Inférer sans loi

Jean-Marc Brossier

Université Grenoble Alpes / Grenoble INP / Ensimag.

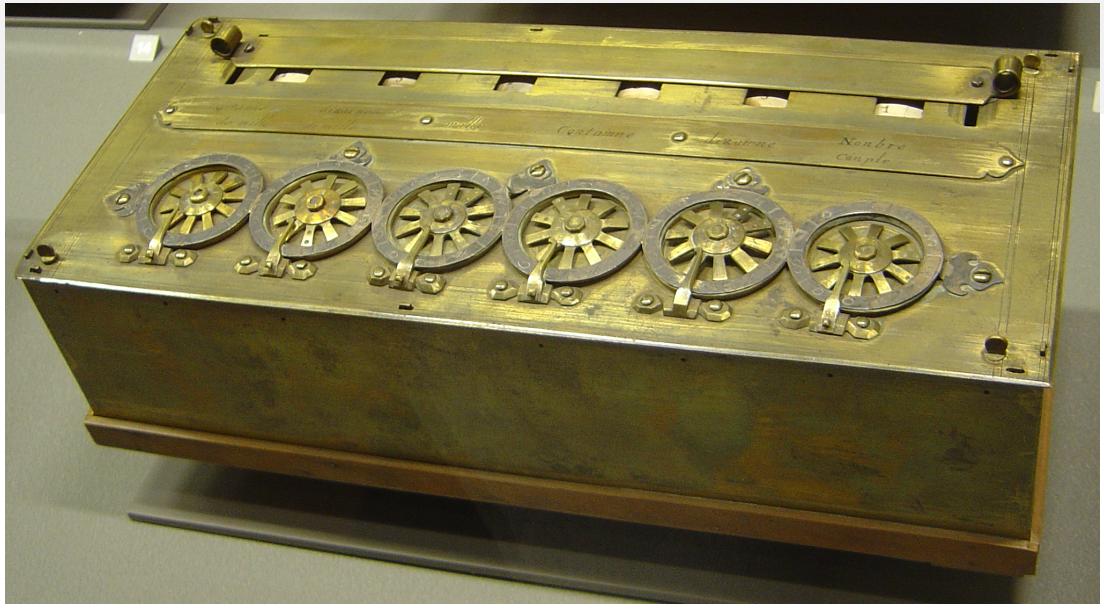
GIPSA-lab (Grenoble Images Parole Signal Automatique) /
GAIA (Géométrie, Apprentissage, Information et Algorithmes)

AFSCET, Paris

25 novembre 2022

Outline

Pascaline



Au musée des arts et métiers du Conservatoire National des Arts et Métiers à Paris, signée par Pascal en 1652.
`https://fr.wikipedia.org/wiki/Pascaline`

La soeur de Pascal : «mon frère a capturé l'esprit»

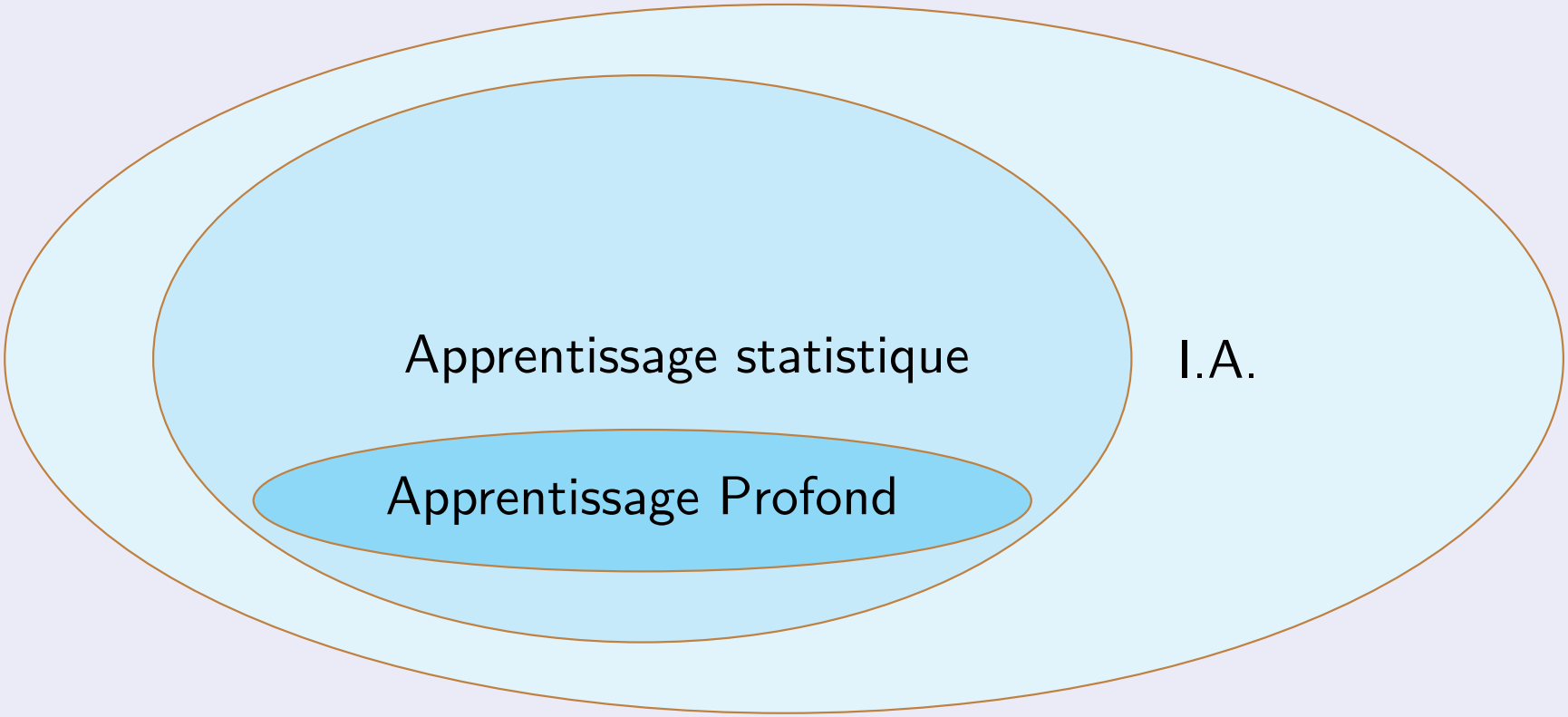
Qu'est-ce que l'intelligence ? Parmi 1000 tentatives :
« la capacité de résoudre par la pensée des problèmes nouveaux. »
Edouard Claparède

AI - Machine Learning - Deep Learning

Intelligence artificielle

⊃ Apprentissage statistique (Machine Learning)

⊃ Réseaux de neurones profonds (Deep Learning)



Ce n'est pas (ou pas seulement)

Des visions de la science-fiction.

`https://www.bostondynamics.com/atlas`

à rapprocher de terminator (sorti en 1984...).

Les méthodes se sont développées au travers de nombreuses applications spécialisées à la fois moins spectaculaires et plus omniprésentes :

- Reconnaissance des caractères (OCR - Optical Character Recognition) dans les années 50 (David H. Shepard),
- Filtres anti spam (du nom d'un personnage de sketch des Monty Python datant de 1970) dans les années 1990.
- De très nombreuses applications spécialisées ont suivi : reconnaissance vocale, assistants vocaux, reconnaissance de l'écriture, des images...

C'est

Une manière particulière de programmer une machine pour qu'elle apprenne à partir des données.

A. Samuel (1959)

Apprendre des données sans programmation explicite.

Tom Mitchell (1997)

Une machine est à même d'apprendre d'une expérience E au sujet d'une tâche T si sa performance P sur la résolution de cette tâche T augmente au fil des expériences E.

C'est

D'une certaine façon,
apprendre transforme des expériences en une expertise.

- L'expérience, ce sont les données (mails étiquetés ou pas, par ex.).
- L'expertise c'est l'algorithme que l'apprentissage produit : capable, à partir de nouvelles données, de prendre une décision (e.g. spam/non-spam)

Des questions se posent sur

- la faisabilité,
- les données (quantité, qualité, pertinence, etc.),
- l'évaluation des performances de la solution obtenue,
- les ressources nécessaires à l'apprentissage,
- ⋮

Pourquoi utiliser de l'apprentissage statistique ?

Complexité ou absence de solution connue :

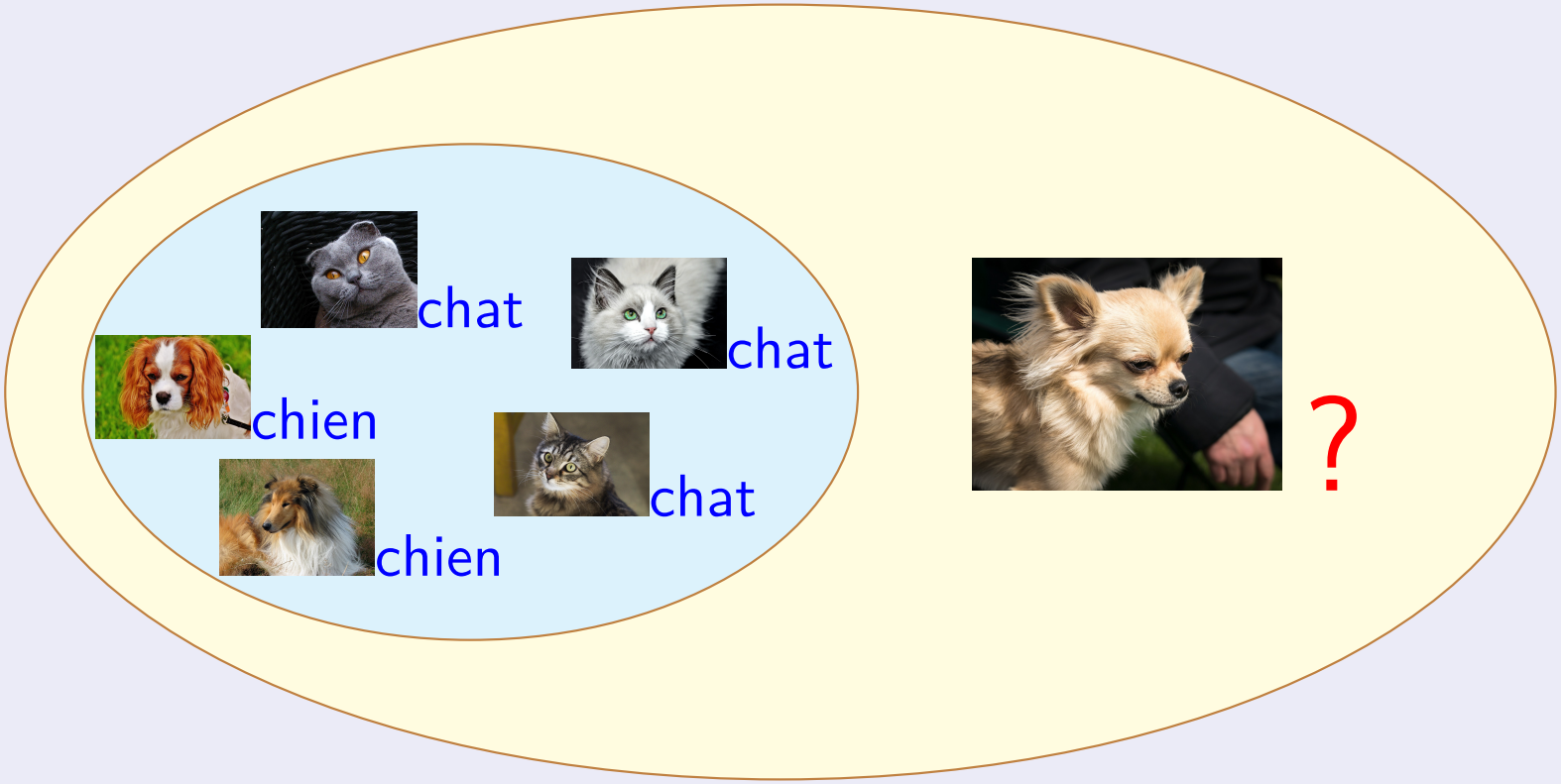
- Distinguer deux mots en utilisant leurs caractéristiques vocales est facile, mais ne passe pas à l'échelle.
  Il est préférable d'apprendre à partir d'une très grande base de mots prononcés par différents locuteurs.
- Même si nous pouvons reconnaître un spam au premier coup d'œil, il est difficile d'expliciter les règles qui nous ont amenés à prendre la décision.

Volumes : les volumes de données dépassent les capacités humaines : exemple des données génétiques.

Adaptativité : Un programme ML peut s'adapter à un environnement changeant. L'utilisateur tague des messages en tant que SPAM, corrige le texte produit par reconnaissance vocale, corrige la reconnaissance d'écriture manuscrite.

Base d'exemples d'apprentissage

On dispose d'exemples (images) de chiens et de chats.
Une nouvelle image arrive, chien ou chat ?



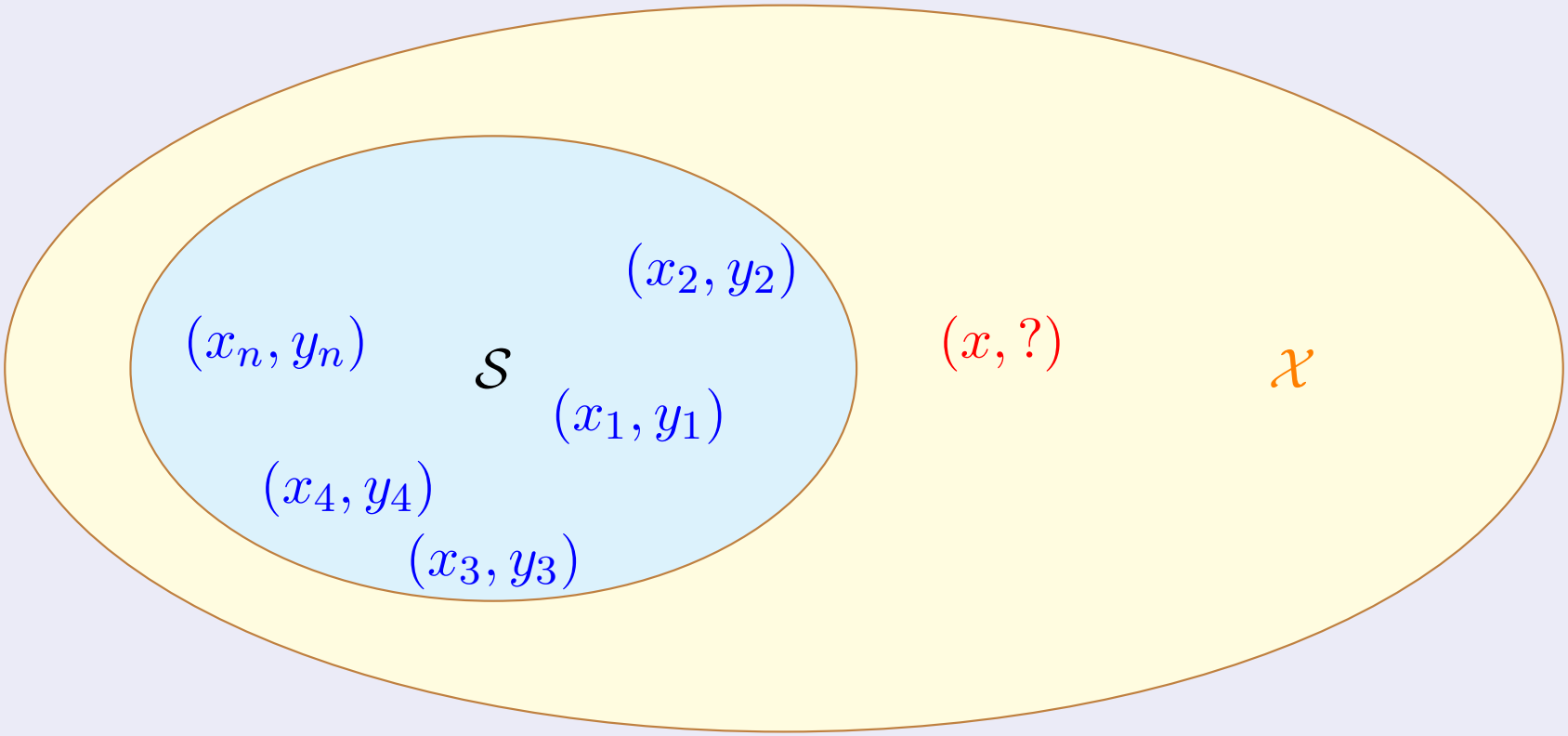
Base d'exemples d'apprentissage

On dispose de n exemples x_i dont l'étiquette y_i est connue :

$$\mathcal{S} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, 1 \leq i \leq n\}.$$

Ici $x \in \mathcal{X} = \mathbb{R}^p$ et $y \in \{-1, +1\}$

But : déterminer l'étiquette y d'exemples x nouveaux (hors de $\mathcal{S}$)



Etiquettes des exemples

A un exemple, de caractéristique $x \in \mathcal{X}$ est associée une étiquette $y \in \mathcal{Y}$.
2 cas :

Classification : les étiquettes sont en nombre fini, ex. $y \in \mathcal{Y} = \{-1, 1\}$.

Régression : les étiquettes sont des réels $y \in \mathcal{Y} = \mathbb{R}$.

Etiquetage

2 cas :

1. Etiquetage dur.
   L'étiquette peut être évaluée en fonction des caractéristiques :
   il existe une fonction f t.q. $\forall x \in \mathcal{X}$, $y = f(x)$
2. Etiquetage souple.
   La loi porte sur les couples (catactéristique, étiquette).

Superstition des pigeons (Skinner, 1947)

Dans une cage, on place des pigeons avec un grand nombre de jouets différents à leur disposition.

En haut de la cage, un distributeur de graines s'active à des instants aléatoires. Les pigeons sont ainsi amenés à faire une association trompeuse entre le fait qu'il joue avec un jouet et le fait que des graines tombent.

Petit à petit, le pigeon se concentre sur un seul jouet qu'il a associé au fait que des graines tombent.

Un rat averti n'en vaut pas deux (Garcia, 1996)

Dans une expérience menée avec des rats, des aliments d'aspect, de goût et d'odeur habituels sont mis à la disposition d'un rat. Sans aucune différence sur ces critères, certains des aliments sont "empoisonnés".

Au moment où un rat s'approche d'un mauvais aliment, une sonnette retentit.

Les rats ne font jamais d'association entre le bruit produit par la sonnette et le fait que l'aliment est empoisonné : le son ne semble pas faire partie des caractéristiques pertinentes pour le rat quant à la sélection d'un bon aliment.

Choix des caractéristiques pertinentes

Réduire la dimension. La grande dimension est source de difficultés : un bruit faible par dimension devient globalement fort, exemples sans plus proches voisins pour un nombre d'exemples raisonnable, certaines coordonnées peuvent être inutiles voire nuisibles.

Choisir des caractéristiques pertinentes. Prêter attention à tout ? Non, tous les cas seraient des cas particuliers. Pour choisir, utiliser la connaissance d'un expert, une modélisation du problème,...

Filtrer les caractéristiques : choix des plus liées aux étiquettes. Mesure du lien par la corrélation ou l'information mutuelle par exemple.

Extraire des caractéristiques : construire un ensemble de fonctions des caractéristiques qui capturent un maximum d'information pertinente quant à la prise de décision.

Choix des caractéristiques avant ou pendant l'apprentissage.

Algorithme d'apprentissage

Un **algorithme d'apprentissage** $\mathcal{A}$ prend en entrée $\mathcal{S}$ et produit en sortie un prédicteur (aussi appelé hypothèse, classifieur) $h_{\mathcal{S}}$

$$\mathcal{A} : \mathcal{S} \to h_{\mathcal{S}} \text{ avec } h_S : \mathcal{X} \to \mathcal{Y}$$

Mesure de performance par un risque
Si $\mathcal{S}$ est échantillonné selon une loi $\mathcal{D}$, le **risque** $\mathcal{L}$ est défini par la probabilité d'**erreur de prédiction théorique** :

$$\mathcal{L}_{\mathcal{D}}(h) = \mathcal{D}\left\{h(\mathbf{x}) \neq y\right\}$$

Le meilleur prédicteur sera celui dont le risque est le plus faible.

On peut supposer une loi sur $\mathcal{X}$ et l'existence d'une fonction d'étiquetage t.q. $y = f(x)$ (étiquetage dur) ou que la loi porte sur $\mathcal{X} \times \mathcal{Y}$ (étiquetage souple).

Deux cadres classiques : Bayésien et PAC.

Prédire un dé

A priori : prédiction

Prédire un dé équilibré

- Faces de même probabilité
- On ne sait pas du tout,
- Maximum d'entropie

Prédire un dé pipé sachant comment il est pipé

Prédire la face la plus probable minimise la probabilité d'erreur.

Prédire un dé

A priori : prédiction

Prédire un dé équilibré

- Faces de même probabilité
- On ne sait pas du tout,
- Maximum d'entropie

Prédire un dé pipé sachant comment il est pipé

Prédire la face la plus probable minimise la probabilité d'erreur.

Prédire un dé

A posteriori : correction

Prédire un dé pipé sachant qu'il est pipé et que le résultat est partiellement connu.

Utilisation du théorème de Bayes :

Prise en compte de l'*a priori* (comment le dé est pipé)
& Correction fonction d'une observation (e.g. la face est paire)

Algorithmes fondés sur Bayes

Dès qu'il y a connaissance *a priori* et observation.

En dehors de l'apprentissage statistique :

- Stratonovich–Kalman-Bucy (1960 après Thorvald N. Thiele et Peter Swerling) et les algorithmes de filtrage
- MAP en décodage (Bahl, Cocke, Jelinek & Raviv BCJR 1974)

En apprentissage statistique :

- Classifieur optimal de Bayes. Le choix de l'étiquette de plus forte probabilité a posteriori (MAP) minimise la probabilité d'erreur :

$$f_\star(x) = \arg\max_{k \in \mathcal{Y}} p(Y = k | X = x)$$

  Si la loi conjointe était connue, le problème serait résolu.

- Sinon ?

Algorithmes fondés sur Bayes

Dès qu'il y a connaissance *a priori* et observation.

En dehors de l'apprentissage statistique :

- Stratonovich–Kalman-Bucy (1960 après Thorvald N. Thiele et Peter Swerling) et les algorithmes de filtrage
- MAP en décodage (Bahl, Cocke, Jelinek & Raviv BCJR 1974)

En apprentissage statistique :

- Classifieur optimal de Bayes. Le choix de l'étiquette de plus forte probabilité a posteriori (MAP) minimise la probabilité d'erreur :

$$f_\star(x) = \arg\max_{k \in \mathcal{Y}} p(Y = k | X = x)$$

  Si la loi conjointe était connue, le problème serait résolu.
- Sinon ?

Restriction sur les lois. Analyse discriminante

$$p(Y = k|X = x) = \frac{p(x|Y = k)p(Y = k)}{p(x)} = \frac{p(x|Y = k)p(Y = k)}{\sum_{j=1}^{K} p(x|Y = j)p(Y = j)}$$

$\{x \in \mathcal{X} : \delta_k(x) = \delta_l(x)\}$ frontière de décision entre les classes, avec $\delta_k(x)$ transformations monotones des $p(Y = k|X = x)$.

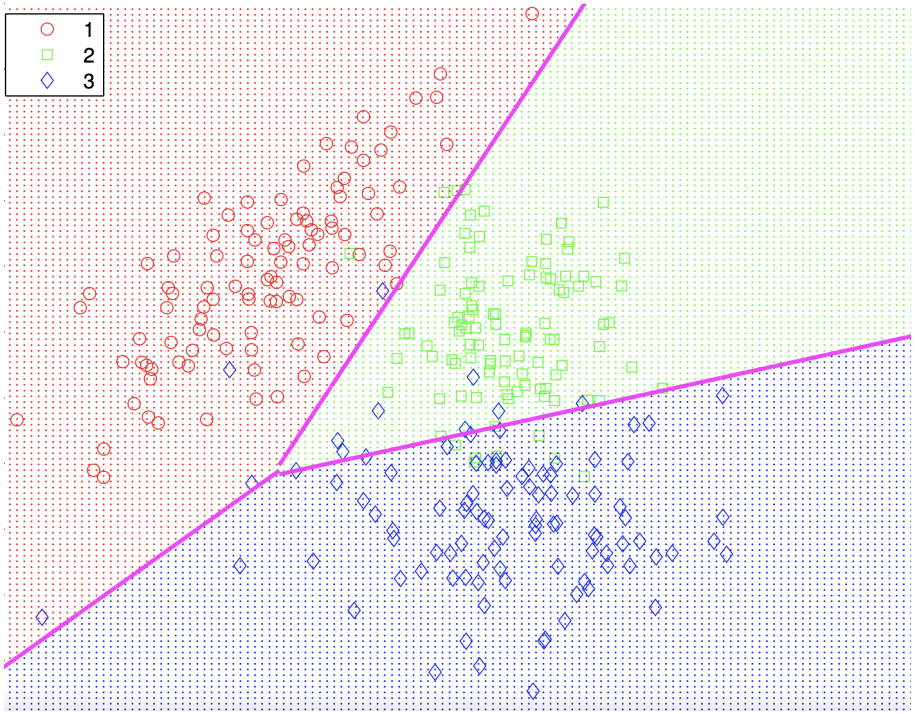
En pratique les quantités suivantes sont inconnues :

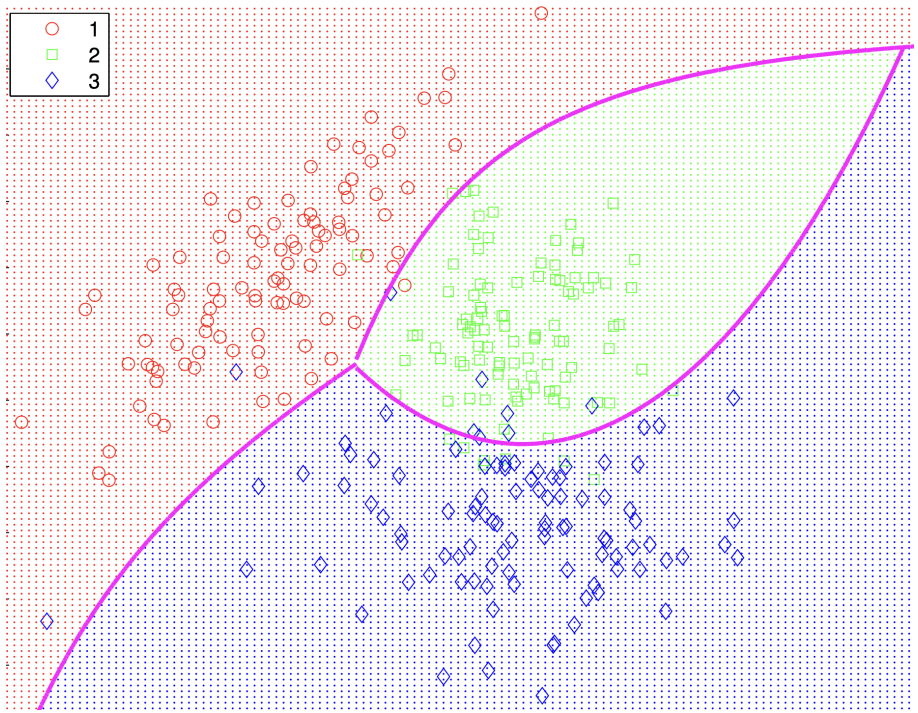
- $\pi_k = p(Y = k)$ la probabilité *a priori* de chaque classe
- $p_k(x) = p(x|Y = k)$ la vraisemblance sous chaque hypothèse

Solutions approchées en assimilant les $p_k(x)$ à des **gaussiennes** et en estimant empiriquement leurs paramètres :

- LDA. Analyse discriminante linéaire. Même covariance.
- QDA. Analyse discriminante quadratique. Covariances différentes.

Analyse discriminante linéaire ou quadratique





La restriction sur les lois impose une restriction sur la classe des frontières.

Extrait du cours de "Machine Learning" de O.J.J. MICHEL et F. CHATELAIN

Pas de restriction sur les lois

Dans l'approche de Bayes, il faut connaître les lois ou faire une hypothèse restrictive forte sur les lois pour se ramener à une estimation paramétrique.

Sans connaissance de la loi :

- L'inférence est-elle toujours possible ?
- Faut-il restreindre autre chose ?
- Que peut-on dire de
  - la quantité de données nécessaire ?
  - la précision du résultat ?
  - la probabilité d'obtenir un résultat avec cette précision ?

Risque empirique

Le risque $\mathcal{L}_{\mathcal{D}}(h)$ est théorique, c'est une espérance :

$$\mathcal{L}_{\mathcal{D}}(h) = \mathcal{D}\left(h(x) \neq y\right) = \mathbb{E}\left(1_{h(x)\neq y}\right)$$

On parle souvent de risque théorique, de risque vrai ou de risque de généralisation (l'espérance ne porte par que sur $\mathcal{S}$).

On en construit un estimateur : le risque empirique

$$L_S\left(h\right) = \frac{1}{|\mathcal{S}|} \left| \, x_i \in \mathcal{S} : h(x_i) \neq y_i \, \right|$$

estimation fluctuante et biaisée du risque.

Minimisation du risque empirique

Le risque étant non calculable, on minimise le risque empirique.

$$h_S = \arg\min_h L_{\mathcal{S}}(h)$$

Quel est le risque de ce prédicteur h_S ? C'est-à-dire, quelles sont ses performances en généralisation (sur des données hors de $\mathcal{S}$) ?
Ça ne va pas toujours bien se passer.

Exemple : $h_{\mathcal{S}}(\mathbf{x}_i) = \begin{cases} y_i \text{ pour } \mathbf{x}_i \in \mathcal{S} \\ 0 \text{ sinon} \end{cases}$

- Ce prédicteur est parfait sur $\mathcal{S}$: $L_{\mathcal{S}}(h_{\mathcal{S}}) = 0$ quel que soit $\mathcal{S}$.
- La valeur prédite est toujours égale à 0 pour de nouvelles données.

Le prédicteur colle trop aux données d'apprentissage et généralise mal : c'est le **surapprentissage**.

Minimisation du risque empirique sous contrainte

Sans contrainte sur les lois, une autre contrainte est nécessaire pour qu'une généralisation soit possible.

Conserver la méthode de minimisation du risque empirique en limitant le risque de surapprentissage.

Régularisation de la minimisation du risque empirique

Restreindre le choix des prédicteurs à $h \in \mathcal{H}$,
où $\mathcal{H}$ est un espace de fonctions à définir *a priori* (avant de voir les données) :
Le prédicteur est estimé par minimisation du risque empirique régularisé :

$$h_S = \arg\min_{h \in \mathcal{H}} L_{\mathcal{S}}(h)$$

Compromis biais variance

Le risque peut se décomposer en une somme de deux termes significatifs :

$$L_{\mathcal{D}}(h_S) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \left[L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \right]$$

Interprétation du premier terme : $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ est le risque minimum qu'il est possible d'obtenir sur la classe $\mathcal{H}$. Ce terme

- est indépendant de la taille de la base d'apprentissage,
- dépend seulement de la taille $|\mathcal{H}|$ de la classe $\mathcal{H}$.

C'est une erreur d'approximation (biais d'induction) qui dépend de la pertinence et de la taille de la famille de prédicteurs $\mathcal{H}$.

Compromis biais variance

Le risque peut se décomposer en une somme de deux termes significatifs :

$$L_{\mathcal{D}}(h_S) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \left[L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \right]$$

Interprétation du second terme : $L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ est l'écart entre le risque de la solution qui "minimise le risque empirique" par rapport à la solution optimale dans $\mathcal{H}$. C'est une erreur d'estimation : la solution obtenue par minimisation du risque empirique dépend des exemples tirés selon la loi $\mathcal{D}$, c'est une variable aléatoire dont les fluctuations dépendent

- du nombre n d'exemples utilisés pour apprendre
  (décroît lorsque n croît)
- de la taille de $\mathcal{H}$
  (croît en $\log |\mathcal{H}|$ pour $|\mathcal{H}| < +\infty$.)

Compromis biais variance

$$L_{\mathcal{D}}(h_S) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \left[L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \right]$$

Compromis biais variance entre ces deux termes contradictoires :

- $\mathcal{H}$ suffisamment *grand* pour que l'approximation soit bonne.
  Trop grand : surapprentissage.
- $\mathcal{H}$ suffisamment *petit* pour que le terme de fluctuation soit faible.
  Trop petit : sous apprentissage.

Importance fondamentale de l'a priori.

- Lorsque peu d'information *a priori* est disponible, beaucoup de données sont nécessaires à l'apprentissage.
- Au contraire lorsque l'on dispose d'énormément d'information *a priori*, peu de données sont nécessaires voire aucune : systèmes experts.

Mise en oeuvre par minimisation sous contrainte l_0,l_1 (LASSO), l_2 (Ridge)

Apprentissage Probablement Approximativement Correct

Que la classe $\mathcal{H}$ puisse être apprise au sens PAC signifie qu'il existe :

- une fonction $n_{\mathcal{H}}(\epsilon, \delta) : (0,1)^2 \to \mathbb{N}$
- un algorithme $\mathcal{A}$ tels que
  - pour tout $\epsilon, \delta > 0$,
  - pour toute distribution $\mathcal{D}$ sur $\mathcal{X}$ et pour toute fonction d'étiquetage $f : \mathcal{X} \to \mathcal{Y}$ (ou pour toute loi sur $\mathcal{X} \times \mathcal{Y}$)

la minimisation du risque empirique en utilisant un nombre d'exemples $n \geq n_{\mathcal{H}}(\epsilon, \delta)$ tirés de manière i.i.d. selon la distribution $\mathcal{D}$, donne un prédicteur h tel que

$$\mathcal{L}_{\mathcal{D}}(h) < \epsilon \text{ avec probabilité supérieure à } 1 - \delta.$$

Apprentissage Probablement Approximativement Correct

Probablement (avec probabilité supérieure à $1 - \delta$)
le prédicteur calculé est
approximativement correct (précision meilleure que ϵ).

Probablement correct et *approximativement correct sont inévitables* :

- L'échantillon $\mathcal{S}$ peut être non représentatif.
- L'échantillon $\mathcal{S}$ peut être trop petit pour capturer certains détails.

Ordre de grandeur pour le nombre d'exemples nécessaire :

$$\frac{\log(1/\delta)}{\epsilon^2}\text{(étiquetage souple)} \quad \frac{\log(1/\delta)}{\epsilon}\text{(étiquetage dur)}$$

Une classe finie peut être apprise au sens PAC

La minimisation du risque empirique sur $\mathcal{S}$ calcule un prédicteur h_S dont le risque de généralisation est faible avec une forte probabilité pour peu que la taille de la base d'apprentissage soit suffisante.

- Une différence importante avec l'approche statistique classique est que l'hypothèse h est choisie à partir d'un ensemble S de données.
- Seul l'espace $\mathcal{H}$ est fixé *a priori*.
- A contrario, en statistique classique, l'hypothèse est fixée avant de voir les données, puis cette hypothèse est testée.

Une classe infinie peut parfois être apprise au sens PAC

Pour une classe infinie, l'apprentissage n'est pas toujours possible mais il l'est parfois.

L'important n'est pas le cardinal infini de la classe $\mathcal{H}$ mais l'aptitude de ses éléments à séparerer plus ou moins de points :
la quantité importante est la dimension de Vapnik-Chervonekis : si elle est finie, l'apprentissage est possible ; sinon, il ne l'est pas.

Il n'existe pas de meilleur algorithme universel.

Pour tout classifieur binaire et pour $n < |\mathcal{X}|/2$,
il existe une distribution $\mathcal{D}$ sur $\mathcal{X} \times \{0;1\}$ telle que

1. $\exists\, f : \mathcal{X} \to \{0;1\}$ avec $L_{\mathcal{D}}(f) = 0$
2. $\Pr\left[L_{\mathcal{D}}(h_{\mathcal{S}}) \geq \frac{1}{8}\right] \geq \frac{1}{7}$

Pas d'apprenant universel : il existe pour chaque apprenant une distribution sur laquelle il échoue alors que pour la même distribution, il existe un autre apprenant qui produira une hypothèse avec un petit risque.

Aucun apprenant ne peut réussir sur toutes les tâches apprenables - chaque apprenant a des tâches sur lesquelles il échoue alors que d'autres apprenants réussissent.

Ecueils classiques

Lorsque un **algorithme** est entraîné sur des **données**, deux problèmes peuvent se poser :

1. les données sont mauvaises,
2. l'algorithme est mauvais.

Ecueils classiques

Lorsque un **algorithme** est entraîné sur des **données**,
deux problèmes peuvent se poser :

1. les données sont mauvaises,
   - Le volume de données disponibles est insuffisant.
   - Les données ne sont pas représentatives.
   - Les données sont de mauvaise qualité : données aberrantes, bruit, caractéristiques incomplètes.
   - Les caractéristiques ne sont pas pertinentes. La phase de sélection et d'extraction des caractéristiques pertinentes est essentielle.
2. l'algorithme est mauvais.

Ecueils classiques

Lorsque un **algorithme** est entraîné sur des **données**, deux problèmes peuvent se poser :

1. les données sont mauvaises,
2. l'algorithme est mauvais.

Ecueils classiques

Lorsque un **algorithme** est entraîné sur des **données**,
deux problèmes peuvent se poser :

1. les données sont mauvaises,
2. l'algorithme est mauvais.
   - Trop forte attache aux données d'apprentissage.
     C'est le surapprentissage.
   - Approximation trop simple pour représenter suffisamment finement le lien entre les caractéristiques et le résultat.
     C'est le sous apprentissage

Ecueils classiques

Lorsque un **algorithme** est entraîné sur des **données**, deux problèmes peuvent se poser :

1. les données sont mauvaises,
2. l'algorithme est mauvais.

Garbage in / Garbage out.

Et les neurones dans tout ça ?

Approche montante et descendante :

↓ Réseaux de neurones. Calcule la meilleure représentation (extractions de caractéristiques) en réduisant la dimension d'une représentation initiale grande : extraction de caractéristiques et même proposition d'un espace de dimension réduite. Recherche d'une représentation minimale dans lesquelle par approximation non linéaire on cherche une frontière .

↑ Approche classique. (SVM à noyau typiquement) Choix *a priori* des caractéristiques : augmentation de la dimension pour séparer linéairement dans un espace qui est une représentation non linéaire des données d'entrée.

Problème linéaire dans le "grand espace",
obligation de non linéaire dans le "petit espace".

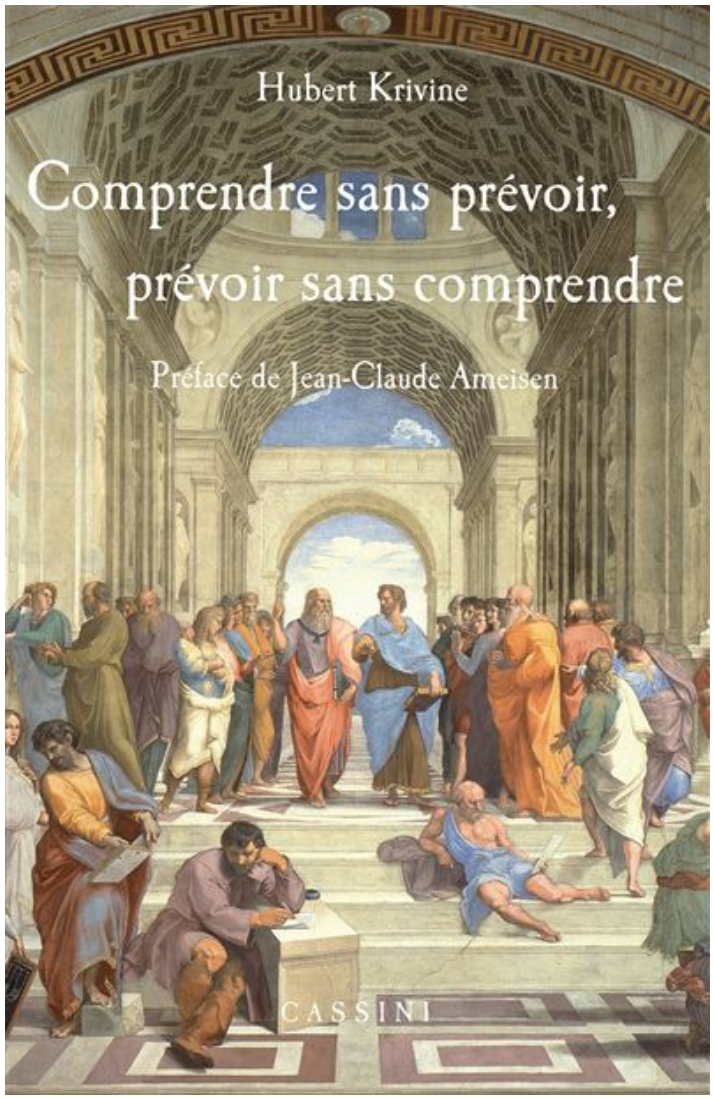
Explicabilité

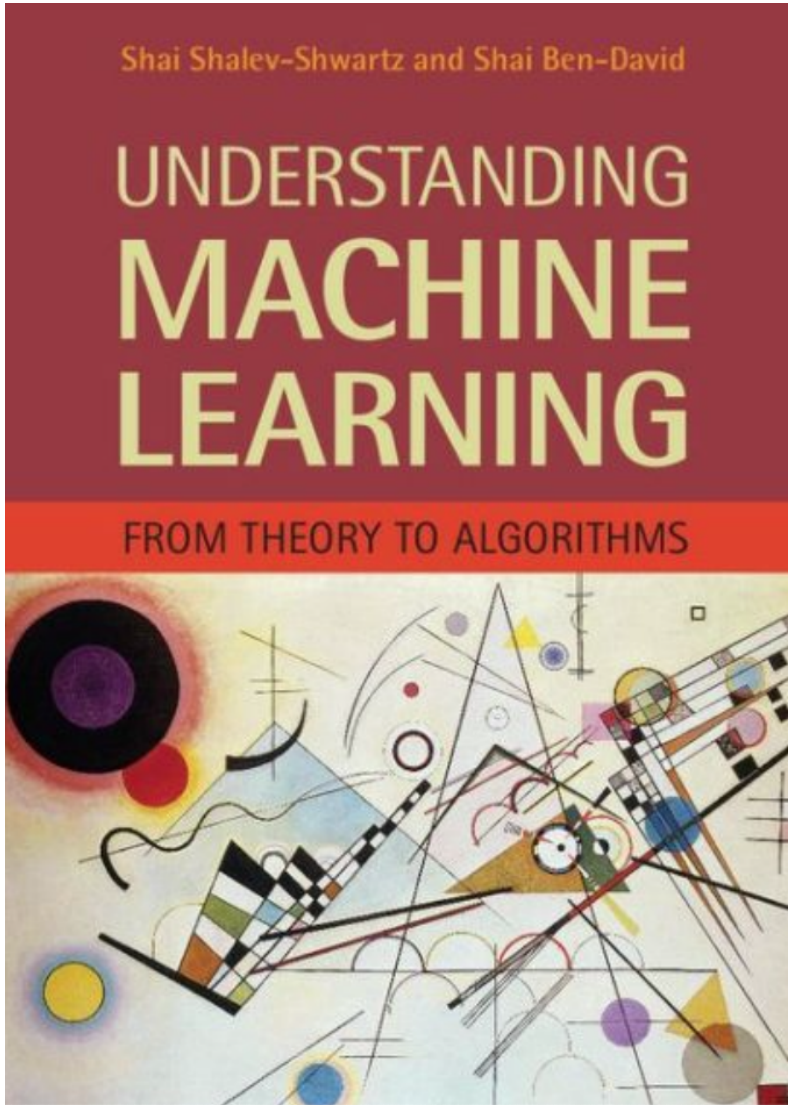
Talon d'Achille des méthodes neuronales : aucune garantie (inacceptable dans certaines applications : libération de prisonniers, défense, etc.)

Jusqu'aux années 90 : des ingénieurs, physiciens (et autres) déterminent seuls les observations pertinentes.

Mettre en place un système qui
fasse son choix de caractéristiques de manière plus automatique à partir des données brutes
et aussi,
parmi les caractéristiques choisies par des experts, soit capable d'évaluer *a posteriori* leurs poids dans la décision.

Références utilisées





et discussions avec O.J.J. Michel.