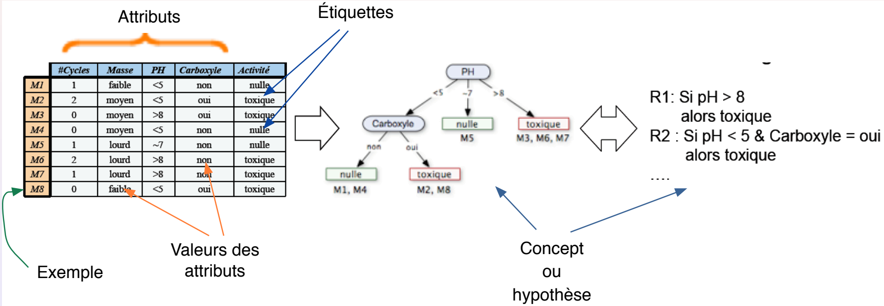
# En introduction à l'**Apprentissage Artificiel**

**Antoine Cornuéjols**

**MMIP, AgroParisTech, Paris**

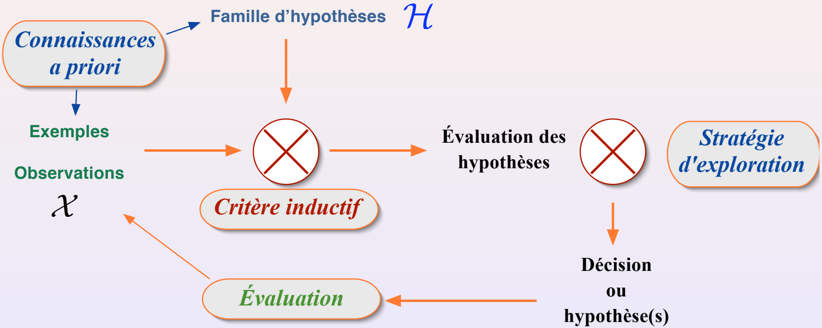
7 février 2008

# Un exemple




| | #Cycles | Masse | PH | Carboxyle | Activité |
|---|---|---|---|---|---|
| M1 | 1 | faible | <5 | non | nulle |
| M2 | 2 | moyen | <5 | oui | toxique |
| M3 | 0 | moyen | >8 | oui | toxique |
| M4 | 0 | moyen | <5 | non | nulle |
| M5 | 1 | lourd | ~7 | non | nulle |
| M6 | 2 | lourd | >8 | non | toxique |
| M7 | 1 | lourd | >8 | non | toxique |
| M8 | 0 | faible | <5 | oui | toxique |

# L'apprentissage
## Ingrédients



# Plan

# Plan

# Que signifie : avoir un bon modèle du monde ?

## Idéalement

# Que signifie : avoir un bon modèle du monde ?

## Idéalement

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{XY}}$
- *Fonction cible* $f : \mathcal{X} \rightarrow \mathcal{Y}$

# Que signifie : avoir un bon modèle du monde ?
## Idéalement

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{XY}}$
- *Fonction cible* $f : \mathcal{X} \to \mathcal{Y}$

**Prédire** correctement

$$\mathcal{X} \to \mathcal{Y}$$

$$x \mapsto \texttt{decision bayésienne}(x)$$

$$\text{ou} \quad x \mapsto \texttt{decision}(x) \; = \; f(x)$$

# Que signifie : avoir un bon modèle du monde ?

## Idéalement

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$
- *Fonction cible* $f : \mathcal{X} \rightarrow \mathcal{Y}$

### **Prédire** correctement

$$\mathcal{X} \rightarrow \mathcal{Y}$$

$$x \mapsto \texttt{decision bayésienne}(x)$$

$$\text{ou} \quad x \mapsto \texttt{decision}(x) \; = \; f(x)$$

### **Expliquer** le monde

$$h(\cdot) \; = \; f(\cdot)$$

# Que signifie : avoir un bon modèle du monde ?

## Idéalement

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{XY}}$
- *Fonction cible* $f : \mathcal{X} \rightarrow \mathcal{Y}$

**Prédire** correctement

$$\mathcal{X} \rightarrow \mathcal{Y}$$

$$x \mapsto \texttt{decision bayésienne}(x)$$

$$\text{ou} \quad x \mapsto \texttt{decision}(x) \; = f(x)$$

**Expliquer** le monde

$$h(\cdot) \; = f(\cdot)$$

Indécidable et illusoire

# Que signifie : avoir un bon modèle du monde ?

## Plus réalistement

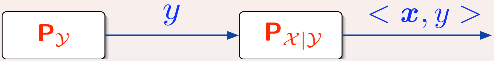



Figure: Modèle de génération des exemples.

# Que signifie : avoir un bon modèle du monde ?

## Plus réalistement

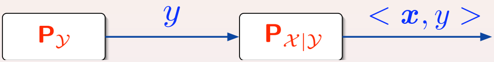



Figure: Modèle de génération des exemples.

**Prédictions correctes** (la plupart du temps)

$$L(h) = \mathbf{P}_{\mathcal{X}\mathcal{Y}}\{h(\boldsymbol{x}) \neq y\}$$

# Que signifie : avoir un bon modèle du monde ?

## Plus réalistement

$$\boxed{\mathbf{P}_{\mathcal{Y}}} \xrightarrow{\;y\;} \boxed{\mathbf{P}_{\mathcal{X}|\mathcal{Y}}} \xrightarrow{<\boldsymbol{x},\, y>}$$

Figure: Modèle de génération des exemples.

**Prédictions correctes** (la plupart du temps)

$$L(h) \;= \mathbf{P}_{\mathcal{X}\mathcal{Y}}\{h(\boldsymbol{x}) \neq y\}$$

**Expliquer le monde**

$$h(\cdot) \approx f(\cdot)$$

# Que signifie : avoir un bon modèle du monde ?

## Le risque réel

*Fonction de perte* :

$$\begin{array}{rcl} \ell(h) : \mathcal{X} \times \mathcal{Y} & \longrightarrow & \mathbb{R}^+ \\ (\boldsymbol{x}, y) & \longmapsto & \ell(h(\boldsymbol{x}), y) \end{array}$$

# Que signifie : avoir un bon modèle du monde ?

## Le risque réel

*Fonction de perte* :

$$\begin{aligned} \ell(h) : \mathcal{X} \times \mathcal{Y} &\rightarrow \mathbb{R}^+ \\ (\boldsymbol{x}, y) &\mapsto \ell(h(\boldsymbol{x}), y) \end{aligned}$$

**Risque réel** : espérance de perte

$$R(h) \;=\; \mathbb{E}[\ell(h(\boldsymbol{x}), y)] \;=\; \int_{\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}} \ell(h(\boldsymbol{x}), y) \; \mathbf{P}_{\mathcal{X}\mathcal{Y}} \, d(\boldsymbol{x}, y)$$

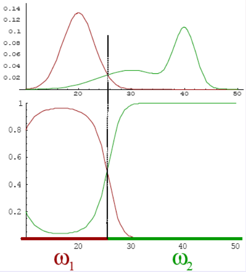
# Décision bayésienne

## Règle de Bayes

$$P(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_i)\, P(\mathcal{C}_i)}{p(\boldsymbol{x})}$$

## Décision bayésienne

$$\mathcal{C}_k = \underset{\mathcal{C}_k \in \mathcal{H}}{\mathrm{ArgMax}}\, P(\mathcal{C}_k|\boldsymbol{x})$$



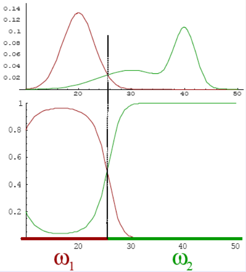
# Décision bayésienne

## Règle de Bayes

$$P(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_i)\, P(\mathcal{C}_i)}{p(\boldsymbol{x})}$$

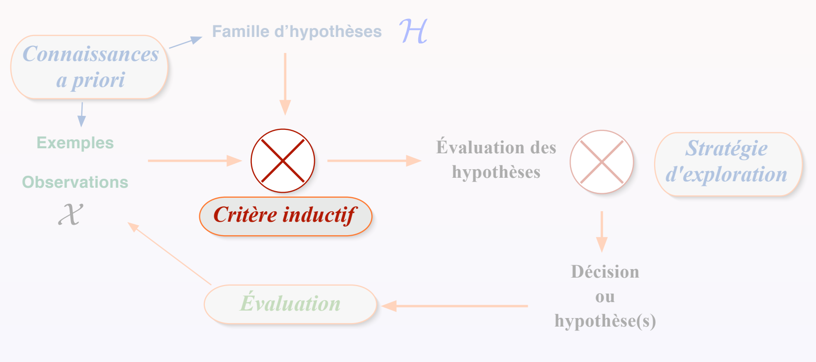
## Décision bayésienne

$$\mathcal{C}_k = \underset{\mathcal{C}_k \in \mathcal{H}}{\mathrm{ArgMax}}\, P(\mathcal{C}_k|\boldsymbol{x})$$




Il faut connaître $P(\mathcal{C}_k)$ et les lois de probabilité $p(\boldsymbol{x}|\mathcal{C}_i)$ !

# L'apprentissage
## Ingrédients



# Les critères inductifs

Mais, on ne connaît pas $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$

***Échantillon d'apprentissage*** supposé représentatif

$$\mathcal{S}_m = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$$

# Les critères inductifs

Mais, on ne connaît pas $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$

***Échantillon d'apprentissage*** supposé représentatif

$$\mathcal{S}_m = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$$

## Minimisation du Risque Empirique

$$R_m(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(\boldsymbol{x}_i), y_i)$$

# Les critères inductifs

## Minimisation du Risque Empirique

### MRE

Choisir l'hypothèse $\hat{h}$ telle que : $\hat{h} \; = \; \mathrm{ArgMin}_{h \in \mathcal{H}} \left[ R_{Emp}(h) \right]$

$$R_{\mathrm{Emp}}(h) \; = \; \frac{1}{m} \sum_{(\boldsymbol{x}_i, u_i) \in S} \ell(h(\boldsymbol{x}_i), u_i)$$

# Les critères inductifs

## Autres critères

### **Compression** maximale d'**information**

Choisir l'hypothèse $\hat{h}$ telle que : $\hat{h} \; = \; \mathrm{ArgMin}_{h \in \mathcal{H}} \left[ L(\mathcal{S}_m) \right]$

$$L(\mathcal{S}_m) \; = \; L(h) + L(\mathcal{S}_m | h))$$

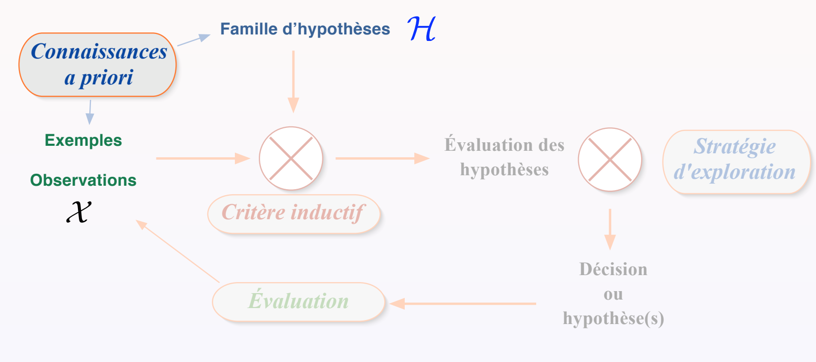
### MLE et MAP

Choisir l'hypothèse $\hat{h}$ telle que :

$$\hat{h} \; = \; \underset{h \in \mathcal{H}}{\mathrm{ArgMax}} \; l(h) = \underset{h \in \mathcal{H}}{\mathrm{ArgMax}} \; \ln \left[ p(\mathcal{S}_m | h) \right] \qquad \text{(MLE)}$$

$$\hat{h} \; = \; \underset{h \in \mathcal{H}}{\mathrm{ArgMax}} \; p(\mathcal{S}_m | h) \, p(h) \qquad \text{(MAP)}$$

# L'apprentissage
## Ingrédients



# Des hypothèses et des modèles

## Nombreuses formes possibles :

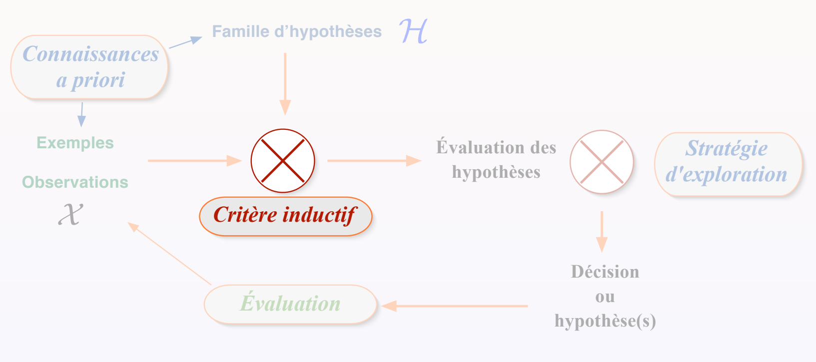
- “simple” plus proche(s) voisin(s)
- SVM

- Modèles linéaires
- Modèles bayésiens

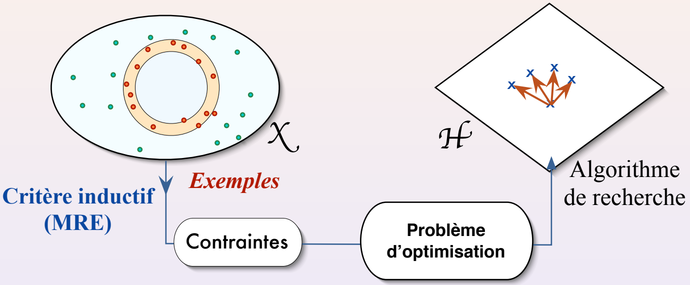
- Réseaux de neurones ; Modèles de Markov à états cachés (HMM)

- Arbres de décision
- Règles (ILP : Induction of Logic Programs)
- Grammaires

# L'apprentissage
## Ingrédients



# Les critères inductifs
## doivent vérifier ...




*Les entrées doivent **se traduire en «différences »** exploitables dans $\mathcal{H}$*

# Un dilemme fondamental

## Le compromis biais-variance

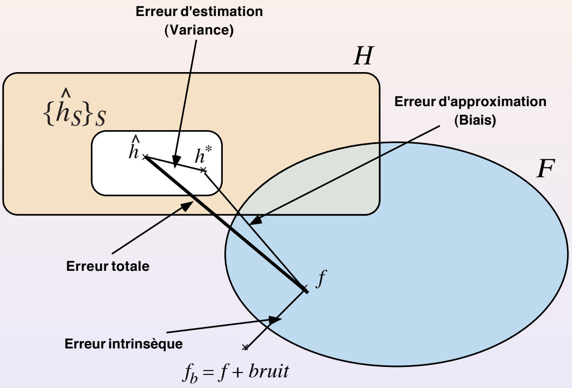



FIG.: Les différents types d'erreurs.

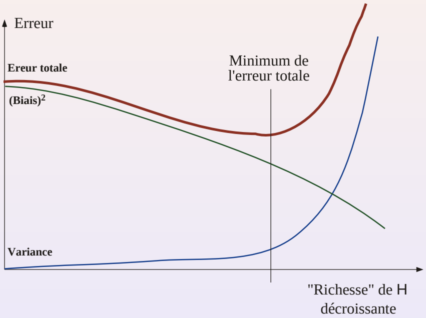
# Capacité de $\mathcal{H}$

## Qualité de l'estimation

$$|R(h) - R_{\mathrm{Emp}}(h)| \;\leq_P\; fct(\mathtt{diversit\acute{e}}_{\mathcal{H}}, m)$$

- Dimension de Vapnik-Chervonenkis
- Complexité de Rademacher
- BIC
- AIC
- ...

# Critère inductif régularisé




Corriger le MRE en contrôlant la capacité de $\mathcal{H}$

(ou la complexité de $h$)

# Critère inductif régularisé

## Contrôler $d_{\mathcal{H}}$

1. "Sélection de modèle"
2. Puis choix de $h \in \mathcal{H}$

$$\hat{h} \;=\; \mathrm{ArgMin}_{h \in \mathcal{H}} \left[ R_{Emp}(h) + \text{Capacité}(\mathcal{H}) \right]$$

# Critère inductif régularisé

## Contrôler $d_{\mathcal{H}}$

1. "Sélection de modèle"
2. Puis choix de $h \in \mathcal{H}$

$$\hat{h} \;=\; \mathrm{ArgMin}_{h \in \mathcal{H}} \left[ R_{Emp}(h) + \text{Capacité}(\mathcal{H}) \right]$$

## Régularisation

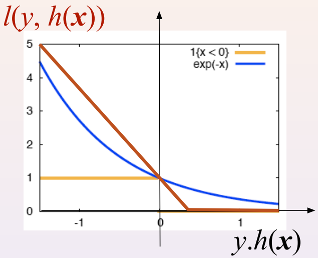
- Contrôler directement la complexité de $h$

$$\hat{h} \;=\; \mathrm{ArgMin}_{h \in \mathcal{H}} \left[ R_{Emp}(h) + \lambda\, \text{Reg}(h) \right]$$

# Optimisation du critère inductif régularisé

Minimisation du $\phi$-risque empirique

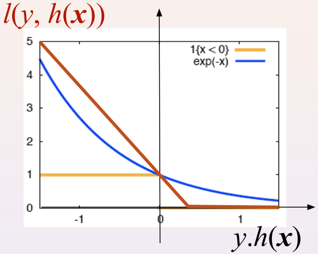
- "hinge loss"
- Fonction de perte exponentielle
- ...



# Optimisation du critère inductif régularisé

Minimisation du $\phi$-risque empirique

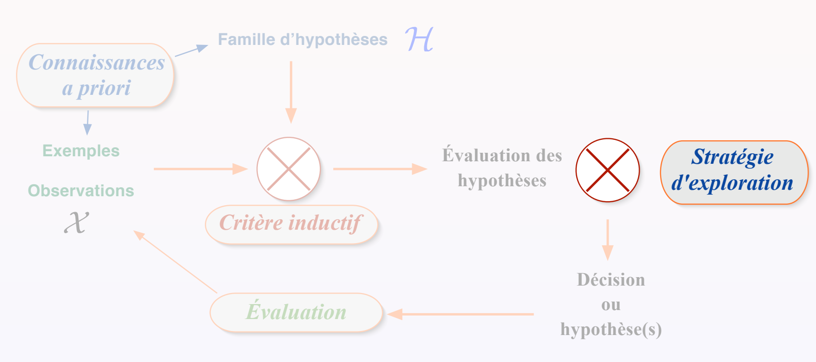
- "hinge loss"
- Fonction de perte exponentielle
- ...




Deux rôles :

- Régulariser
- Faciliter l'optimisation
  - Différentiabilité
  - Convexité

# L'apprentissage
## Ingrédients



# Exploration de $\mathcal{H}$

## Structures sur $\mathcal{H}$

# Exploration de $\mathcal{H}$

## Structures sur $\mathcal{H}$

### Pas de $\mathcal{H}$

Plus-proches-voisins

# Exploration de $\mathcal{H}$

## Structures sur $\mathcal{H}$

### Pas de $\mathcal{H}$

Plus-proches-voisins

### $\mathcal{H}$ muni d'une distance

*Réseaux de neurones ; régression logistique ; modèles bayésiens ; HMM ; ...*

- Optimisation directe (e.g. pseudo-inverse)
- Adaptation itérative = descente de gradient

# Exploration de $\mathcal{H}$

## Structures sur $\mathcal{H}$

### Pas de $\mathcal{H}$

Plus-proches-voisins

### $\mathcal{H}$ muni d'une distance

*Réseaux de neurones ; régression logistique ; modèles bayésiens ; HMM ; ...*

- Optimisation directe (e.g. pseudo-inverse)
- Adaptation itérative = descente de gradient

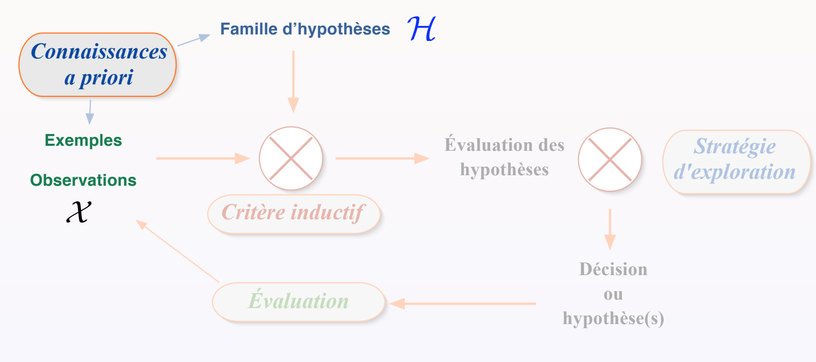
### $\mathcal{H}$ muni d'une relation de généralité

*Inférence grammaticale ; Induction de règles ; Apprentissage relationnel*

- Apprentissage symbolique
- Bruit

# Plan

# L'apprentissage
## Ingrédients



# Choix de $\mathcal{H}$

## Types d'espaces d'hypothèses

# Choix de $\mathcal{H}$

## Types d'espaces d'hypothèses

### Modèles génératifs

Demandent $p(\boldsymbol{x}|\mathcal{C}_k)$ et $p(\mathcal{C}_k)$

# Choix de $\mathcal{H}$

## Types d'espaces d'hypothèses

### Modèles génératifs

Demandent $p(\boldsymbol{x}|\mathcal{C}_k)$ et $p(\mathcal{C}_k)$

### Fonctions de décision

- Fonctions composées de fonctions de base
- Méthodes à Noyaux (*Kernel methods*)

# Choix de $\mathcal{H}$

## Types d'espaces d'hypothèses

### Modèles génératifs

Demandent $p(\boldsymbol{x}|\mathcal{C}_k)$ et $p(\mathcal{C}_k)$

### Fonctions de décision

- Fonctions composées de fonctions de base
- Méthodes à Noyaux (*Kernel methods*)

### Approches constructives

- Spécialisation dans l'espace $\mathcal{X}$
- Modèles hiérarchiques ou “profonds” (*Deep models*)

# Choix de $\mathcal{H}$

## Modèles génératifs

1. **Estimer** $p(\boldsymbol{x}|\mathcal{C}_k)$ (et $p(\mathcal{C}_k)$) $\forall \mathcal{C}_k$

   par le ***Principe du Maximum de Vraisemblance*** (MLE)

# Choix de $\mathcal{H}$

## Modèles génératifs

1. **Estimer** $p(\boldsymbol{x}|\mathcal{C}_k)$ (et $p(\mathcal{C}_k)$) $\forall \mathcal{C}_k$

   par le ***Principe du Maximum de Vraisemblance*** (MLE)

2. **Décider**, en utilisant le *théorème de Bayes* :

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)\, p(\mathcal{C}_k)}{p(\boldsymbol{x})}$$

avec $$p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x}|\mathcal{C}_k)\, p(\mathcal{C}_k)$$

# Choix de $\mathcal{H}$

## Modèles génératifs

1. **Estimer** $p(\boldsymbol{x}|\mathcal{C}_k)$ (et $p(\mathcal{C}_k)$) $\forall \mathcal{C}_k$

   par le ***Principe du Maximum de Vraisemblance*** (MLE)

2. **Décider**, en utilisant le *théorème de Bayes* :

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)\; p(\mathcal{C}_k)}{p(\boldsymbol{x})}$$

avec $$p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x}|\mathcal{C}_k)\; p(\mathcal{C}_k)$$

Exemple : Deux classes supposées gaussiennes de mêmes matrices de covariance

$$p(\boldsymbol{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\mu_k)^\top \Sigma^{-1} (\boldsymbol{x}-\mu_k)\}$$

# Choix de $\mathcal{H}$

## Modèles probabilistes discriminatifs

- **Estimer directement** $p(\mathcal{C}_k|\boldsymbol{x})$ en utilisant un modèle paramétré

**Exemple** : Deux classes et régression logistique

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})} = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$$

1. Fonction de vraisemblance : $p(\boldsymbol{y}|\boldsymbol{w}) = \Pi_{i=1}^{m} \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)^{y_i} \left\{1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right\}^{(1-y_i)}$
2. Fonction d'erreur :
$E(\boldsymbol{w}) = -\ln p(\boldsymbol{y}|\boldsymbol{w}) = -\sum_{i=1}^{m} \{y_i \ln \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)_i + (1-y_i) \ln(1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)_i)\}$
3. Optimisation par ***descente de gradient***

# Choix de $\mathcal{H}$

## Modèles probabilistes discriminatifs

- **Estimer directement** $p(\mathcal{C}_k|\boldsymbol{x})$ en utilisant un modèle paramétré

**Exemple** : Deux classes et régression logistique

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})} = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$$

1. Fonction de vraisemblance : $p(\boldsymbol{y}|\boldsymbol{w}) = \Pi_{i=1}^m \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)^{y_i} \left\{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right\}^{(1-y_i)}$
2. Fonction d'erreur :
$E(\boldsymbol{w}) = -\ln p(\boldsymbol{y}|\boldsymbol{w}) = -\sum_{i=1}^m \{y_i \ln \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)_i + (1-y_i)\ln(1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)_i)\}$
3. Optimisation par ***descente de gradient***

Moins de paramètres à estimer.

# Choix de $\mathcal{H}$

## Fonctions de décision à base de dictionnaire

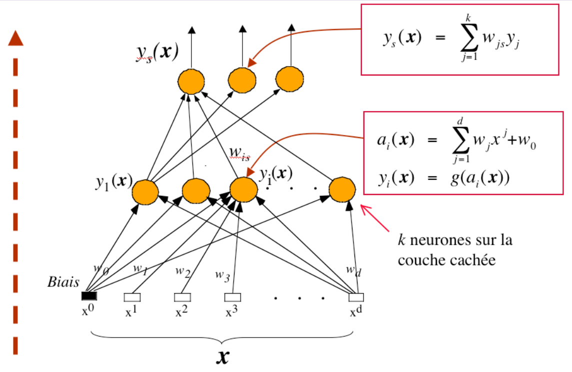
- $h(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{n} w_i \, g_i(\boldsymbol{x}) + w_0$
- où les $g_i(\boldsymbol{x})$ sont des ***fonctions de base***

# Choix de $\mathcal{H}$

## Fonctions de décision à base de dictionnaire

- $h(\boldsymbol{x}, \boldsymbol{w}) \ = \ \sum_{i=1}^{n} w_i \, g_i(\boldsymbol{x}) + w_0$
- où les $g_i(\boldsymbol{x})$ sont des ***fonctions de base***

**Exemple** : Perceptron multi-couches



# Choix de $\mathcal{H}$

## Fonctions de décision par noyaux

- $h(\boldsymbol{x}) \;=\; \sum_{i \text{ ``critiques''}} \alpha_i \, y_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) + \alpha_0$
- où les $K_i(\cdot,\cdot)$ sont des ***fonctions noyaux***

# Choix de $\mathcal{H}$

## Fonctions de décision par noyaux

- $h(\boldsymbol{x}) \;=\; \sum_{i \text{ ``critiques''}} \alpha_i \, y_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) \,+\, \alpha_0$
- où les $K_i(\cdot,\cdot)$ sont des ***fonctions noyaux***

$$K_G(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{||\boldsymbol{x}, \boldsymbol{x}_i||^2}{2\,\sigma^2}\right)$$

$$K_L(\boldsymbol{x}, \boldsymbol{x}_i) = \boldsymbol{x}^\top \boldsymbol{x}_i$$

$$K_{Poly1}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i)^d$$

$$K_{Poly2}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i + c)^d$$

$$K_{sig}(\boldsymbol{x}, \boldsymbol{x}_i) = \tanh(\kappa \boldsymbol{x}^\top \boldsymbol{x}_i + \theta)$$

# Choix de $\mathcal{H}$

## Fonctions de décision par noyaux

- $h(\boldsymbol{x}) \;=\; \sum_{i \text{ ``critiques''}} \alpha_i \, y_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) + \alpha_0$
- où les $K_i(\cdot,\cdot)$ sont des ***fonctions noyaux***

$$K_G(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{||\boldsymbol{x}, \boldsymbol{x}_i||^2}{2\,\sigma^2}\right)$$

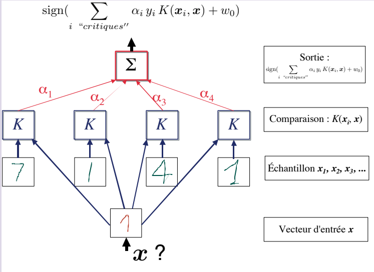
$$K_L(\boldsymbol{x}, \boldsymbol{x}_i) = \boldsymbol{x}^\top \boldsymbol{x}_i$$

$$K_{Poly1}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i)^d$$

$$K_{Poly2}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i + c)^d$$

$$K_{sig}(\boldsymbol{x}, \boldsymbol{x}_i) = \tanh(\kappa \boldsymbol{x}^\top \boldsymbol{x}_i + \theta)$$

**Exemple** : Séparateurs à Vastes Marges (SVM)

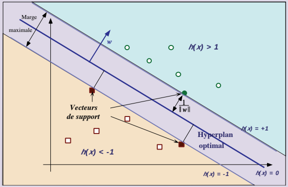


# Choix de $\mathcal{H}$

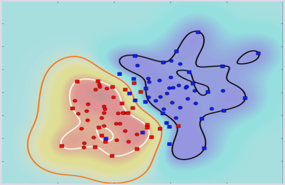
## Fonctions de décision par noyaux : Les SVM

$$h^*(\mathbf{x}) \;=\; (\boldsymbol{w}^\star\;\mathbf{x}) \,+\, w_0^\star \;=\; \sum_{i=1}^{m} \alpha_i^\star\, u_i \,.\, \langle \phi(\mathbf{x}_i)\,,\, \phi(\mathbf{x})\rangle \,+\, w_0^\star$$

### SVM : espace des redescripteurs $\Phi(\mathcal{X})$




### SVM : espace initial $\mathcal{X}$

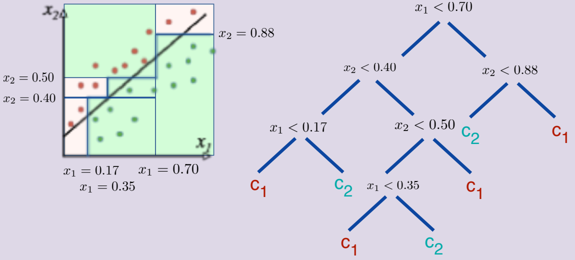


# Choix de $\mathcal{H}$

## Fonctions de décision par spécialisation

$$
\begin{aligned}
h_{\mathcal{X}} &= h_{\mathcal{X}_1} \times h_{\mathcal{X}_2} \times \ldots \times h_{\mathcal{X}_n} \\
\mathcal{X} &= h_{\mathcal{X}_1} \cup h_{\mathcal{X}_2} \cup \ldots \cup h_{\mathcal{X}_n}
\end{aligned}
$$

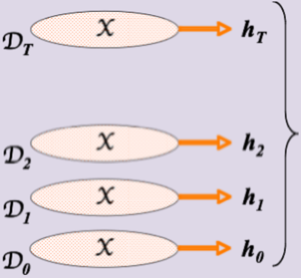
**Exemple** : Arbres de décisions



# Choix de $\mathcal{H}$

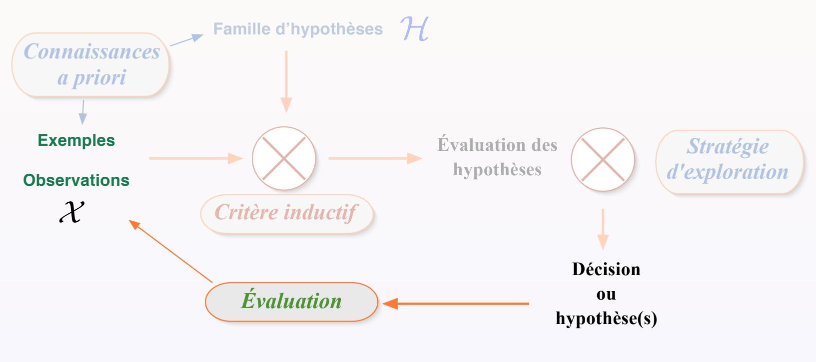
## Fonctions de décision par vote

**Exemple** : Boosting




$$h(\boldsymbol{x}) \;=\; \operatorname{sign}\left[\sum_{t=0}^{T} \alpha_t \cdot h_t(\boldsymbol{x})\right]$$

# L'apprentissage
## Ingrédients



# Les données

## Caractéristiques à prendre en considération

### Représentativité

- $P_{\mathcal{X}}$ biaisé : *classe sous échantillonnée*
- $P_{\mathcal{Y}|\mathcal{X}}$ biaisé : *bruit ; dérive de concept*

### Dimension de l'espace

- Réduction de dimension

### Déséquilibre des classes

- Rééquilibrer

### Déséquilibre faux positifs / faux négatifs

- Techniques ad hoc

# Évaluation des résultats

## Précision en prédiction

Précision ; Rappel ; Validation croisée ; ... ; Courbe ROC ; test d'hypothèses

Souci principal : l'*overfitting*

# Évaluation des résultats

## Précision en prédiction

Précision ; Rappel ; Validation croisée ; ... ; Courbe ROC ; test d'hypothèses

Souci principal : l'*overfitting*

## Avec degré de confiance ou de probabilité

- Fct de coût adaptée (e.g. entropie croisée)
- Distance à la frontière de décision
- ...

# Évaluation des résultats

## Précision en prédiction

Précision ; Rappel ; Validation croisée ; ... ; Courbe ROC ; test d'hypothèses

Souci principal : l'*overfitting*

## Avec degré de confiance ou de probabilité

- Fct de coût adaptée (e.g. entropie croisée)
- Distance à la frontière de décision
- ...

## Interprétabilité de l'hypothèse

- Symbolique
- Simplicité

# Évaluation des résultats

## Contrôle de l'overfitting

### En **théorie**

- Fonctions de décision
  - Mesure de capacité ($d_{VC}$, ...)
- Modèles génératifs
  - **AIC** : Pénalisation = 2 $\times$ nb de paramètres libres
    (suppose que le modèle est correct, sinon modèle trop complexe)
  - **BIC** : Pénalisation = $\log m \times$ nb de paramètres libres

# Évaluation des résultats

## Contrôle de l'overfitting

### En **théorie**

- Fonctions de décision
  - Mesure de capacité ($d_{VC}$, ...)
- Modèles génératifs
  - **AIC** : Pénalisation = 2 $\times$ nb de paramètres libres
    (suppose que le modèle est correct, sinon modèle trop complexe)
  - **BIC** : Pénalisation = $\log m \times$ nb de paramètres libres

### En **pratique**

Validation croisée

- Nombre de **paramètres** (ou AIC ou BIC)
- Nombre de **concepts** (apprentissage relationnel ; arbre de décision)
- Nombre d'**itérations** (RNs, Boosting)

# Plan

1. Les fondements
2. En pratique
3. Conclusions et perspectives
   - Le paradigme et ses frontières
   - Les directions

# Autres tâches d'apprentissage

1. Non supervisé
2. Semi-supervisé
3. Transduction
4. Apprentissage de tri (*ranking*)
5. Apprentissage de recommandations
6. Apprentissage de politique d'action (par renforcement)
7. Apprentissage à partir d'explications
8. ...

# Le paradigme ... et ses limites

- Lien entre passé et futur :

  distributions $P_{\mathcal{X}}$ et $P_{\mathcal{Y}|\mathcal{X}}$ supposées stationnaires

- Données i.i.d.

# L'avenir

## Les tendances

### Dans le paradigme

- Très grosses Bases de Données
- Recherche de "Deep models" (modèles causaux hiérarchiques)
- Apprentissage et jeux (adversaire ; transfert)

# L'avenir
## Les tendances

### Dans le paradigme

- Très grosses Bases de Données
- Recherche de “Deep models” (modèles causaux hiérarchiques)
- Apprentissage et jeux (adversaire ; transfert)

### Renouvellement

- **Apprentissage-en-ligne et flux de données**
  - Pas de stockage possible : “one-pass learning”
  - Dérive de concept possible
  - Information dans la séquence d'exemples
- **Rationalité limitée**
  - *Intelligence ambiante*

# L'avenir ...

... commence ici !

MERCI !