

$\mathcal{H}$ =cercles centrés en O

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Tous ces exples générés
avec le m̂ concept inconnu c.

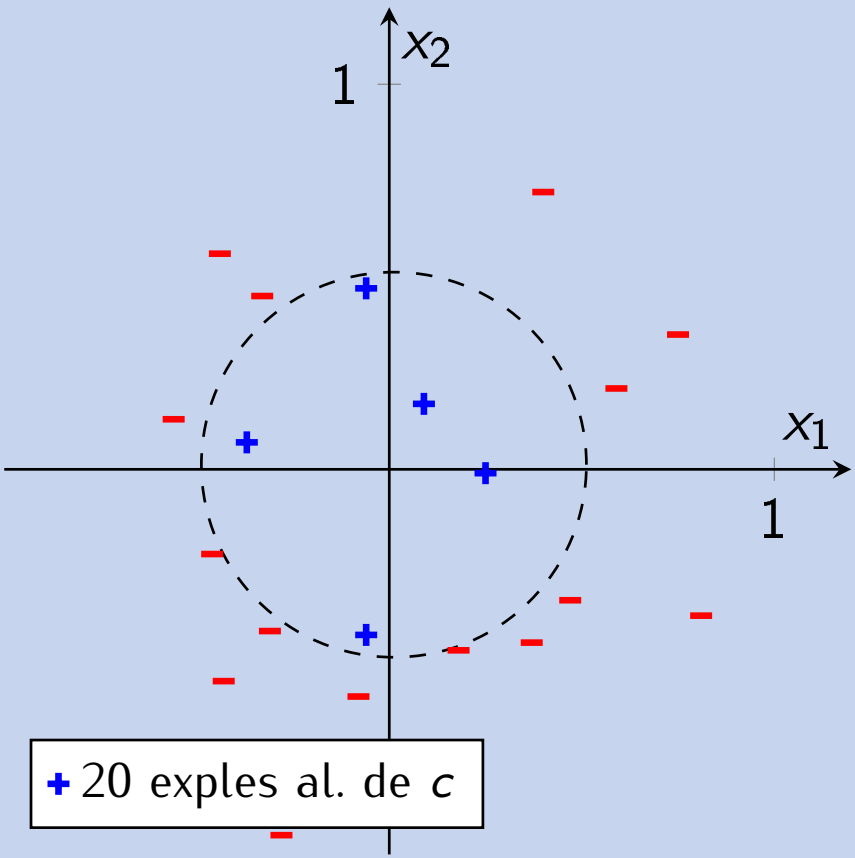
Intuitivement, on apprend mieux :

- avec plus d'exemples
- avec un espace d'hypothèses plus contraint, par exple :

$\mathcal{H}$ =cercles centrés en O

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



Tous ces exples générés
avec le m̂ concept inconnu c.

Intuitivement, on apprend mieux :

- avec plus d'exemples
- avec un espace d'hypothèses plus contraint, par exple :

$\mathcal{H}$ =cercles centrés en O

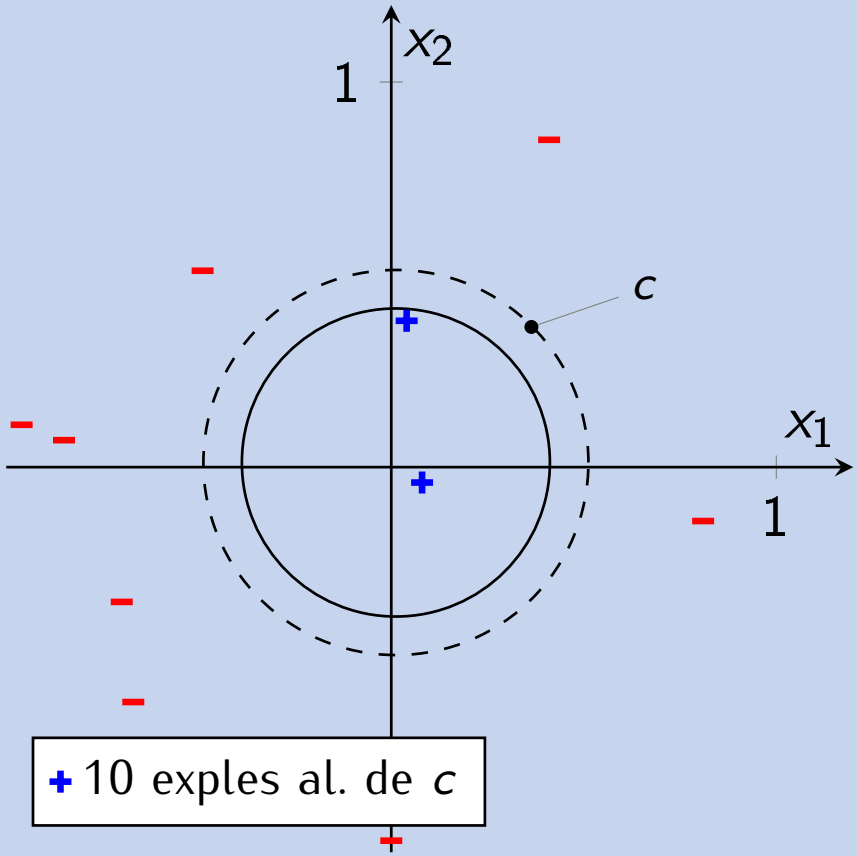
Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu

$\mathcal{H}$ =cercles centrés en O

S = ens. de m exples (x_1, x_2, y)



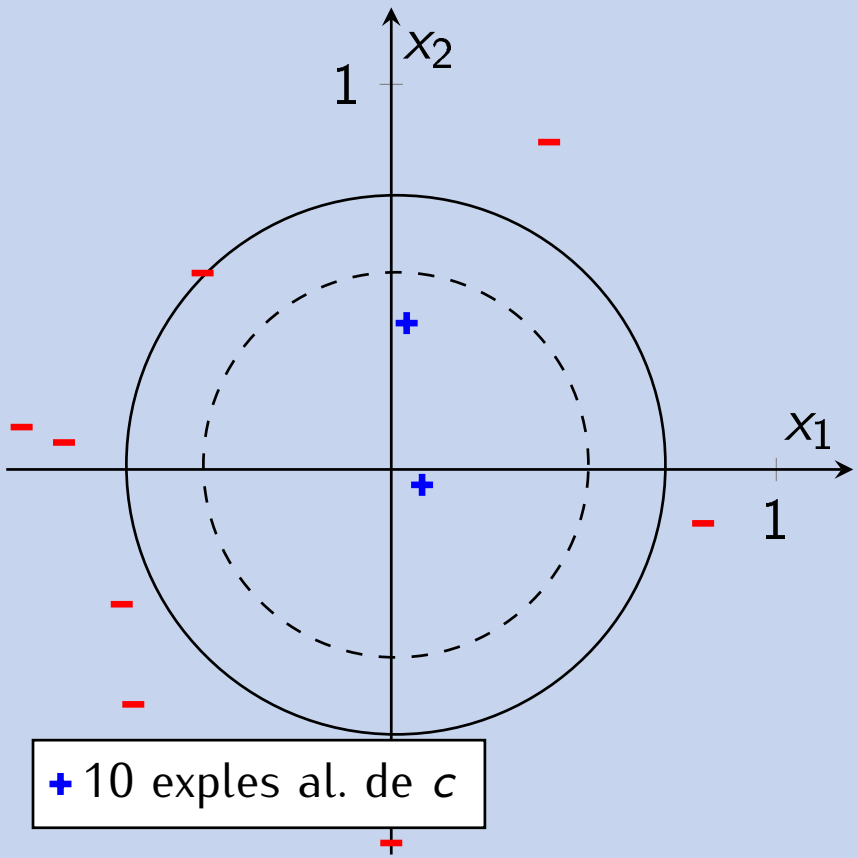
Algo. d'apprentissage :

- Le + petit cercle (centré en O) qui contient tous les +
  1. $r \leftarrow 0$
  2. for $(x_1, x_2, +1)$ in S :
     a. $d \leftarrow x_1^2 + x_2^2$
     b. if $d < r^2$: $r \leftarrow \sqrt{d}$

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



$\mathcal{H}$ =cercles centrés en O

S = ens. de m exples (x_1, x_2, y)

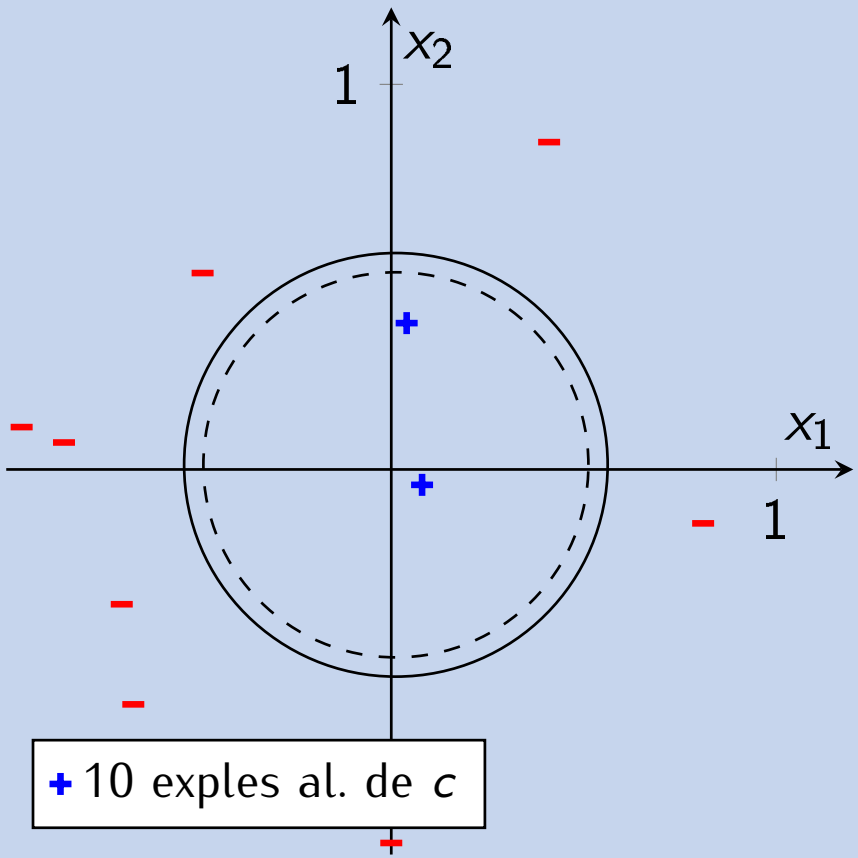
Algo. d'apprentissage :

- Le + petit cercle (centré en O) qui contient tous les +
- Le + ***grand*** cercle (tjs en O) qui ne contient aucun −

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. binaire $\mathcal{Y} = \{-1, +1\}$ Concept c inconnu



$\mathcal{H}$ =cercles centrés en O

S = ens. de m exples (x_1, x_2, y)

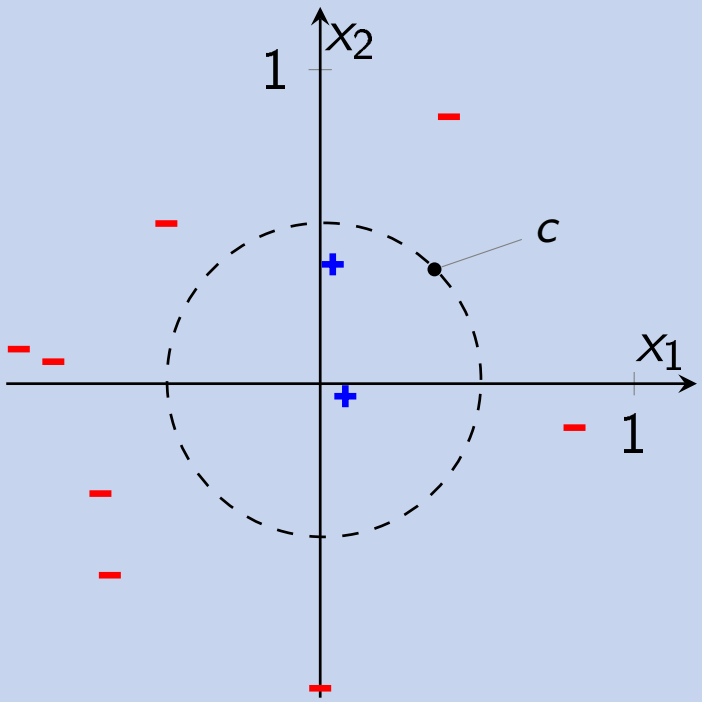
Algo. d'apprentissage :

- Le + petit cercle (centré en O) qui contient tous les +
- Le + ***grand*** cercle (tjs en O) qui ne contient aucun −
- au milieu de ces 2 cercles
- ...

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



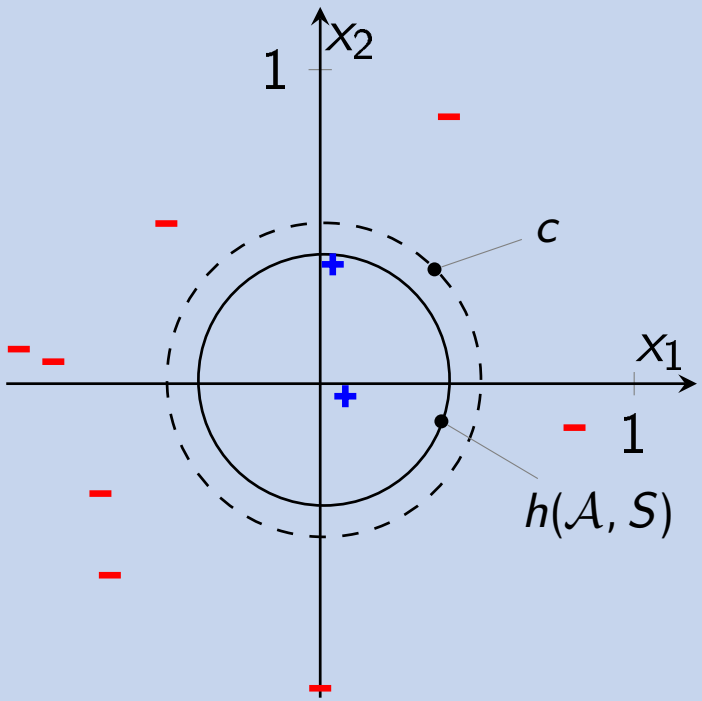
Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

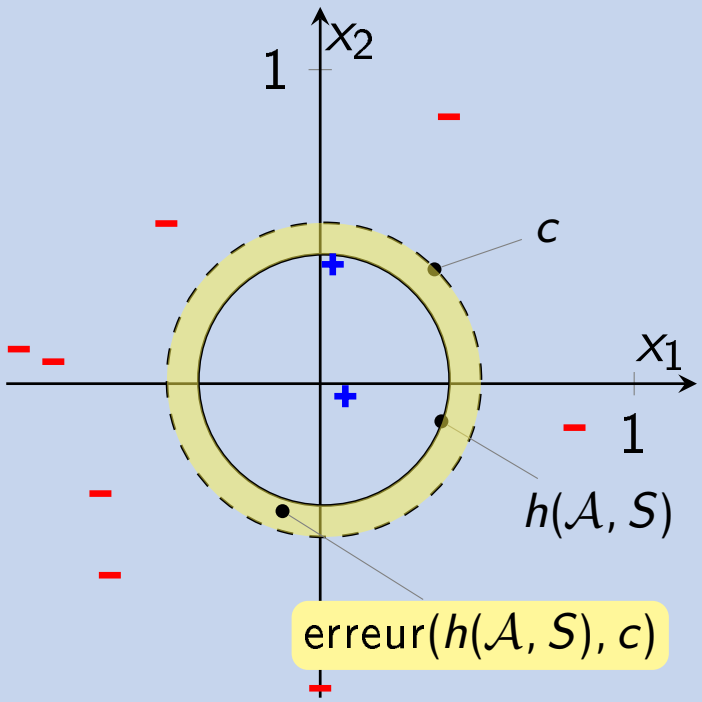
- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

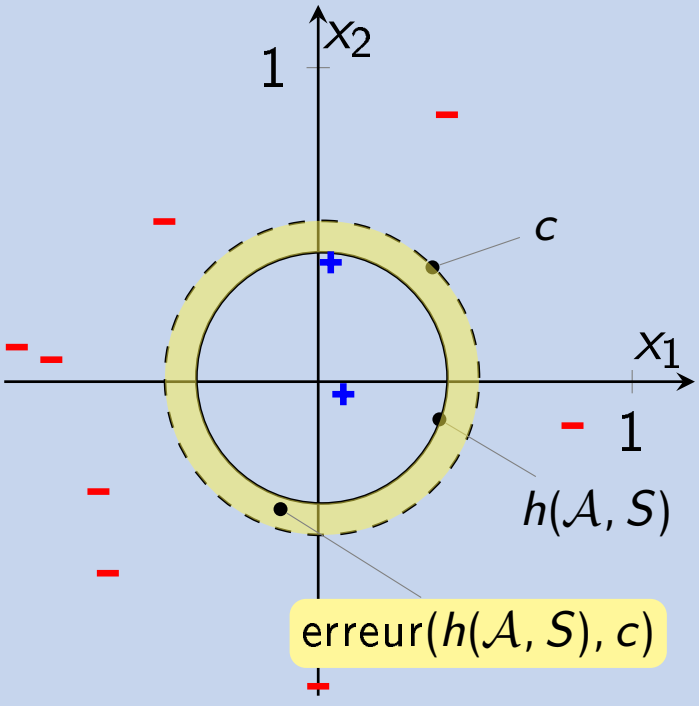
Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = proba. de se tromper sur une instance aléatoire

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

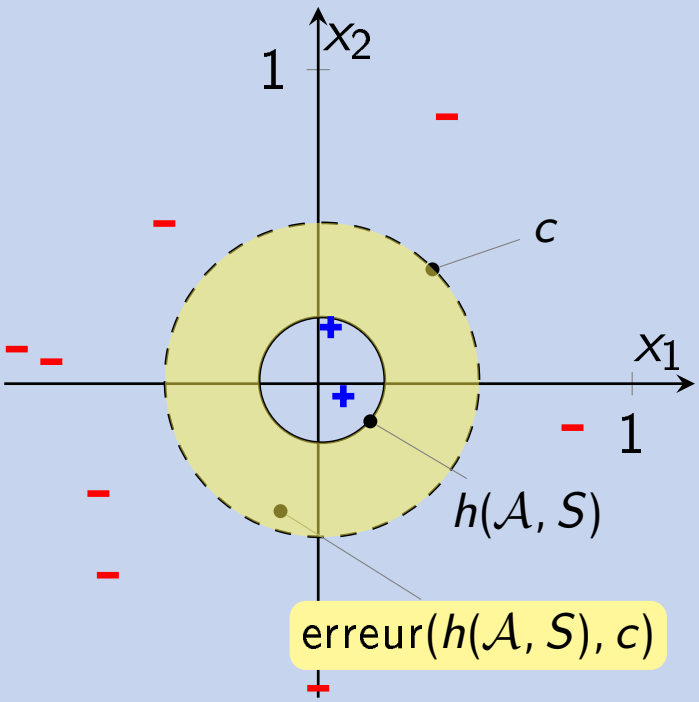
Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_{\mathcal{D}}(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
  - si x en dehors de cette zone alors $h(\mathcal{A}, S)(x) = c(x)$
  - $\mathcal{D}$ = distrib. de proba. sur $\mathcal{X}$ (souvent inconnue)

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

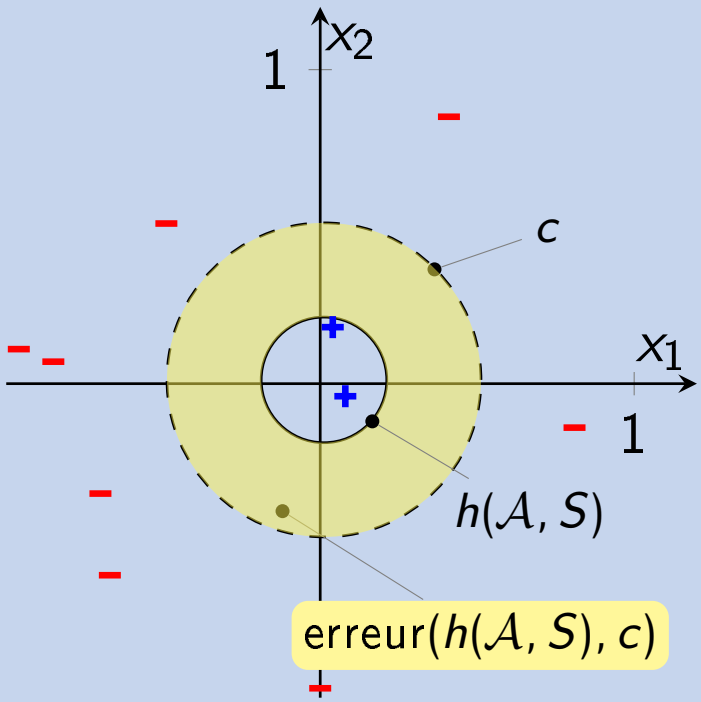
Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_{\mathcal{D}}(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_{\mathcal{D}}(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

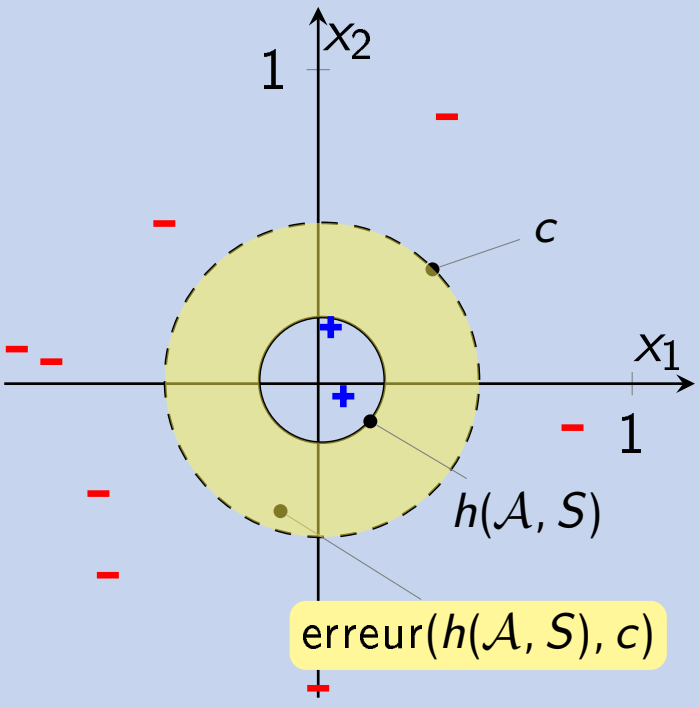
Pas de contrôle sur les exemples (apprentissage ***passif***)

$\Rightarrow$ Quelle est la probabilité d'avoir une erreur faible ?

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_{\mathcal{D}}(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

Pas de contrôle sur les exemples (apprentissage ***passif***)

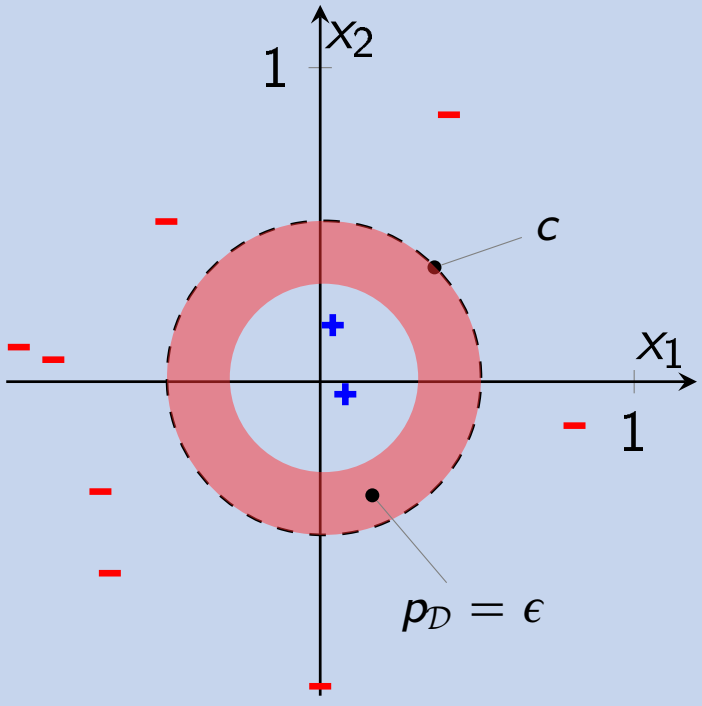
$\Rightarrow$ Quelle est la probabilité d'avoir une erreur faible ?

Ou encore : quelle est la proba. d'avoir un «***mauvais***» S ?

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_D(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

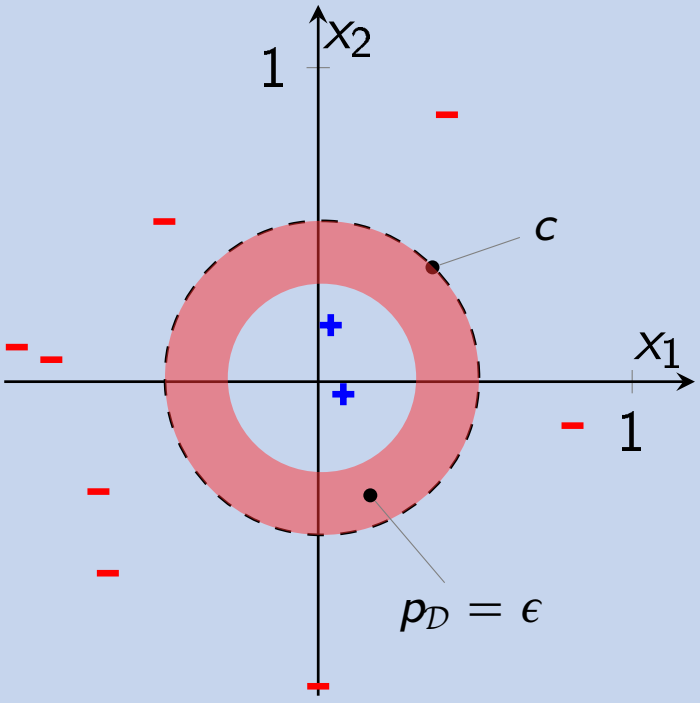
- $p_D(\text{erreur}(h(\mathcal{A}, S), c) > \epsilon) = p_D(\text{aucun exple de } S \in \text{ zone rouge})$

  où «*zone rouge*» = zone de proba. ϵ à l'intérieur de c

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Erreur *réelle* de $h(\mathcal{A})$:

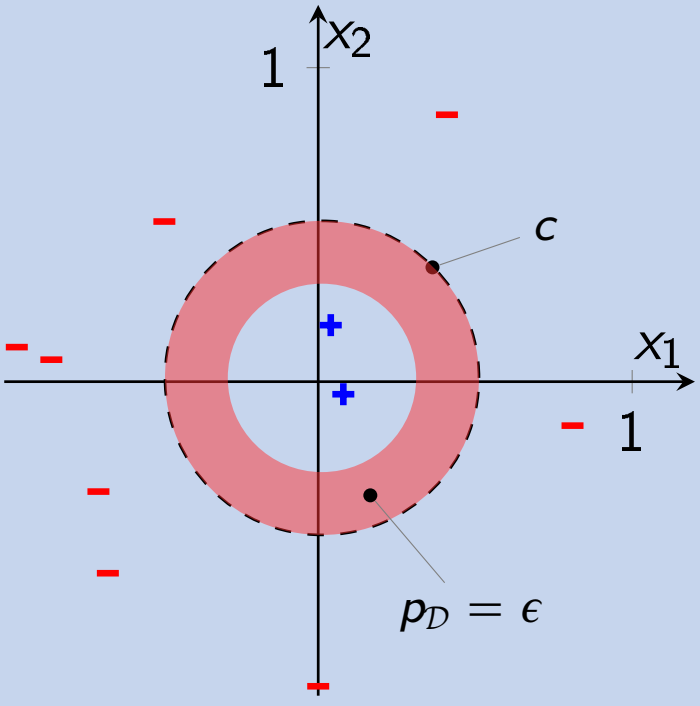
- erreur$(h(\mathcal{A}, S), c)$ = $p_D(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

- $p_D(\text{erreur}(h(\mathcal{A}, S), c) > \epsilon) = p_D(\text{aucun exple de } S \in \text{zone rouge})$
- pour chaque exemple (x, y) : $p_D(x \notin \text{zone rouge}) = 1 - \epsilon$

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_D(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

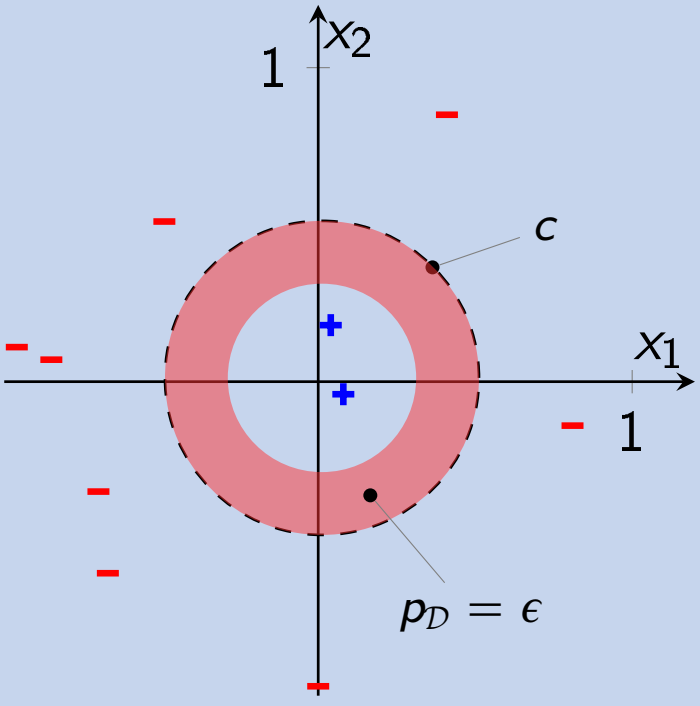
- $p_D(\text{erreur}(h(\mathcal{A}, S), c) > \epsilon) = p_D(\text{aucun exple de } S \in \text{zone rouge})$
- pour chaque exemple (x, y) : $p_D(x \notin \text{zone rouge}) = 1 - \epsilon$

$\Rightarrow$ pour m exemples : $p_D(S \cap \text{zone rouge} = \varnothing) = (1 - \epsilon)^m$
si exemples i.i.d. (indépendants identiquement distribués)

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_D(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

- $p_D(\text{erreur}(h(\mathcal{A}, S), c) > \epsilon) = p_D(\text{aucun exple de } S \in \text{ zone rouge})$
- pour chaque exemple (x, y) : $p_D(x \notin \text{zone rouge}) = 1 - \epsilon$

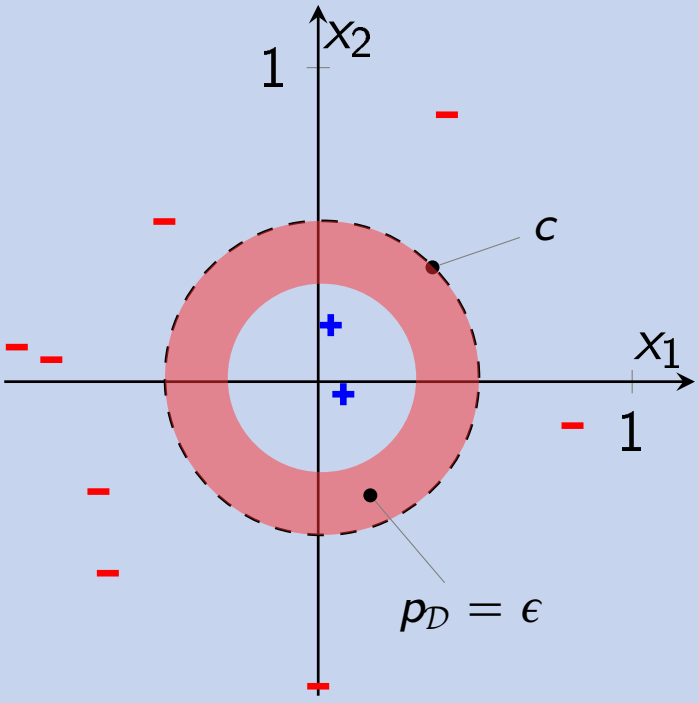
$\Rightarrow$ pour m exemples : $p_D(S \cap \text{zone rouge} = \varnothing) = (1 - \epsilon)^m$

$\Rightarrow$ $p_D(\text{erreur}(h(\mathcal{A}, S), c) > \epsilon) = (1 - \epsilon)^m$

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O



Concept inconnu $c \in \mathcal{H}$: rayon $r(c)$

S = ens. de m exemples (x_1, x_2, y)

Algo. d'apprentissage $\mathcal{A}$:

- Le + petit cercle (centré en O) qui contient tous les +

$\Rightarrow$ $h(\mathcal{A}, S)$ = cercle de rayon $r(\mathcal{A}, S)$

Erreur *réelle* de $h(\mathcal{A})$:

- erreur$(h(\mathcal{A}, S), c)$ = $p_D(x \in \mathbb{R}^2 \mid x$ entre les 2 cercles$)$
- dépend de S

- $p_D(\text{erreur}(h(\mathcal{A}, S), c) > \epsilon) = p_D(\text{aucun exple de } S \in \text{ zone rouge})$
- pour chaque exemple (x, y) : $p_D(x \notin \text{zone rouge}) = 1 - \epsilon$

$\Rightarrow$ pour m exemples : $p_D(S \cap \text{zone rouge} = \varnothing) = (1 - \epsilon)^m$

ou encore $\boxed{p_D(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) = 1 - (1 - \epsilon)^m}$

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O
Concept inconnu $c \in \mathcal{H}$
Ensemble S de m exemples Apprentissage passif
Algo. $\mathcal{A}$: + petit cercle contenant tout les $(x, +1)$ de S

$$p_{\mathcal{D}}(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) = 1 - (1 - \epsilon)^m$$

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O
Concept inconnu $c \in \mathcal{H}$
Ensemble S de m exemples Apprentissage passif
Algo. $\mathcal{A}$: + petit cercle contenant tout les $(x, +1)$ de S

$$p_{\mathcal{D}}(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) = 1 - (1 - \epsilon)^m$$

Ce qu'on retient de cet exemple d'apprentissage **passif** :

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O
Concept inconnu $c \in \mathcal{H}$
Ensemble S de m exemples Apprentissage passif
Algo. $\mathcal{A}$: + petit cercle contenant tout les $(x, +1)$ de S

$$p_{\mathcal{D}}(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) = 1 - (1 - \epsilon)^m$$

Ce qu'on retient de cet exemple d'apprentissage **passif** :

- On n'a pas d'influence sur le tirage des exemples de S

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O
Concept inconnu $c \in \mathcal{H}$
Ensemble S de m exemples Apprentissage passif
Algo. $\mathcal{A}$: + petit cercle contenant tout les $(x, +1)$ de S

$$p_{\mathcal{D}}(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) = 1 - (1 - \epsilon)^m$$

Ce qu'on retient de cet exemple d'apprentissage **passif** :

- On n'a pas d'influence sur le tirage des exemples de S

$\Rightarrow$ On ne peut pas ***garantir*** qu'on aura une erreur réelle faible

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O
Concept inconnu $c \in \mathcal{H}$
Ensemble S de m exemples Apprentissage passif
Algo. $\mathcal{A}$: + petit cercle contenant tout les $(x, +1)$ de S

$$p_{\mathcal{D}}(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) = 1 - (1 - \epsilon)^m$$

Ce qu'on retient de cet exemple d'apprentissage **passif** :

- On n'a pas d'influence sur le tirage des exemples de S

$\Rightarrow$ On ne peut pas ***garantir*** qu'on aura une erreur réelle faible

- ***Mais*** on peut garantir qu'avec suffisamment d'exemples, on aura ***probablement*** une erreur faible :

Introduction : Un exemple

Example (Classif. binaire dans $\mathbb{R}^2$ – apprentissage passif)

Instance $\mathcal{X} = \mathbb{R}^2$ Classif. bin. $\mathcal{Y} = \{-1, +1\}$ $\mathcal{H}$ = cercles centrés en O
Concept inconnu $c \in \mathcal{H}$
Ensemble S de m exemples Apprentissage passif
Algo. $\mathcal{A}$: + petit cercle contenant tout les $(x, +1)$ de S

$$p_{\mathcal{D}}(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) = 1 - (1 - \epsilon)^m$$

Ce qu'on retient de cet exemple d'apprentissage **passif** :

- On n'a pas d'influence sur le tirage des exemples de S

$\Rightarrow$ On ne peut pas ***garantir*** qu'on aura une erreur réelle faible

- ***Mais*** on peut garantir qu'avec suffisamment d'exemples, on aura ***probablement*** une erreur faible :
  - quand $m \to +\infty$, alors $p_{\mathcal{D}}(\text{erreur}(h(\mathcal{A}, S), c) \leq \epsilon) \to 1$

Apprentissage PAC

On parle d'apprentissage Probablement Approximativement Correct :

Approximativement Correct on obtient une hypothèse de risque réel faible mais non nul ;

Probablement on obtient probablement une telle hypothèse, mais ça n'est pas certain.

Apprentissage PAC : Définitions

- Ensemble d'instances $\mathcal{X}$
- Deux labels $\mathcal{Y} = \{-1, +1\}$

Definition (Erreur réelle)

- Concept $c : \mathcal{X} \to \{-1, +1\}$
- Hypothèse $h : \mathcal{X} \to \{-1, +1\}$
- Distribution de proba. $\mathcal{D}$ sur $\mathcal{X}$

$$\Rightarrow \text{erreur}_{\mathcal{D}}(h, c) = p_{\mathcal{D}}(h(x) \neq c(x))$$

Apprentissage PAC : Définitions

- Ensemble d'instances $\mathcal{X}$
- Deux labels $\mathcal{Y} = \{-1, +1\}$

Definition (Erreur réelle)

- Concept $c : \mathcal{X} \to \{-1, +1\}$
- Hypothèse $h : \mathcal{X} \to \{-1, +1\}$
- Distribution de proba. $\mathcal{D}$ sur $\mathcal{X}$

$$\Rightarrow \text{erreur}_{\mathcal{D}}(h, c) = p_{\mathcal{D}}(h(x) \neq c(x))$$

Definition (Erreur empirique)

- Ensemble de m exemples $S \subseteq \mathcal{X} \times \{-1, +1\}$
- Hypothèse $h : \mathcal{X} \to \{-1, +1\}$

$$\Rightarrow \text{erreur}(h, S) = \frac{1}{m} |\{(x, y) \in S | h(x) \neq y\}|$$

Apprentissage PAC : Définitions

- Ensemble d'instances $\mathcal{X}$
- Deux labels $\mathcal{Y} = \{-1, +1\}$

Definition (Erreur réelle)

- Concept $c : \mathcal{X} \to \{-1, +1\}$
- Hypothèse $h : \mathcal{X} \to \{-1, +1\}$
- Distribution de proba. $\mathcal{D}$ sur $\mathcal{X}$

$$\Rightarrow \text{erreur}_{\mathcal{D}}(h, c) = p_{\mathcal{D}}(h(x) \neq c(x))$$

Definition (Erreur empirique)

- Ensemble de m exemples $S \subseteq \mathcal{X} \times \{-1, +1\}$
- Hypothèse $h : \mathcal{X} \to \{-1, +1\}$

$$\Rightarrow \text{erreur}(h, S) = \frac{1}{m} |\{(x, y) \in S | h(x) \neq y\}|$$

- En général, on ne peut pas calculer $\text{erreur}_{\mathcal{D}}(h, c)$
- On peut calculer $\text{erreur}(h, S)$ dès qu'on connaît h et S

Apprentissage PAC : Définitions

- Ensemble d'instances $\mathcal{X}$
- Deux labels $\mathcal{Y} = \{-1, +1\}$

Definition (Sample complexity)

- Ensemble d'hypothèses $\mathcal{H} : \mathcal{X} \to \mathcal{Y} = \{-1, +1\}$
- Algo. d'apprentissage $\mathcal{A}$: pour ens. d'exples S, retourne $h(\mathcal{A}, S)$

$\Rightarrow$ la ***complexité en nombre d'exemples*** de $\mathcal{A}$ =
plus petite fonction $m :]0,1[^2 \to \mathbb{N}$ telle que

- pour tout concept $c \in \mathcal{H}$,
- pour tous $\epsilon, \delta \in]0,1[$
- pour toute distrib. de proba. $\mathcal{D}$ sur $\mathcal{X}$

$$p_{\mathcal{D}}(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq 1 - \delta$$

dès que

S ens. d'exples tirés selon $\mathcal{D}$ avec $|S| \geq m(\epsilon, \delta)$.

(Si une telle fonction m n'existe pas, on dit que la complexité en nombre d'exemples de $\mathcal{A}$ est infinie.)

Apprentissage PAC : Définitions

- Ensemble d'instances $\mathcal{X}$
- Deux labels $\mathcal{Y} = \{-1, +1\}$

Definition (PAC apprenable)

- Ensemble d'hypothèses $\mathcal{H} : \mathcal{X} \to \mathcal{Y} = \{-1, +1\}$

$\Rightarrow$ $\mathcal{H}$ est ***PAC apprenable*** s'il existe un algorithme $\mathcal{A}$ d'apprentissage pour $\mathcal{H}$ qui a une complexité en nombre d'exemples finie.

Apprentissage PAC : Ensemble d'hypothèse est fini

Theorem

Si

- *l'ensemble des hypothèses $\mathcal{H}$ est fini ;*
- *les exemples de S sont i.i.d. non bruités ;*
- *le concept cible est dans $\mathcal{H}$;*
- *$\mathcal{A}$ retourne une hypothèse $h(\mathcal{A}, S)$ d'erreur empirique nulle ;*

alors :

$$p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq 1 - \delta \quad \text{dès que} \quad |S| \geq \frac{1}{\epsilon}(\ln \frac{1}{\delta} + \ln |\mathcal{H}|)$$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. Si $\forall h \in \mathcal{H}(c, \epsilon)$: erreur$(h, S) > 0$, alors $h(\mathcal{A}, S) \notin \mathcal{H}(c, \epsilon)$
   car on suppose $\mathcal{A}$ tel que erreur$(h(\mathcal{A}, S), S) = 0$ (erreur emp. nulle)

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. Si $\forall h \in \mathcal{H}(c, \epsilon)$: erreur$(h, S) > 0$, alors $h(\mathcal{A}, S) \notin \mathcal{H}(c, \epsilon)$
   et donc erreur$_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon$.

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. Si $\forall h \in \mathcal{H}(c, \epsilon)$: erreur$(h, S) > 0$, alors $h(\mathcal{A}, S) \notin \mathcal{H}(c, \epsilon)$
   et donc erreur$_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon$.

   Donc
   $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$

□

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$

□

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c, \epsilon)} p(\forall x \in S, c(x) = h(x))$
   car $p(A \vee B) \leq p(A) + p(B)$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c, \epsilon)} p(\forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c, \epsilon)} p(c(x) = h(x))^m \quad$ où $m = |S|$
   car les x sont i.i.d.

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c,\epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} p(\forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} p(c(x) = h(x))^m \quad \text{où } m = |S|$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} (1-\epsilon)^m$
   car $h \in \mathcal{H}(c,\epsilon) \Rightarrow p(c(x) = h(x)) \leq 1 - \epsilon$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} p(\forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} p(c(x) = h(x))^m \quad \text{où } m = |S|$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} (1 - \epsilon)^m$
   $\leq \sum_{h \in \mathcal{H}} (1 - \epsilon)^m$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c,\epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} p(\forall x \in S, c(x) = h(x))$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} p(c(x) = h(x))^m \quad \text{où } m = |S|$
   $\leq \sum_{h \in \mathcal{H}(c,\epsilon)} (1-\epsilon)^m$
   $\leq \sum_{h \in \mathcal{H}} (1-\epsilon)^m$
   $\leq |\mathcal{H}|(1-\epsilon)^m$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x)) \leq |\mathcal{H}|(1 - \epsilon)^m$
5. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x)) \geq 1 - |\mathcal{H}|(1 - \epsilon)^m$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c, \epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c, \epsilon) : \forall x \in S, c(x) = h(x)) \leq |\mathcal{H}|(1 - \epsilon)^m$
5. $p(\forall h \in \mathcal{H}(c, \epsilon) : \exists x \in S, c(x) \neq h(x)) \geq 1 - |\mathcal{H}|(1 - \epsilon)^m$
6. On veut $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq 1 - \delta$
   on sait maintenant que $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A}, S), c) \leq \epsilon) \geq 1 - |\mathcal{H}|(1 - \epsilon)^m$
   il suffit donc d'avoir $\delta \geq |\mathcal{H}|(1 - \epsilon)^m$

□

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c,\epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x)) \leq |\mathcal{H}|(1-\epsilon)^m$
5. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x)) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
6. $\boxed{\text{On veut } p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1-\delta}$
   on sait maintenant que $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
   $\boxed{\text{il suffit donc d'avoir } \delta \geq |\mathcal{H}|(1-\epsilon)^m}$
7. $\delta \geq |\mathcal{H}|(1-\epsilon)^m \quad \Leftrightarrow \ln \delta \geq \ln |\mathcal{H}| + m \ln(1-\epsilon)$

□

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c,\epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x)) \leq |\mathcal{H}|(1-\epsilon)^m$
5. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x)) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
6. On veut $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1-\delta$
   on sait maintenant que $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
   il suffit donc d'avoir $\delta \geq |\mathcal{H}|(1-\epsilon)^m$
7. $\delta \geq |\mathcal{H}|(1-\epsilon)^m \Leftrightarrow \ln\delta \geq \ln|\mathcal{H}| + m\ln(1-\epsilon)$
   $\Leftarrow \ln\delta \geq \ln|\mathcal{H}| - m\epsilon$ car $-\epsilon \geq \ln(1-\epsilon)$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c,\epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x)) \leq |\mathcal{H}|(1-\epsilon)^m$
5. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x)) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
6. On veut $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1-\delta$
   on sait maintenant que $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
   il suffit donc d'avoir $\delta \geq |\mathcal{H}|(1-\epsilon)^m$
7. $\delta \geq |\mathcal{H}|(1-\epsilon)^m \Leftrightarrow \ln\delta \geq \ln|\mathcal{H}| + m\ln(1-\epsilon)$
   $\Leftarrow \ln\delta \geq \ln|\mathcal{H}| - m\epsilon$ car $-\epsilon \geq \ln(1-\epsilon)$
   $\Leftarrow -\ln\delta \leq -\ln|\mathcal{H}| + m\epsilon$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c,\epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x)) \leq |\mathcal{H}|(1-\epsilon)^m$
5. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x)) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
6. On veut $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1-\delta$
   on sait maintenant que $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
   il suffit donc d'avoir $\delta \geq |\mathcal{H}|(1-\epsilon)^m$
7. $$\begin{aligned} \delta \geq |\mathcal{H}|(1-\epsilon)^m &\Leftrightarrow \ln\delta \geq \ln|\mathcal{H}| + m\ln(1-\epsilon) \\ &\Leftarrow \ln\delta \geq \ln|\mathcal{H}| - m\epsilon \qquad \text{car } -\epsilon \geq \ln(1-\epsilon) \\ &\Leftarrow -\ln\delta \leq -\ln|\mathcal{H}| + m\epsilon \\ &\Leftarrow \ln\tfrac{1}{\delta} + \ln|\mathcal{H}| \leq m\epsilon \end{aligned}$$

Apprentissage PAC : Ensemble d'hypothèse est fini

Démonstration.

1. On définit : $\mathcal{H}(c,\epsilon) =$ hyp. de $\mathcal{H}$ d'erreur ***réelle*** $> \epsilon$
2. $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$.
3. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x))$
   $= 1 - p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x))$
4. $p(\exists h \in \mathcal{H}(c,\epsilon) : \forall x \in S, c(x) = h(x)) \leq |\mathcal{H}|(1-\epsilon)^m$
5. $p(\forall h \in \mathcal{H}(c,\epsilon) : \exists x \in S, c(x) \neq h(x)) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
6. On veut $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1-\delta$
   on sait maintenant que $p(\text{erreur}_{\mathcal{D}}(h(\mathcal{A},S),c) \leq \epsilon) \geq 1 - |\mathcal{H}|(1-\epsilon)^m$
   il suffit donc d'avoir $\delta \geq |\mathcal{H}|(1-\epsilon)^m$
7. $$\begin{aligned} \delta \geq |\mathcal{H}|(1-\epsilon)^m &\Leftrightarrow \ln\delta \geq \ln|\mathcal{H}| + m\ln(1-\epsilon) \\ &\Leftarrow \ln\delta \geq \ln|\mathcal{H}| - m\epsilon \qquad \text{car } -\epsilon \geq \ln(1-\epsilon) \\ &\Leftarrow -\ln\delta \leq -\ln|\mathcal{H}| + m\epsilon \\ &\Leftarrow \ln\tfrac{1}{\delta} + \ln|\mathcal{H}| \leq m\epsilon \\ &\Leftarrow \tfrac{1}{\epsilon}(\ln\tfrac{1}{\delta} + \ln|\mathcal{H}|) \leq m \end{aligned}$$

□

Apprentissage PAC : Ensemble d'hypothèse est fini

Example

Conjonction de littéraux booléens :

- n attributs booléens
- $\mathcal{H}$ = ensembles des conjonction de littéraux sur ces attributs

Apprentissage PAC : Ensemble d'hypothèse est fini

Example

Conjonction de littéraux booléens :

- n attributs booléens
- $\mathcal{H}$ = ensembles des conjonction de littéraux sur ces attributs

$\Rightarrow$ $|\mathcal{H}| = \ldots \Rightarrow$ sample complexity en $\ldots$

Apprentissage PAC : Ensemble d'hypothèse est fini

Example

Conjonction de littéraux booléens :

- n attributs booléens
- $\mathcal{H}$ = ensembles des conjonction de littéraux sur ces attributs

$\Rightarrow$ $|\mathcal{H}| = \ldots \Rightarrow$ sample complexity en $\ldots$

- Pour $S = \{x_1, \ldots, x_m\}$ donné,
  on peut calculer h_S d'erreur empirique nulle en temps en $\Theta(n)$:
  1. $h \leftarrow x_1$
  2. Pour i allant de 2 à m faire $h \leftarrow \sup(h, x_i)$;
  3. Retourner h.

  (Temps de calcul de $\sup(h, x_i)$ *en ...)*

Apprentissage PAC : Ensemble d'hypothèse est fini

Example

Conjonction de littéraux booléens :

- n attributs booléens
- $\mathcal{H}$ = ensembles des conjonction de littéraux sur ces attributs

$\Rightarrow$ $|\mathcal{H}| = \ldots \Rightarrow$ sample complexity en $\ldots$

- Pour $S = \{x_1, \ldots, x_m\}$ donné,
  on peut calculer h_S d'erreur empirique nulle en temps en $\Theta(n)$:
  1. $h \leftarrow x_1$
  2. Pour i allant de 2 à m faire $h \leftarrow \sup(h, x_i)$;
  3. Retourner h.

  (Temps de calcul de $\sup(h, x_i)$ *en ...)*

$\Rightarrow$ On peut apprendre une hypothèse ayant probablement une erreur faible en un temps qui augmente linéairement avec n.

Apprentissage PAC : Temps

Apprentissage PAC : Temps

Definition

La classe d'hypothèses $\mathcal{H}$ définie sur un ensemble d'instances $\mathcal{X}$ de dimension n est ***probablement approximativement apprenable en temps polynomial*** s'il existe un algorithme qui, pour tout concept $c \in \mathcal{H}$, pour toute distribution de probabilité sur $\mathcal{X}$, et pour tous $0 < \delta, \epsilon < 1/2$, retourne une hypothèse $h \in \mathcal{H}$ dont l'erreur réelle est $\leq \epsilon$ avec une probabilité $> 1 - \delta$ en un temps en $O(1/\delta, 1/\epsilon, n)$ (en utilisant un ***oracle*** pour obtenir des exemples correctement étiquetés «*à la demande*»).

Apprentissage PAC : Temps

On a vu sur un exemple comment on peut montrer qu'une classe de concept est PAC-apprenable en temps polynomial :

1. on montre qu'un nombre d'exemples suffisant pour apprendre un concept augment linéairement avec la taille des instances et des concepts ;
2. on exhibe un algorithme consistant qui traite chaque exemple en temps polynomial.

Il y a des classes d'hypothèses

- pour lesquelles le nombre d'exemples suffisant augmente de manière polynomial,
- mais qui ne sont pas PAC-apprenable en temps polynomial car on ne peut trouver d'algorithme consistant qui traite chaque exemple en temps polynomial.

Apprentissage PAC : Temps

Example

k-termes DNF, pour k fixé

- n attributs booléens ; $k \in \mathbb{N}$ fixé
- $\mathcal{H}$ = ensembles des disjonctions de k conjonctions de littéraux

$$\text{si } k = 3 : (l_1 \wedge \ldots \wedge l_i) \vee (l_{i+1} \wedge \ldots \wedge l_j) \vee (l_{j+1} \wedge \ldots \wedge l_h)$$

$\Rightarrow$ $|\mathcal{H}| = \ldots \Rightarrow$ sample complexity en $\ldots$

Apprentissage PAC : Temps

Example

k-termes DNF, pour k fixé

- n attributs booléens ; $k \in \mathbb{N}$ fixé
- $\mathcal{H}$ = ensembles des disjonctions de k conjonctions de littéraux

$$\text{si } k = 3 : (l_1 \wedge \ldots \wedge l_i) \vee (l_{i+1} \wedge \ldots \wedge l_j) \vee (l_{j+1} \wedge \ldots \wedge l_h)$$

$\Rightarrow$ $|\mathcal{H}| = 3^n \Rightarrow$ sample complexity en ...

Apprentissage PAC : Temps

Example

k-termes DNF, pour k fixé

- n attributs booléens ; $k \in \mathbb{N}$ fixé
- $\mathcal{H}$ = ensembles des disjonctions de k conjonctions de littéraux

$$\text{si } k = 3 : (l_1 \wedge \ldots \wedge l_i) \vee (l_{i+1} \wedge \ldots \wedge l_j) \vee (l_{j+1} \wedge \ldots \wedge l_h)$$

$\Rightarrow$ $|\mathcal{H}| = 3^n \Rightarrow$ sample complexity en $\frac{1}{\epsilon}(\ln \frac{1}{\delta} + n \ln 3)$

Apprentissage PAC : Temps

Example

k-termes DNF, pour k fixé

- n attributs booléens ; $k \in \mathbb{N}$ fixé
- $\mathcal{H}$ = ensembles des disjonctions de k conjonctions de littéraux

$$\text{si } k = 3 : (l_1 \wedge \ldots \wedge l_i) \vee (l_{i+1} \wedge \ldots \wedge l_j) \vee (l_{j+1} \wedge \ldots \wedge l_h)$$

$\Rightarrow$ $|\mathcal{H}| = 3^n \Rightarrow$ sample complexity en $\frac{1}{\epsilon}(\ln \frac{1}{\delta} + n \ln 3)$

- Si (R)P $\neq$ NP : il n'y a pas d'algorithme qui retourne une hypothèse d'erreur empirique nulle en temps polynomial !

Espaces d'hypothèses infinis : Dimension de Vapnik-Chervonenkis

Definition

- Ensemble d'hypothèses $\mathcal{H}$
- Ensemble d'instances $X = \{x_1, \ldots, x_m\}$

$\Rightarrow$ $\mathcal{H}$ peut ***pulvériser*** X si :

$\forall$ étiquetage $S = \{(x_i, y_i) \mid 1 \leq i \leq m\}$

$\exists h \in H$, t.q. $\text{erreur}_S(h) = 0$

Alors :

VC-dimension de $H = \max\{|X| \mid \mathcal{H}$ peut pulveriser $X\}$.

Espaces d'hypothèses infinis : Dimension de Vapnik-Chervonenkis

Example

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ = droites / sép. linéaires :VC-dim($\mathcal{H}$) = :

Espaces d'hypothèses infinis : Dimension de Vapnik-Chervonenkis

Example

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ = droites / sép. linéaires :VC-dim($\mathcal{H}$) = 3 :
on peut toujours classer toute configuration de 3 points non alignés, mais jamais 4.

Espaces d'hypothèses infinis : Dimension de Vapnik-Chervonenkis

Example

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ = droites / sép. linéaires :VC-dim($\mathcal{H}$) = 3 :
on peut toujours classer toute configuration de 3 points non alignés, mais jamais 4.

Example

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ = cercles centrés à l'origine :
$h(x) = signe(qx^2 - r^2)$, où $q \in \{+1, -1\}$, $r \in \mathbb{R}$,
$\Rightarrow$ VC-dim($\mathcal{H}$) = .

Espaces d'hypothèses infinis : Dimension de Vapnik-Chervonenkis

Example

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ = droites / sép. linéaires :VC-dim($\mathcal{H}$) = 3 :
on peut toujours classer toute configuration de 3 points non alignés, mais jamais 4.

Example

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ = cercles centrés à l'origine :
$h(x) = signe(qx^2 - r^2)$, où $q \in \{+1, -1\}$, $r \in \mathbb{R}$,
$\Rightarrow$ VC-dim($\mathcal{H}$) = 2.

Espaces d'hypothèses infinis : Dimension de Vapnik-Chervonenkis

Example

$\mathcal{X} = \mathbb{R}$; $\mathcal{H}$ =hyp. de la forme $h_w(x) = \text{signe}(\sin(wx))$, pour $w \in \mathbb{R}$.

$\Rightarrow VC(\mathcal{H}) = +\infty$: pour M fixé

- $X_M = \{2, 2^2, \ldots, 2^M\}$
- pour $y_1, \ldots, y_M \in \{-1, +1\}$: $w = -(0, y_1 \ldots y_M)_2 \times \pi$
  alors $h_w(2^i) = y_i$

Donc $VC(\mathcal{H}) \geq M$

Espaces d'hypothèses infinis : Dimension de Vapnik-Chervonenkis

Example

Classification linéaire en dimension n : VC-dimension est $n + 1$.

Example

Soit $\mathcal{H}$ l'ensemble des hypothèses représentables par un réseau de neuronnes acyclique à n entrées et s unités internes, ayant chacune au plus r entrées, alors

$$VC(\mathcal{H}) \leq 2s(1 + \ln s) \times VC(\mathcal{H}_U)$$

où $\mathcal{H}_U$ est l'ensemble des hypothèses (sur $\mathbb{R}^r$) représentables par la famille des classifiieurs choisie pour les unités.

Espaces d'hypothèses infinis : Un résultat d'apprenabilité

Theorem

Soit $\mathcal{H}$ un ensemble d'hypothèses de VC-dimension > 2, S un ensemble d'exemples tel qu'il existe une hypothèse h_S de $\mathcal{H}$ consistante avec S[a], alors

$$p(erreur_{réelle}(h_S) \leq \epsilon) \geq 1 - \delta$$

dès que

$$|S| > \frac{1}{\epsilon}(4 \log_2 \frac{2}{\delta} + 8 VC(H) \log_2 \frac{13}{\epsilon})$$

[a] de risque empirique nul sur S

Espaces d'hypothèses infinis : Majoration du risque réel

L'inégalité de Vapnik-Chervnonenkis permet de majorer le risque réel à partir du risque empirique sur un échantillon de taille m :

$$\text{erreur}_{\text{réelle}}(h) \leq \text{erreur}_S(h) + \sqrt{\frac{VC(\mathcal{H})(1 + \ln \frac{2|S|}{VC(\mathcal{H})}) - \ln \frac{4}{\delta}}{m}}$$

avec une proba $1 - \delta$.

Apprentissage «*agnostique*»

Les résultats vus jusqu'ici reposent, entre autres, sur 2 hypothèses peu réalistes :

- les exemples de S sont non bruités ;
- le concept cible est dans $\mathcal{H}$.

Apprentissage «*agnostique*»

Les résultats vus jusqu'ici reposent, entre autres, sur 2 hypothèses peu réalistes :

- les exemples de S sont non bruités ;
- le concept cible est dans $\mathcal{H}$.

Example (Shalev-Shwartz and Ben-David, 2014)

Des papayes, 2 indices / attributs numériques

- couleur : 0=vert clair,. . ., 10 = marron foncé
- dureté : 0=très mou, . . ., 10 = dur comme du bois

On veut apprendre à reconnaître les bonnes papayes :

Apprentissage «*agnostique*»

Les résultats vus jusqu'ici reposent, entre autres, sur 2 hypothèses peu réalistes :

- les exemples de S sont non bruités ;
- le concept cible est dans $\mathcal{H}$.

Example (Shalev-Shwartz and Ben-David, 2014)

Des papayes, 2 indices / attributs numériques

- couleur : 0=vert clair,. . ., 10 = marron foncé
- dureté : 0=très mou, . . ., 10 = dur comme du bois

On veut apprendre à reconnaître les bonnes papayes :

- on cherche un rectangle de $[0, 10] \times [0, 10]$;
- le concept cible n'est sans doute pas aussi régulier
- il est possible qu'il y ait deux papayes :
  - l'une bonne, l'autre non
  - ayant les même caractéristiques

Apprentissage «*agnostique*» : Un cadre plus général

non-agnostique	agnostique

Apprentissage «*agnostique*» : Un cadre plus général

	non-agnostique	agnostique
distribution	sur $\mathcal{X}$	sur $\mathcal{X} \times \mathcal{Y}$

Apprentissage «*agnostique*» : Un cadre plus général

	non-agnostique	agnostique
distribution	sur $\mathcal{X}$	sur $\mathcal{X} \times \mathcal{Y}$
cible	$c \in \mathcal{H}$	$c \notin \mathcal{H}$ (ou pas de c)

Apprentissage «*agnostique*» : Un cadre plus général

	non-agnostique	agnostique
distribution	sur $\mathcal{X}$	sur $\mathcal{X} \times \mathcal{Y}$
cible	$c \in \mathcal{H}$	$c \notin \mathcal{H}$ (ou pas de c)
exemples	$((x_i, c(x_i)))_{1 \leq i \leq n}$	$((x_i, y_i))_{1 \leq i \leq n}$

Apprentissage «*agnostique*» : Un cadre plus général

	non-agnostique	agnostique
distribution	sur $\mathcal{X}$	sur $\mathcal{X} \times \mathcal{Y}$
cible	$c \in \mathcal{H}$	$c \notin \mathcal{H}$ (ou pas de c)
exemples	$((x_i, c(x_i)))_{1 \leq i \leq n}$	$((x_i, y_i))_{1 \leq i \leq n}$
risque réel	$p(c(x) \neq h(x))$	$p(y \neq h(x))$

Apprentissage «*agnostique*» : Un cadre plus général

	non-agnostique	agnostique
distribution	sur $\mathcal{X}$	sur $\mathcal{X} \times \mathcal{Y}$
cible	$c \in \mathcal{H}$	$c \notin \mathcal{H}$ (ou pas de c)
exemples	$((x_i, c(x_i)))_{1 \leq i \leq n}$	$((x_i, y_i))_{1 \leq i \leq n}$
risque réel	$p(c(x) \neq h(x))$	$p(y \neq h(x))$
objectif	$\text{erreur}_c(h) \leq \epsilon$	$\text{erreur}_c(h) \leq \min_{h' \in \mathcal{H}}(\text{erreur}_c(h')) + \epsilon$

Apprentissage «*agnostique*» : Définition et résultats

Definition

Un ensemble d'hypothèses $\mathcal{H}$ sur un espace $\mathcal{X} \times \mathcal{Y}$ est PAC-apprenable dans le cadre agnostique s'il existe une fonction $m :]0,1[^2 \to \mathbb{N}$ et un algorithme tels que pour tous $\delta, \epsilon \in]0,1[^2$ et toute distribution sur $\mathcal{X} \times \mathcal{Y}$, si on fournit à l'algorithme un ensemble d'au moins $m(\delta, \epsilon)$ exemples alors il retourne avec une probabilité $\geq 1 - \delta$ une hypothèse h telle que :

$$\text{erreur}_c(h) \leq \min_{h' \in \mathcal{H}} (\text{erreur}_c(h')) + \epsilon$$

Apprentissage «*agnostique*» : Définition et résultats

Theorem

Si $\mathcal{H}$ est finie, alors $\mathcal{H}$ est PAC-apprenable dans le cadre agnostique, avec une sample complexity d'au plus

$$\frac{2}{\epsilon^2}(\log\frac{1}{\delta} + \log(2|\mathcal{H}|)$$

Apprentissage «*agnostique*» : Définition et résultats

Theorem

Si $\mathcal{H}$ est finie, alors $\mathcal{H}$ est PAC-apprenable dans le cadre agnostique, avec une sample complexity d'au plus

$$\frac{2}{\epsilon^2}(\log\frac{1}{\delta} + \log(2|\mathcal{H}|)$$

Theorem

Les trois propriétés suivantes sont équivalentes :

- *$\mathcal{H}$ est PAC-apprenable*
- *$\mathcal{H}$ est PAC-apprenable dans le cadre agnostic*
- *$VC(\mathcal{H})$ est finie*

(Dans le cadre de la classification binaire.)

Apprentissage «*en ligne*»

Situations où les exemples doivent être présentés en séquence à l'apprenant

- observation des utilisateurs d'un système d'aide à la décision
  on traite immédiatement chaque information
- ensemble d'exemples trop grand pour être conservé / traité d'un seul coup

$\Rightarrow$ on adapte le modèle au fur et à mesure de l'arrivée des exemples

$\Rightarrow$ apprentissage incrémental

Apprentissage «*en ligne*» : Le perceptron incrémental

- Ensemble d'instances $\mathcal{X} = \mathbb{R}^d$
- Deux labels $\mathcal{Y} = \{-1, +1\}$
- Séquence S d'exemples (x, y)
- $\mathcal{H}$ = séparateurs linéaires

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Apprentissage «*en ligne*» : Le perceptron incrémental

- Ensemble d'instances $\mathcal{X} = \mathbb{R}^d$
- Deux labels $\mathcal{Y} = \{-1, +1\}$
- Séquence S d'exemples (x, y)
- $\mathcal{H}$ = séparateurs linéaires

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Theorem

- *Si* $\|x\| = 1$, *pour tout* $(x, y) \in S$,
- *et si* $\exists h_{a,b} \in \mathcal{H}$ *d'erreur empirique nulle sur* S,

alors l'algorithme du perceptron incrémental fait au plus $1/\gamma^2$ *erreurs sur* S, *où*

$$\gamma = \min_{(x,y) \in S} \frac{|a \cdot x + b|}{\|a, b\|}$$

Apprentissage «*en ligne*» : Le perceptron incrémental

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Démonstration.

Notations :

- $w = (a, b)$ = vecteur d'erreur empirique nulle sur S, avec $\|w\| = 1$
- $w_t = (a_t, b_t)$ = avant la t-ième erreur
- (x_t, u_t) exemple qui cause la t-ième erreur

□

Apprentissage «*en ligne*» : Le perceptron incrémental

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Démonstration.

- $\gamma = \min_{(x,y)\in S} |a \cdot x + b|/\|a, b\| = \min_{(x,y)\in S} |w \cdot (x, 1)|/\|w\|$

Apprentissage «*en ligne*» : Le perceptron incrémental

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Démonstration.

- $\gamma = \min_{(x,y)\in S} |a \cdot x + b|/\|a, b\| = \min_{(x,y)\in S} |w \cdot (x, 1)|/\|w\|$
- si $y_t = 1$: $w \cdot (x_t, 1) \geq 0$ (erreur empirique nulle)

Apprentissage «*en ligne*» : Le perceptron incrémental

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Démonstration.

- $\gamma = \min_{(x,y)\in S} |a \cdot x + b|/\|a, b\| = \min_{(x,y)\in S} |w \cdot (x, 1)|/\|w\|$
- si $y_t = 1$: $w \cdot (x_t, 1) \geq 0$ (erreur empirique nulle)
- donc $w \cdot (x_t, 1) = |w \cdot (x_t, 1)| = (|w \cdot (x_t, 1)|/\|w\|)\|w\| \geq \gamma\|w\|$

Apprentissage «*en ligne*» : Le perceptron incrémental

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Démonstration.

- $\gamma = \min_{(x,y)\in S} |a \cdot x + b|/\|a, b\| = \min_{(x,y)\in S} |w \cdot (x, 1)|/\|w\|$
- si $y_t = 1$: $w \cdot (x_t, 1) \geq 0$ (erreur empirique nulle)
- donc $w \cdot (x_t, 1) = |w \cdot (x_t, 1)| = (|w \cdot (x_t, 1)|/\|w\|)\|w\| \geq \gamma\|w\|$
- donc $w_{t+1} \cdot w = w_t \cdot w + (x_t, 1) \cdot w \geq w_t \cdot w + \gamma\|w\|$
  $\Rightarrow$ comme $w_0 = 0$, après M erreurs : $w_{M+1} \cdot w \geq M\gamma\|w\|$;

Apprentissage «*en ligne*» : Le perceptron incrémental

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Démonstration.

- $\gamma = \min_{(x,y)\in S} |a \cdot x + b|/\|a, b\| = \min_{(x,y)\in S} |w \cdot (x, 1)|/\|w\|$
- si $y_t = 1$: $w \cdot (x_t, 1) \geq 0$ (erreur empirique nulle)
- donc $w \cdot (x_t, 1) = |w \cdot (x_t, 1)| = (|w \cdot (x_t, 1)|/\|w\|)\|w\| \geq \gamma\|w\|$
- donc $w_{t+1} \cdot w = w_t \cdot w + (x_t, 1) \cdot w \geq w_t \cdot w + \gamma\|w\|$
  $\Rightarrow$ comme $w_0 = 0$, après M erreurs : $w_{M+1} \cdot w \geq M\gamma\|w\|$;
- si $y_t = 1$: $w_t \cdot x_t \leq 0$ (car erreur avec w_t sur (x_t, y_t)),
  donc $\|w_t + x_t\|^2 = \|w_t\|^2 + \|x\|_t^2 + 2w_t \cdot x_t \leq \|w_t\|^2 + 1$
  donc $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$ et donc, comme $\|w_0\| = 0$: $\|w_{M+1}\| \leq \sqrt{M}$.

□

Apprentissage «*en ligne*» : Le perceptron incrémental

Perceptron incrémental

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. pour chaque nouvel exemple (x, y) reçu, si $(a \cdot x + b) \times y \leq 0$ faire :
   $(a, b) \leftarrow (a, b) + y \times (x, 1)$;
3. retourner (a, b).

Démonstration.

- $\gamma = \min_{(x,y)\in S} |a \cdot x + b|/\|a, b\| = \min_{(x,y)\in S} |w \cdot (x, 1)|/\|w\|$
- si $y_t = 1$: $w \cdot (x_t, 1) \geq 0$ (erreur empirique nulle)
- donc $w \cdot (x_t, 1) = |w \cdot (x_t, 1)| = (|w \cdot (x_t, 1)|/\|w\|)\|w\| \geq \gamma\|w\|$
- donc $w_{t+1} \cdot w = w_t \cdot w + (x_t, 1) \cdot w \geq w_t \cdot w + \gamma\|w\|$
  $\Rightarrow$ comme $w_0 = 0$, après M erreurs : $w_{M+1} \cdot w \geq M\gamma\|w\|$;
- si $y_t = 1$: $w_t \cdot x_t \leq 0$ (car erreur avec w_t sur (x_t, y_t)),
  donc $\|w_t + x_t\|^2 = \|w_t\|^2 + \|x\|_t^2 + 2w_t \cdot x_t \leq \|w_t\|^2 + 1$
  donc $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$ et donc, comme $\|w_0\| = 0$: $\|w_{M+1}\| \leq \sqrt{M}$.
- or $w_{M+1} \cdot w = \|w_{M+1}\|\|w\| \cos(w_{M+1}, w) \leq \|w_{M+1}\|$, donc $\sqrt{M} \geq \gamma M$.

□

Apprentissage «*en ligne*» : Le perceptron incrémental

- Au pire, γ peut être exponentiellement petit en n
  $\Rightarrow$ Perceptron efficace lorsque les exemples sont « bien séparés » (vaste marge)
- On peut utiliser des fonctions noyaux :
  à tout instant (a, b) est une somme des x_t

Apprentissage «*en ligne*» : Le perceptron incrémental

Cas non séparable

1. $a \leftarrow (0, \ldots, 0)$; $b \leftarrow 0$; pas_fini $\leftarrow$ vrai ;
2. Pour chaque nouvel exemple (x, u) reçu :
   si $(a \cdot x + b)/\|(a, b)\| \times u \leq -\gamma/2$ faire :
   $a \leftarrow a + u \times x$; $b \leftarrow b + u$;
3. Retourner (a, b).

Dans ce cas, le nombre d'erreurs est au pire $8/\gamma^2$.

Apprentissage actif

Dasgupta [2011] « *As digital storage gets cheaper, and sensing devices proliferate, and the web grows ever larger, it gets easier to amass vast quantities of unlabeled data – raw speech, images, text documents, and so on. But to build classifiers from these data, labels are needed, and obtaining them can be costly and time consuming. An active learner would try to get the most out of a limited budget by choosing its query points in an intelligent and adaptive manner.* »

Aussi : apprendre les préférences d'un utilisateur d'un système d'aide à la décision

$\Rightarrow$ lui poser le moins de questions possible (risque d'abandon)

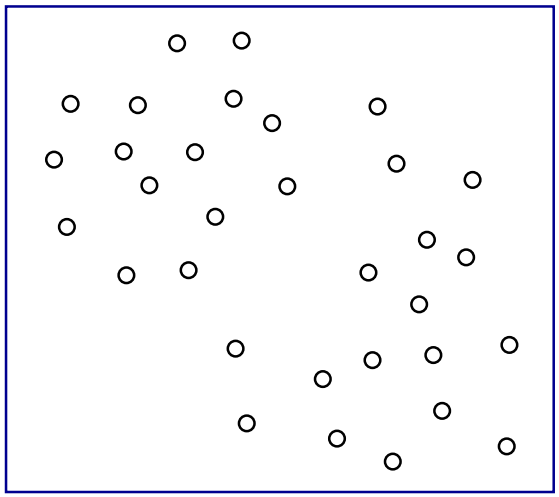
Apprentissage actif

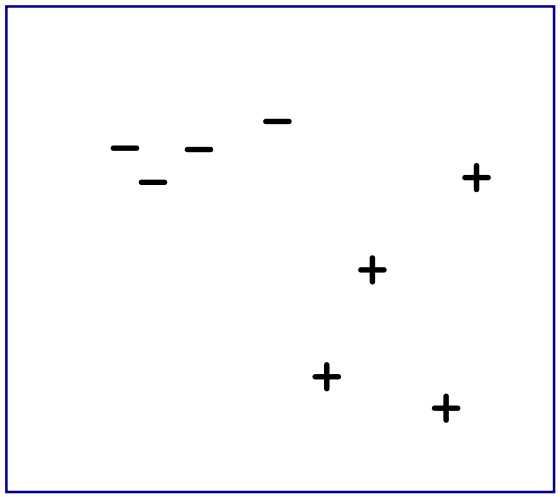
Il s'agit maintenant d'apprendre avec le moins possible d'exemples, en ***choisissant*** bien les exemples.

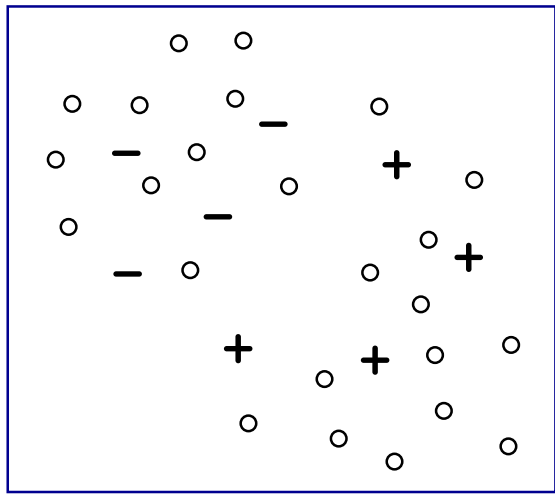
Par exemple, pour apprendre un classifieur linéaire en dimension 1, il faut de l'ordre de $1/\epsilon$ exemples en apprentissage passif, mais $\log_2(1/\epsilon)$ si on adopte une recherche dichotomique.

Apprentissage actif

Apprentissage non-supervisé / supervisé / semi-supervisé ou actif :







Apprentissage actif : Un algorithme pas agressif

Si on ne contrôle pas l'arrivée des instances, mais on veut demander le moins possible les étiquettes :

Cohn et al. [1994] Apprentissage actif – cas séparable

1. Initialiser $H \leftarrow \mathcal{H}$
2. Tant que...
   a. $x \leftarrow$ nouvelle instance non étiquetée ;
   b. s'il y a désaccord dans H sur la classe de x :
   demander la classe $c(x)$ de x ; $H \leftarrow \{h \in H : h(x) = c(x)\}$

Apprentissage actif : Un algorithme pas agressif

Version où H est représenté par les exemples rencontrés :

Cohn et al. [1994] Apprentissage actif – cas séparable

1. $S \leftarrow \emptyset$
2. Tant que...
   a. $x \leftarrow$ nouvelle instance non étiquetée
   b. si $S \cup \{(x, +1)\}$ n'est pas séparable : $y \leftarrow -1$;
      sinon si $S \cup \{(x, -1)\}$ n'est pas séparable : $y \leftarrow +1$;
      sinon : demander la classe $c(x)$ de x ; $y \leftarrow c(x)$;
   c. $S \leftarrow S \cup \{(x, y)$.

On n'a pas besoin de mettre S à jour lorsqu'on n'a pas eu à demander la classe de x.

Apprentissage actif : Un algorithme pas agressif

Soit x_t l'instance reçue à la t-ième itération, sa classe est connue après cette itération – soit donnée par H, soit demandée. Donc, après t itérations, on connaît la classe de t instances.

Apprentissage actif : Un algorithme pas agressif

Complexité en nombre d'exemples : on cherche le nombre d'itérations nécessaires pour un apprentissage PAC.

Soit $T_{\min}(\epsilon, \delta) =$ le plus petit entier T tel que pour tout $t \geq T$:

$$p(\text{il existe } h \in H_t, \text{erreur}_{\mathcal{D}}(h) > \epsilon) \leq \delta$$

Theorem

Si $VC(\mathcal{H})$ est finie, et si le concept cible est représentable par une hypothèse de $\mathcal{H}$, et si les exemples sont corrects, alors

$$T_{\min}(\epsilon) \leq O\left(\theta VC(\mathcal{H}) \log \frac{1}{\epsilon}\right)$$

où θ est le coefficient de désaccord.

En apprentissage actif, le nombre d'exemples nécessaires pour atteindre une précision ϵ est donc en $\log \frac{1}{\epsilon}$, alors qu'on a vu qu'avec un algorithme passif, il est en $\frac{1}{\epsilon}$: on a une amélioration exponentielle.

Apprentissage actif : Un algorithme pas agressif

Coefficient de désaccord entre deux ensembles d'hypothèses :

- pour un ensemble d'hypothèse $V \subseteq \mathcal{H}$:
  $\text{désac}(V) = |\{x \in \mathcal{X} \mid \exists h, h' \in V, h(x) \neq h'(x)\}|$
- si h^* est une hypothèse optimale (pour c) de $\mathcal{H}$, et $r \in \mathbb{R}$:
  $B(h^*, r) = \{h \in \mathcal{H} \mid |\{x \in \mathcal{X} \mid h(x) \neq h^*(x)\}| < r\}$

Alors :

- $p_{\mathcal{D}}(\text{désac}(B(h^*, r))$ est la probabilité qu'il faudra demander la classe d'un exemple si l'espace des version est contenu dans $B(h^*, r)$.
- $\theta = \sup_{r>0} \frac{p_{\mathcal{D}}(\text{désac}(B(h^*, r))}{r}$

Exemple de coefficient de désaccord : si l'ensemble des instances est $\mathbb{R}$, et si les hyptothèses sont les seuils dans $\mathbb{R}$, alors $\theta = 2$

David A. Cohn, Les E. Atlas, and Richard E. Ladner. Improving generalization with active learning. ***Machine Learning***, 15(2) :201–221, 1994. URL `http://dx.doi.org/10.1007/BF00993277`.

Sanjoy Dasgupta. Two faces of active learning. ***Theoretical Computer Science***, 412(19) :1767–1781, 2011.

Shai Shalev-Shwartz and Shai Ben-David. ***Understanding machine learning : From theory to algorithms***. Cambridge university press, 2014.