IFD1, 2011-2012

Initiation à la Fouille de Données et à l'Apprentissage

Valentin Emiya*,

François-Xavier Dupé,

Pierre Machart

prénom.nom@lif.univ-mrs.fr

http://www.lif.univ-mrs.fr/~vemiya/teaching/

Cours largement inspiré par ceux de R. Eyraud, F. Denis et L. Miclet.

IFD1 : objectifs

- Aborder l'apprentissage automatique comme élément central de la FD
- Assimiler les notions et outils de base en apprentissage automatique
- Mise en pratique : apprentissage de SAS
- Préparer le module AN

Organisation de l'UE

- Enseignements
  - 7 cours-TD de 4h (VE & FXD)
  - 7 TP de 2h (PM)
- Évaluation
  - TP noté
  - Examen : 2h, une feuille recto-verso manuscrite
- Site web : transparents, sujets TD/TP, poly de F. Denis, etc.

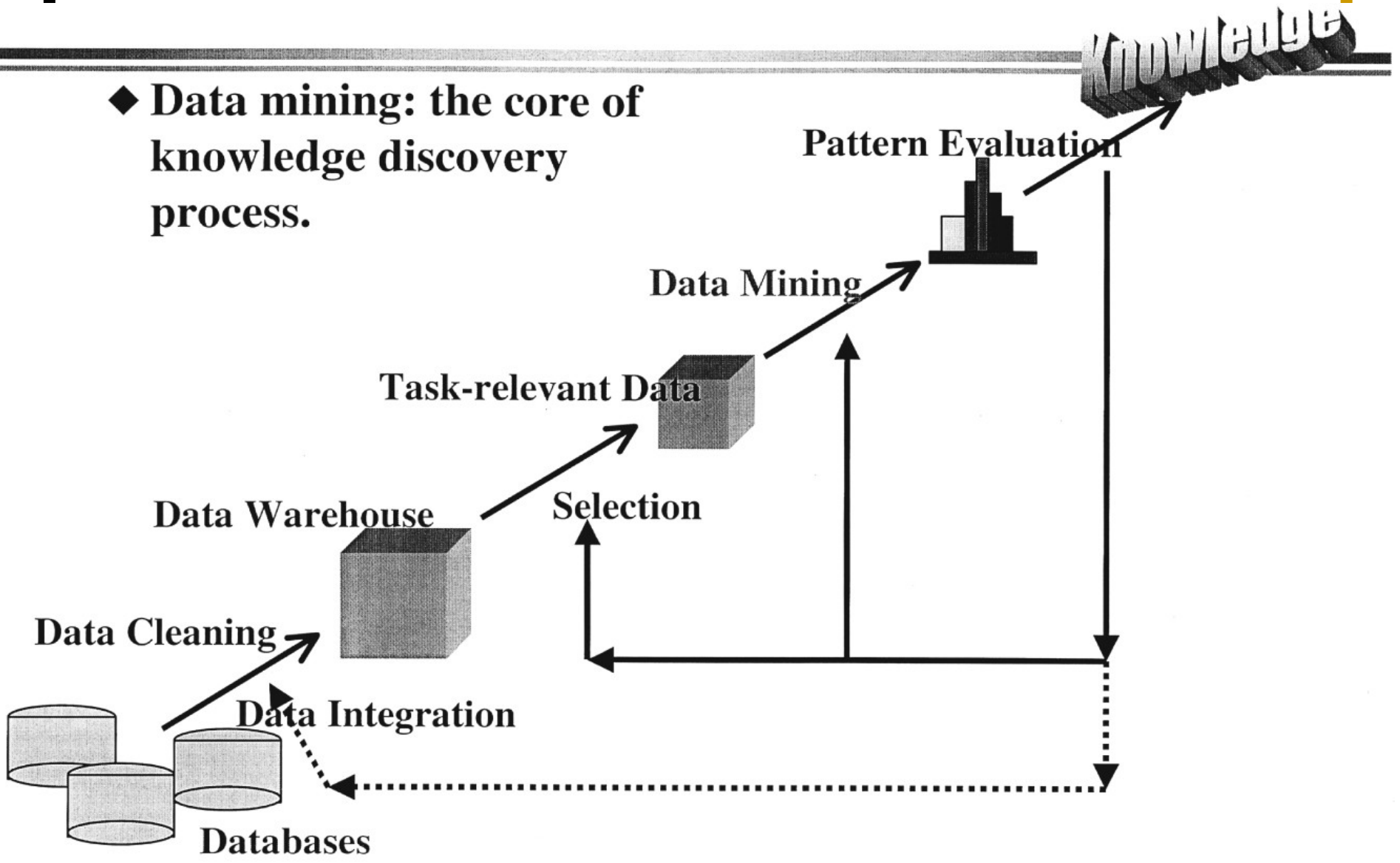
Plan général du cours

- Fouille de données et apprentissage
- Exemples introductifs
- Les différents types d'apprentissage
- Modèle de l'apprentissage statistique
  - Classification, apprentissage de concept
  - Régression, estimation de densité (CM2)
- Minimisation du risque empirique (CM2)
- Les arbres de décision (CM3-4)
- Apprentissage et statistiques (CM5)
- Régression linéaire et logistique (CM6)

La fouille de données et l'apprentissage automatique

Schéma général

- **Data mining: the core of knowledge discovery process.**



Source: Han and Kamber (2001)

L'apprentissage automatique

- « Moteur » de la fouille
- Tirer des règles de portée générale à partir d'observations particulières.
- Quelques exemples d'objectifs :
  - Apprendre à filtrer l'information
  - Apprendre les préférences d'un utilisateur
  - Apprendre à faire des résumés
  - Découverte scientifique (découverte de régularités en biochimie, …)
  - ...

2 types d'apprentissage

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

2 types d'apprentissage

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

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

Un exemple introductif

Alors que vous venez juste d'atterrir au Groland pour la première fois, vous apercevez un mouton noir. Quelles conclusions en tirer ?

- Au Groland, il existe un mouton dont une partie du corps est noir (fait)
- Il y a un et un seul mouton noir au Groland (apprentissage par coeur, *overfitting*)
- Certains moutons sont noirs au Groland
- Tous les moutons du Groland sont noirs (surgénéralisation)

Un deuxième exemple

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

Un deuxième exemple

- Quelques solutions valides :
  - a=6. Argument : c'est la suite des entiers sauf 4.
  - a=7. Argument 1 : c'est la suite des nombres premiers.
  - a=7. Argument2 : suite binaire 1(1), 10(2), 11(3), 101(5), 111(7), 1011(11), 1111(15), 10111(23), 11111(31)...
  - a=8. Argument : c'est la suite de Fibonacci.
  - a=2π. Argument : la liste ordonnée des racines du polynôme : $x^5 - (11+a)x^4 + (41+11a)x^3 - (61-41a)x^2 + (30+61a)x - 30a$ qui est le développement de $(x-1)(x-2)(x-3)(x-5)(x-a)$

    ⇒ a peut être n'importe quel nombre réel supérieur ou égal à 5)
- Généralisation : il est facile de montrer que n'importe quel nombre est la suite d'une séquence de nombre...

De vrais exemples (1)

Classification supervisée

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

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

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

But : reconnaissance de chiffres manuscrits.

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

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

De vrais exemples (2)

Classification non-supervisée (*clustering*)

But : Identifier des profils parmi les clients d'une entreprise, les usagers des transports en commun ou les spectateurs d'une chaîne de télévision : la fameuse ménagère de plus de 40 ans, les bobos, les couples en voie d'acheter un bien immobilier, les “grands voyageurs", ...

Données : Base de données clients, résultats médiamétrie, ...

De vrais exemples (3)

Régression supervisée

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

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

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

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

De vrais exemples (4)

Estimation de densité

But : Différencier deux auteurs à partir des documents qu'ils ont produits pour estimer la probabilité qu'un nouveau document ait été produit par l'un ou l'autre auteur (par exemple en étudiant les fréquences de mots apparaissant dans leurs œuvres).

But : Déterminer si la distribution des nucléotides est la même dans les parties codantes et les parties non codantes d'un gène.

But : Caractériser une famille de protéines par la distribution des acides aminés

Références Bibliographiques

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

Modélisation de l'apprentissage (supervisé)

Modéliser : quoi ? comment ?

- Comment modéliser les données ?

  → *distribution statistique*

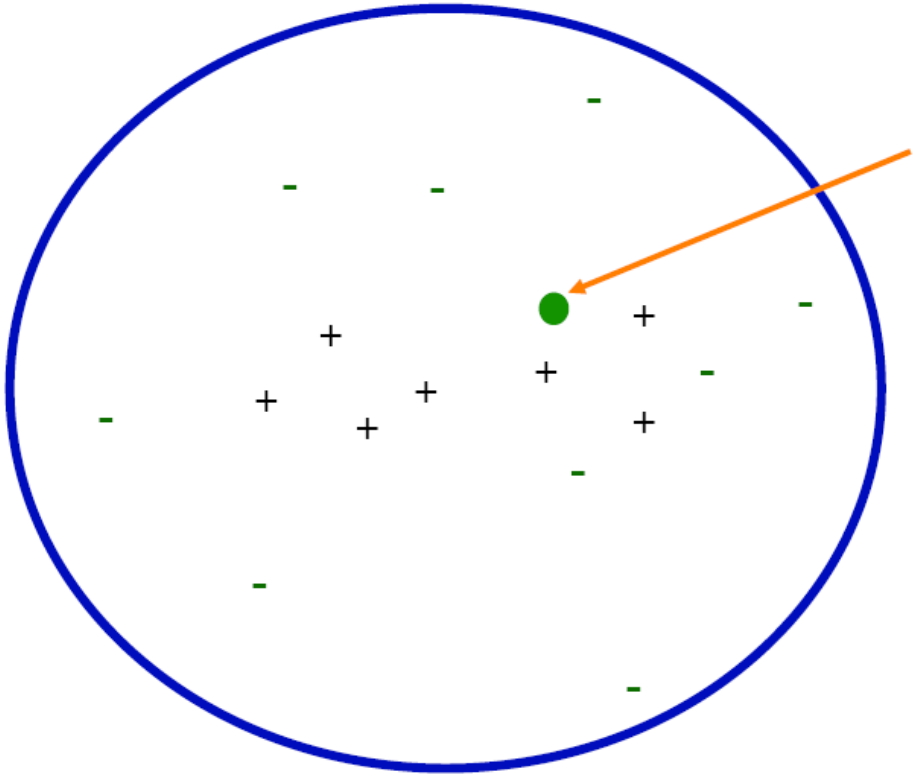
- Que veut-on apprendre ?

  → *une fonction*

- Comment évaluer la performance de l'apprentissage obtenu ?

  → *notion de risque (=erreur)*

Exemple : apprentissage de concept



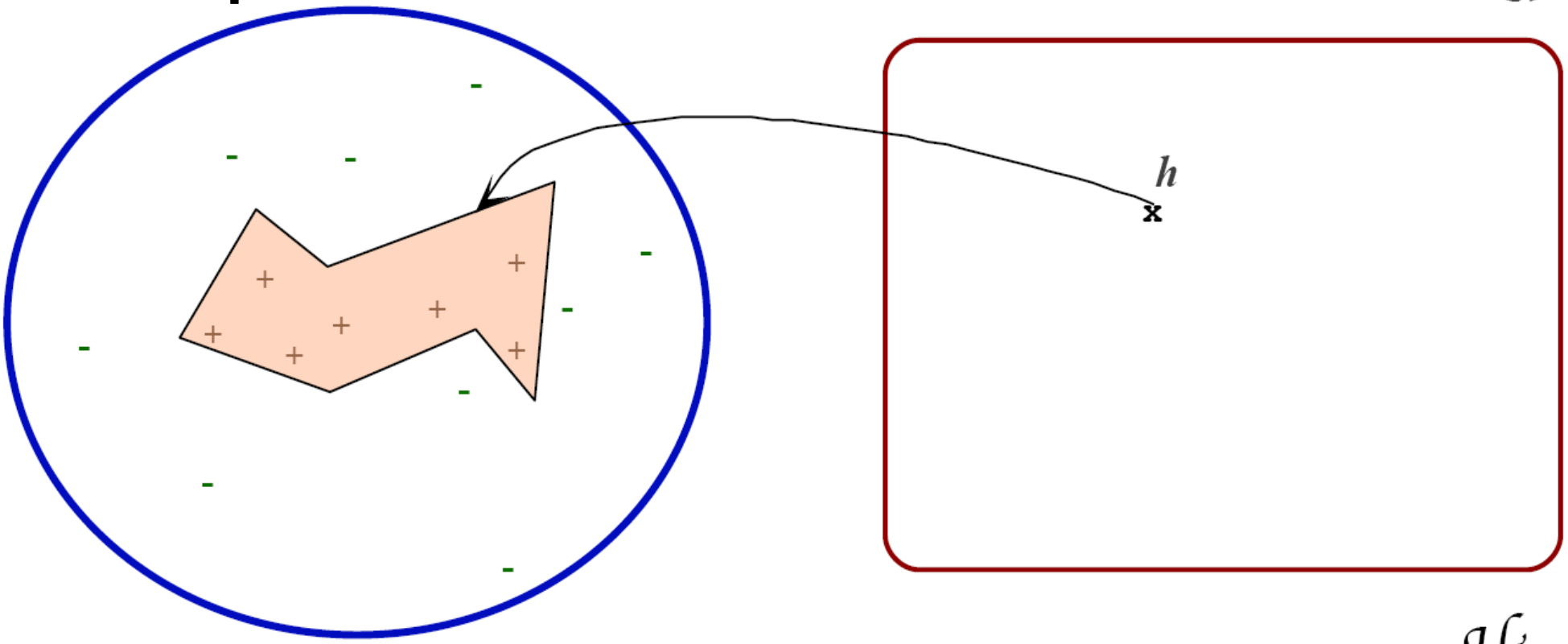
- Méthodes par ***plus proches voisins***
- Nécessité d'une *notion de distance*

Espace des exemples : $\mathcal{X}$

→ **Hypothèse de continuité dans $\mathcal{X}$**

Apprentissage : un jeu entre espaces

$\mathcal{L}_{\mathcal{H}}$

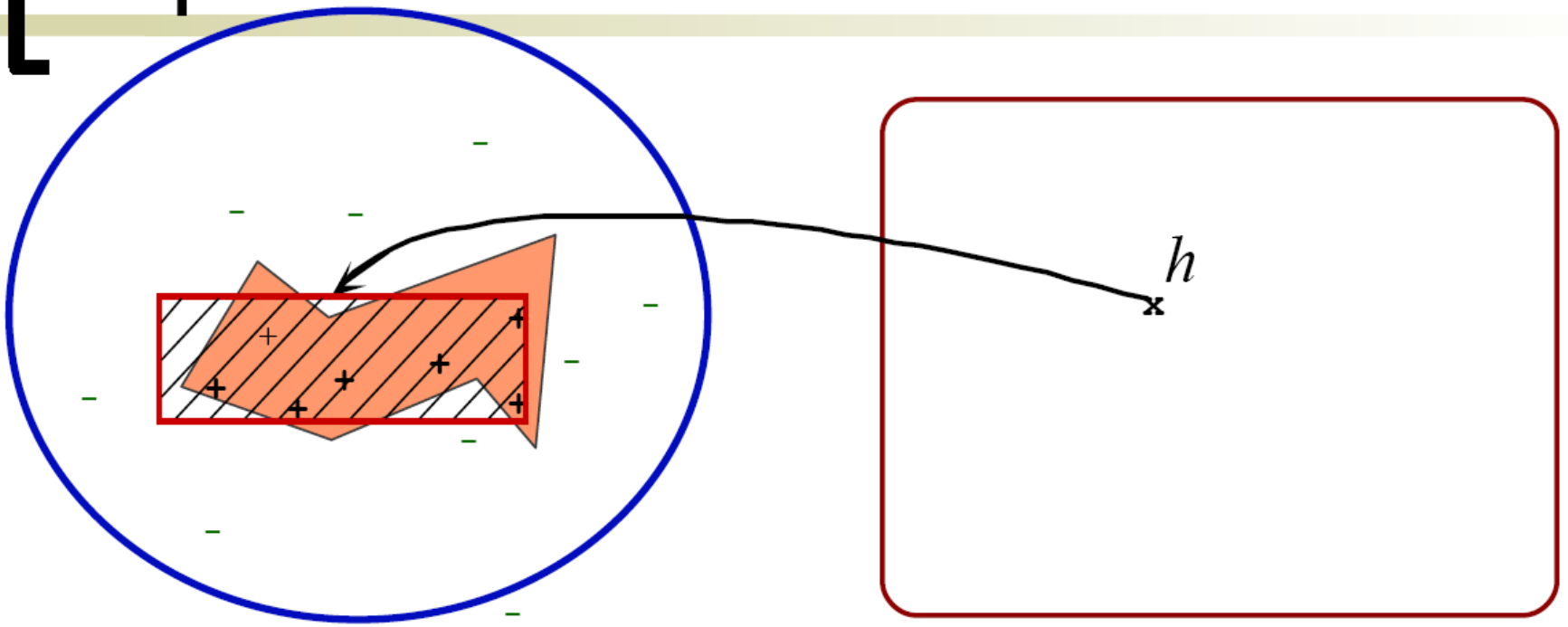


Espace des hypothèses : $\mathcal{H}$

Espace des exemples : $\mathcal{X}$

↙ Comment choisir l'espace des hypothèses (i.e. le langage $\mathcal{L}_{\mathcal{H}}$) ?

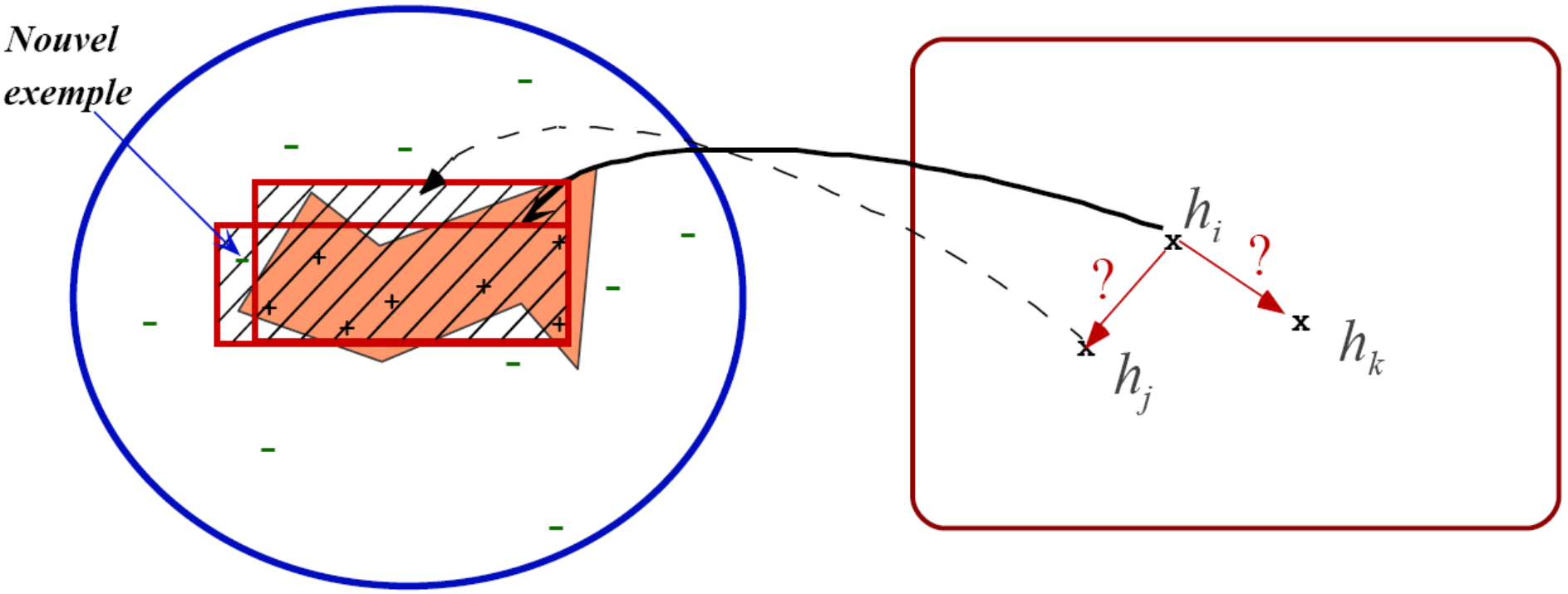
Apprendre = un jeu entre espaces



Espace des exemples : $\mathcal{X}$ Espace des hypothèses : $\mathcal{H}$

- Comment choisir une hypothèse ?
- Quel **critère inductif** ?

Apprendre = un jeu entre espaces



Espace des exemples : $\mathcal{X}$ Espace des hypothèses : $\mathcal{H}$

↙ Comment explorer l'espace des hypothèses ?

Modélisation : les données

- **Notations :**
  - **entrées** : vecteurs de n **attributs**

$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{X} = \mathbf{X}_1 \times \ldots \times \mathbf{X}_n$$

  - **sorties** $y \in \mathbf{Y}$

Exemple

Entrée à $n = 2$ attributs :

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

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

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

Modélisation : les données

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

  variables aléatoires $(X, Y) \in \mathbf{X} \times \mathbf{Y}$
  distribuées selon $P(X, Y)$ (inconnu en g^{al})
- Base d'apprentissage, « **échantillon** » :
  ensemble de l **exemples** $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$
  avec $(\mathbf{x}_i, y_i)$ **i.i.d.** selon $P(X, Y)$

Principe de l'apprent. supervisé

- **À partir des données**

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

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

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

- [Si $\mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$, facile : apprent. par coeur.]
- Défi : **généraliser** pour tout $\mathbf{x} \in \mathbf{X}$
- h est aussi appelée aussi hypothèse, règle, classifieur, etc.

Principe de l'apprent. supervisé

- **À partir des données**

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

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

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

Exemple

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

Performance : notion de risque

- **Fonction de perte** / *loss function (classif.)* :

$$L(y, h(\mathbf{x})) = 1 \text{ si } y \neq h(\mathbf{x}),\ 0 \text{ sinon.}$$

(rq : autres fonctions possibles)

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

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

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

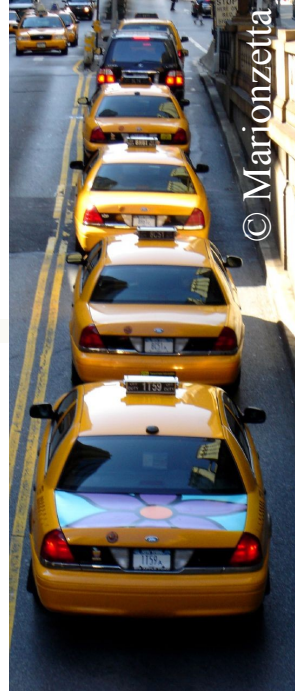
Synthèse

Problème général de l'apprentissage supervisé :

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

Risque et tirage iid - Exemple

	Taxis (T)	Autres voitures (A)
Manhattan	1/3, 100% jaunes (J)	5% J, 95% pas J
Brooklyn	1/5, 100% jaunes (J)	5% J, 95% pas J
Londres	1/10, 100% noirs (N)	20% N, 80% pas N



Fonction de détection des taxis par la couleur :

$$\text{pour une couleur } c,\ h_c(x) = \begin{cases} T & \text{si } x = c \\ A & \text{sinon} \end{cases}$$

Calculer le risque de h_J et h_N à chaque endroit et commenter.

Règles de classification

Règle majoritaire

- Pour toute nouvelle instance, retourner la classe y_{maj} majoritaire, c'est-à-dire pour laquelle P(y) est maximum :

$$\text{pour } \mathbf{x} \in \mathbf{X}, f_{\text{maj}}(\mathbf{x}) = \text{argmax}_y P(y) = y_{\text{maj}}$$
$$R\left(f_{\text{maj}}\right) = 1 - P\left(y_{\text{maj}}\right)$$

- Rq : $f_{maj}(x)$ ne dépend pas de x !
- Exemple : sur un échantillon de 100 étudiants, 80 ont un portable, 10 un fixe et 10 rien du tout. En utilisant la règle majoritaire, on considérera qu'un nouvel étudiant possède un portable avec un risque de 20%.

Règle du maximum de vraisemblance

- *Maximum likelihood*
- Pour chaque instance x, retourner la classe y pour laquelle x est la valeur la plus observée :

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

- Souvent utilisée...
- Mais ne peut pas être estimée à partir de l'échantillon sans hypothèse supplémentaire.

Règle de classification de Bayes

- Pour chaque instance x, retourner la classe y dont l'observation est la plus probable, ayant observé x :

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

- La règle de Bayes est la règle de risque minimal [demonstration].
- Ne peut pas être estimée à partir de l'échantillon.

Un exemple : énoncé

- Une banque souhaite proposer une offre à certains de ses clients : gérer son compte par internet. Comment cibler les clients le plus susceptibles d'êtres intéressés ? Elle se rend compte que certains indiquent une adresse e-mail : c'est peut-être un critère sur lequel baser le mailing. Un sondage réalisé sur un échantillon supposé représentatif de sa clientèle, indique que :
  - 40% sont intéressés parmi lesquels 80% ont indiqué leur e-mail,
  - 60% ne sont pas intéressés parmi lesquels 40% ont indiqué leur e-mail.
- Modélisation ?

Un exemple : modélisation

- Modèle :
  - Données : les clients
  - Attribut(s) des données : un seul dont le domaine est ***X***= {email, ~~email~~},
  - Classes : deux : ***Y***={interesse,~~interesse~~}.

- Règle majoritaire :

  La majorité des clients n'est pas intéressée : la règle retourne ~~intéressé~~ pour tout instance (risque de 0.4).

Un exemple : règle de maximum de vraissemblance

- Il nous faut regarder P(x|y) :

  On a P(email|intéressé)=0.8 > P(email|~~intéressé~~). D'après la règle, un client ayant indiqué un email est intéressé.

  D'autre part, P(~~email~~|intéressé) < P(~~email~~|~~intéressé~~) : la règle indique qu'un client n'ayant pas donné d'email n'est pas intéressé.

- Le risque est : P(~~email,~~interéressé) + P(email,~~intéressé~~)=0.4*0.2 + 0.6*0.4 = 0.32

Un exemple : règle de Bayes

- Il nous faut évaluer P(y|x) :

  Pour cela il nous faut P(email)=0.8 * 0.4 + 0.6 * 0.4 = 0.56. Du coup :

  P(intéressé|email) = P(email|intéressé)*P(intéressé)/P(email) = 32/56 = 4/7 > 0,5 > P(~~intéressé~~|email) donc la règle décide qu'un client avec un email est intéressé

  De la même manière, P(intéressé | ~~email~~) = 4/11 < 0,5 : le classifieur décide que si un client n'a pas d'email il n'est pas intéressé.

- Les décisions étant les mêmes que la règle précédente, le risque est identique (=0,32).

Plan du cours

- Fouille de données et apprentissage
- Exemples introductifs
- Les différents types d'apprentissage
- Modèle de l'apprentissage statistique
  - Classification, apprentissage de concept
  - Régression, estimation de densité (CM2)
- Minimisation du risque empirique (CM2)
- Les arbres de décision (CM3-4)
- Apprentissage et statistiques (CM5)
- Régression linéaire et logistique (CM6)