Introduction à la classification supervisée

Fabien Torre

Université de Lille

Mercredi 23 septembre 2009

Exemples de tâches

On veut répondre automatiquement à des questions comme :

- le patient aura-t-il un accident cardio-vasculaire ?
- la molécule que je désire commercialiser est-elle cancérigène ?
- qui est l'auteur de cette page HTML ?
- cette phrase est-elle grammaticalement correcte ?
- quelle sera la taille de cet enfant à l'âge adulte ?

Ne pas écrire des programmes qui répondent à ces questions... mais les *découvrir* automatiquement, par apprentissage (observation d'exemples et de contre-exemples).
En vue de prédire (classer de nouveaux exemples), on ne peut donc pas apprendre par cœur !

Historique : la motivation venue des systèmes experts

- Volonté de remplacer les experts humains ;
- mise en place de raisonnements automatiques sur des faits et des règles ;
- mais où trouver les règles ? les demander à l'expert.

- mais l'humain peut difficilement expliciter son expertise (*cf. hiver de l'Intelligence Artificielle*) ;
- on peut simplement lui demander de faire ce qu'il sait faire ;
- l'observer et apprendre.

Aujourd'hui : la motivation de la découverte

- INDANA [Colombet, 2002] :
  - prédiction du risque cardio-vasculaire après un examen minimal ;
  - des économies réalisées...
- Skicat [Fayyad, 1995] :
  - quel secteur du ciel regardé ? plusieurs téraoctets de données ;
  - 40 fois plus d'objets découverts par nuit d'observation, dépasse l'humain sur les objets faiblement lumineux ;
- molécules cancérigènes [Srinivasan et al., 1994] :
  - décider si un produit peut être diffusé, expérimentations de plusieurs années sur des animaux ;
  - bonnes performances, au croisement de plusieurs disciplines.

La motivation de la découverte (2)

Bilan

Découverte scientifique là où il n'y a pas ou peu d'expertise humaine.
Vocabulaire : règles/hypothèses et classifieurs/théorie.

À vous de jouer ?

- Loto : prédire le gain d'une grille ?
- football : prédire les résultats.
- etc.

Comment modéliser ces problèmes ?

Les points à définir

1. Ce que l'on veut prédire ;
2. modalité d'obtention des exemples ;
3. nature des exemples, nature des classifieurs ;
4. mode d'évaluation des prédictions ;
5. méthode d'apprentissage.

Étiquetage des exemples

Étiquettes et prédictions possibles : $\mathcal{Y}$

- binaires (-1 et $+1$) ou discrètes à plus de deux valeurs ;
- continues (*régression*) ;
- plus complexe (*sortie structurée*),

parfois avec une valeur de confiance.

Différents étiquetages des exemples

- tous étiquetés : *apprentissage supervisé* ;
- aucun étiqueté : *apprentissage non supervisé* ;
- un peu des deux : *semi-supervisé* (ex : les pages web) ;
- à la demande : *apprentissage actif* ;
- et aussi *apprentissage par positifs seuls* ! (ex : parents-enfants).

Arrivée des exemples

Parallèle avec les interactions enseignant-élève.

- Échantillon fixé à l'avance : cadre expérimental, A un ensemble d'exemples disponibles pour l'apprentissage, éventuellement T un ensemble de test indépendant de A.
- obtention incrémentale des exemples :
  - un par un et disponible à l'infini : *modèle de [Gold, 1967]* (influence de l'apprentissage d'une langue par le bébé humain)
  - disponible en nombre polynomial : *PAC, modèle de [Valiant, 1984]* ;
  - soumission d'un exemple et étiquetage par un oracle : *apprentissage actif* ;

Difficultés possibles dans les exemples

- Bruit dans les données (classes, description) ;
- valeurs manquantes dans les descriptions ;
- déséquilibre de classes ;
- complexité de la description même des exemples (langage $\mathcal{X}$) :
  - vecteur,
  - séquences,
  - arbres,
  - graphes.

Vecteurs en attributs-valeurs

Les exemples sont des vecteurs de valeurs discrètes ou continues :

ClumpThickness	5	5
UniformitySize	1	3
UniformityShape	1	3
MarginalAdhesion	1	3
BareNuclei	1	3
BlandChromatin	3	4
NormalNucleoli	1	4
Mitoses	1	1
Conclusion	**B**	**M**

if ClumpThickness ≤ 6.0
then class B `[weight=0.07]`

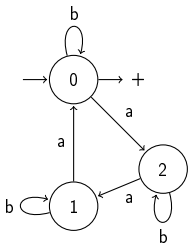
if MarginalAdhesion ≤ 8.0
then class M `[weight=0.05]`

Méthodes attributs-valeurs (*arbres de décision* vus en M1).

Séquences et IG

Les exemples sont des séquences :

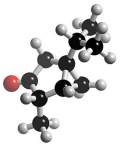
- $+$: ϵ, *aaa*, *b*, *abbaa*, *aaaaaa*
- $-$: *bba*, *aab*, *bbaababa*



Inférence Grammaticale.

Graphes et ILP

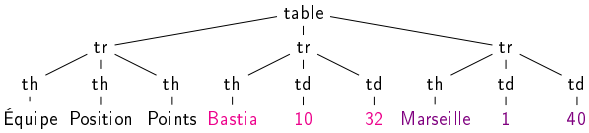
Les exemples sont des graphes :



active(M) ← atom(M,A1,carbon,22),
atom(M,A2,carbon,10),
bond(M,A1,A2,1).

Programmation Logique Inductive.

Arbres



Problématiques liées aux documents semi-structurés (XML) :

- classification d'arbres et inférence de DTD ;
- extraction n-aire et apprentissage de transformations.

Au croisement de l'IG, de l'ILP et des langages formels d'arbres.

Difficultés possibles avec les hypothèses

- garantie de l'existence du concept-cible ?
- cible évoluant dans le temps (*révision de théorie*) ;
- concept disjonctif...
- ... nous cherchons une unique hypothèse de $\mathcal{H}$, une disjonction ou un ensemble ?
- complexité de la description même des hypothèses (langage $\mathcal{H}$) :
  - règles ou arbres de décisions,
  - automates ou grammaires,
  - programmes logiques.

Évaluations possibles

- La compréhensibilité !
- évaluation théorique, l'erreur en généralisation : probabilité de se tromper en classant un nouvel exemple ;
- modèle de Gold : découverte du concept cible, erreur nulle, à la limite ;
- *probably approximately correct* (PAC) : erreur inférieure à ϵ avec une probabilité $1-\delta$ après observation d'un nombre polynomial d'exemples ;
- évaluation expérimentale, approximation de l'erreur en généralisation : erreur en apprentissage (calculée sur A), erreur de test (calculée sur T).

Évaluations expérimentales

- Matrice de confusion calculée sur un ensemble de test :

Prédiction \ Classe	-1	$+1$
-1	vn	fn
$+1$	fp	vp

- à partir de cette matrice :
  - *accuracy* : $\frac{vp+vn}{vp+vn+fp+fn}$;
  - rappel : $R = \frac{vp}{vp+fn}$ et précision $P = \frac{vp}{vp+fp}$;
  - f-mesure : $F = 2 \times \frac{P \times R}{P+R}$;
  - sensibilité $Se = R$ et spécificité $Sp = \frac{vn}{vn+fp}$;
- validations croisées et *leave-one-out* ;
- test statistique 5x2cv [Dietterich, 1998].

No free lunch theorem [Wolpert and Macready, 1995]

- Soit un problème d'apprentissage donné par (A, T) ;
- construction d'un problème dual (A, T') : T' contient les mêmes exemples que T mais les classes sont inversées ;
- on apprend dans les deux cas sur A, on apprend donc la même théorie ;
- l'erreur mesurée sur T et celle mesurée sur T' moyennent à 0.5.

En moyenne sur l'ensemble des problèmes possibles, une méthode d'apprentissage ne fait pas mieux que le hasard.

Les méthodes

- Selon $\mathcal{X}$ le langage de description des exemples ;
- selon $\mathcal{H}$ le langage de description des hypothèses ;
- paramétriques (on suppose un modèle de distribution des données et l'apprentissage consiste à chercher ses paramètres) vs non paramétriques (arbres de décision par exemple) ;
- optimisation vs symbolique : on va vers le *learning as search*.

Cadre et notations (1)

- Classification supervisée à deux classes $\mathcal{Y} = \{+1, -1\}$;
- des langages : $\mathcal{X}$ pour les exemples, $\mathcal{H}$ pour les hypothèses :

$$h \in \mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}$$

- données parfaites (descriptions complètes et sans bruit) ;
- modèle expérimental : un ensemble d'apprentissage A de n exemples étiquetés $(x_i \in \mathcal{X}, y_i \in \mathcal{Y})$, et un de test T ;
- et modèles théoriques ;

Cadre et notations (2)

- erreur en généralisation :

$$e(h) = P_{(x,y)}(h(x) \neq y)$$

- approchée par l'erreur en apprentissage :

$$e_A(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h(x_i)|$$

  et par l'erreur en test e_T ;
- souci de compréhensibilité ;
- représentations d'exemples et les méthodes associées ;
- mais que faire du *no free lunch* ? ? ?

Plan du cours

Aujourd'hui :

- Introduction à la classification supervisée ;
- comment poser un problème d'apprentissage ?
- l'apprentissage comme un problème de recherche ;
- *retour sur les arbres de décision*.

Ensuite, cours chaque mercredi à 8h45, aternance théorique/algorithmique :

- PAC, un modèle d'apprentissage particulier ;
- passage en revue des représentations :
  - attributs-valeurs ,
  - séquences et automates ;
  - *arbres* ?
  - graphes et programmes logiques ;
- méthodes d'ensemble.

Évaluation

- Un travail personnel, en dehors des cours, au choix :
  - constitution d'un jeu de données ;
  - codage en Java d'un algorithme vu en cours ;
  - expérimentations (loto, football, participation à un challenge) ;
  - résumé d'un cours ;
  - synthèse d'articles ;
  - travail théorique (écriture d'une preuve non vue en cours) ;
- (une pause d'une semaine à mi-parcours) ;
- une interrogation finale en décembre liée à un article cité en cours et fourni à l'avance.

Mais... qui êtes vous ?

- Parcours dans le master ?
- parcours avant le master ?
- *rigueur scientifique* ?
- connaissances en apprentissage ? arbres de décision ?
- IA et problèmes de recherche dans les graphes ?
- statistiques ?
- complexité et calculabilité ?
- théorie des langages ?
- langages de programmation : perl ? java ? prolog ?
- LaTeX ?

Bibliographie I

Colombet, I. (2002).
Aspects méthodologiques de la prédiction du risque cardiovasculaire : apports de l'apprentissage automatique.
PhD thesis, SPIM Inserm.

Dietterich, T. G. (1998).
Approximate statistical tests for comparing supervised classification learning algorithms.
Neural Computation, 10(7) :1895–1924.

Fayyad, U. M. (1995).
Skicat : Sky image cataloging and analysis tool.
In *IJCAI*, pages 2067–2068.

Bibliographie II

- Gold, E. M. (1967).
  Language identification in the limit.
  Information and Control, 10(5) :447–474.
- Srinivasan, A., Muggleton, S., King, R. D., and Sternberg, M. J. E. (1994).
  Mutagenesis : ILP experiments in a non-determinate biological domain.
  In Wrobel, S., editor, *Proceedings of the 4th International Workshop on Inductive Logic Programming*, volume 237 of *GMD-Studien*, pages 217–232. Gesellschaft für Mathematik und Datenverarbeitung MBH.

Bibliographie III

- Valiant, L. G. (1984).
  A theory of the learnable.
  Communications of the ACM, 27 :1134–1142.
- Wolpert, D. H. and Macready, W. G. (1995).
  No free lunch theorems for search.
  Technical Report SFI-TR-95-02-010, The Santa Fe Institute.