

En introduction à
l'Apprentissage Actif

Antoine Cornuéjols

MMIP, AgroParisTech, Paris

2008-2009

Plan

- 1 Introduction
- 2 Approches constructives
- 3 Approches sélectives

Plan

- 1 Introduction
 - Exemple
 - Les approches
- 2 Approches constructives
- 3 Approches sélectives

Apprentissage actif

Définition : L'apprentissage est dit **actif** si l'apprenant peut influencer le choix des exemples d'apprentissage

Exemples :


- Le cas du MasterMind
- Activité scientifique en général

Questions

- Quel gain potentiel en terme de **nombre d'exemples d'apprentissage** ?
- Quel gain potentiel en terme de **concept apprenable** ? (peut-on apprendre plus de concept ?)
- **Comment** ?
- **Quel coût** ?

Un exemple

Distribution uniforme sur intervalle $[0,1]$

$$h_w(x) = \begin{cases} 1 & \text{if } x \geq w \\ 0 & \text{if } x < w \end{cases}$$


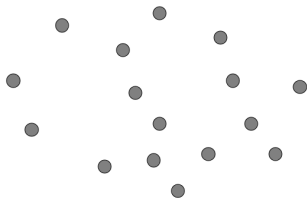
Tirage au hasard des points : $m = \mathcal{O}(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon})$

Sélection active : $m = \mathcal{O}(\log \frac{1}{\varepsilon})$

Amélioration exponentielle en terme d'échantillonnage !!

Le problème

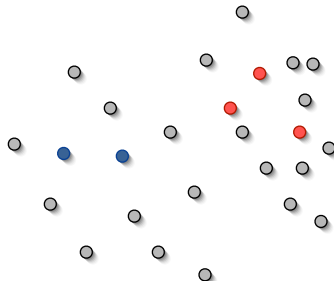
Illustration : approche sélective



Quels points examiner ?

Le problème

Illustration : approche sélective



Deux grandes approches

1 Approches **constructives**

- L'apprenant « **construit** » ses questions

2 Approches **sélectives**

- L'apprenant **sélectionne** des exemples dans un échantillon non étiqueté

Comment sélectionner les exemples ?

Scenarii

Approches constructives

- Construction d'exemple par l'apprenant
 - Nuances critiques (« *Near-Misses* »)
 - Construction d'exemples à partir de valeurs d'attributs

Approches sélectives

- ***Pool-based***
 - Sélection d'exemples parmi des exemples non supervisés
- ***Stream-based***
 - Sélection d'exemples à la volée dans un flux de données

Plan

- 1 Introduction
- 2 Approches constructives**
- 3 Approches sélectives

Comment sélectionner les exemples ?

Heuristiques

Principe général

Sélectionner les **exemples les plus informatifs**

- **Permettent de construire un modèle du monde** (apprentissage symbolique)
 - Nuances critiques (« *Near-Misses* »)
 - Apprentissage constructif (incrémental). Généralement guidé par un professeur.
- **Heuristique d'estimation d'espérance de gain d'information**
 - Réduction d'incertitude
 - Minimisation de l'espérance d'erreur
 - Réduction de l'espace des versions
- **Approche par comités** (« *Query-by-Committee* »)

Notion de Near-Miss

[Winston, 75]



(a)



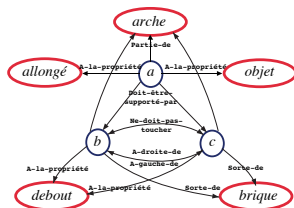
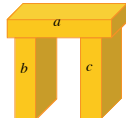
(b)



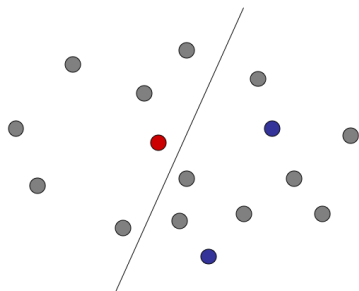
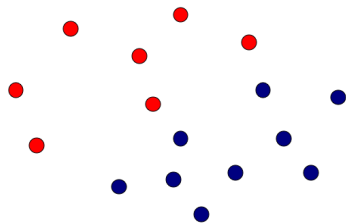
(c)



(d)



Un exemple



Identification exacte

[Angluin, 1992]

$$\mathcal{H}_0 = \mathcal{H}$$

For $t = 1, 2, \dots$

Choisir un exemple $x \in \mathcal{X}$ et demander son étiquette $f(x)$

$\mathcal{H}_t =$ toutes les hypothèses de \mathcal{H}_{t-1} cohérentes avec $(x, f(x))$

Question

Quel est le nombre minimal de question d'appartenance (*membership queries*) sont nécessaires pour réduire \mathcal{H} à f ?

Cas où $\mathcal{H} = \mathcal{F}$ (cas réalisable)

Identification exacte

Illustration

$$\mathcal{X} = \{0, 1\}^N$$

\mathcal{H} = conjonction de littéraux (e.g. $x_1 \wedge x_3 \wedge x_34$)

$S = 0$ (ensemble des indices des conjuncts)

For $i = 1, 2, \dots, N$

Demander l'étiquette de $(1, \dots, 1, 0, 1, \dots, 1)$ 0 à la position i

Si étiquette négative : $S := S \setminus \{i\}$

Total : N requêtes

Idée générale

Synthétiser des points les plus informatifs

Chaque requête coupe l'espace des versions en 2

Identification exacte

Difficulté

Nombreux résultats dans ce cadre, même pour des classes d'hypothèses complexes.

MAIS :

[Baum and Lang, 1991] tried fitting a neural net to handwritten characters.

Synthetic instances created were **incomprehensible to humans!**

[Lewis and Gale, 1992] tried training text classifiers.

“an artificial text created by a learning algorithm is **unlikely to be a legitimate natural language expression**, and probably would be uninterpretable by a human teacher.”

Donc l'oracle ne peut pas être un humain !!

[BL91]

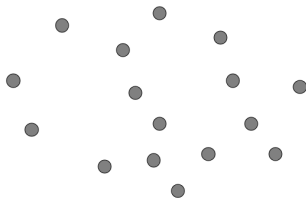
Baum, E. Lang, K. *Neural network algorithms that learn in polynomial time from examples and queries.* IEEE Trans. Neural Networks, 2.

Plan

- 1 Introduction
- 2 Approches constructives
- 3 Approches sélectives**
 - Réduction de l'Espace des Versions
 - SVM actif
 - Réduction d'incertitude
 - Echantillonnage par comité de modèles
 - Réduction d'erreur

Le problème

Illustration : approche sélective



Quels points examiner ?

Algorithme général

Algorithme 1 : Algorithme générique d'échantillonnage actif

Notations :

- h : une hypothèse prédictive munie d'un algorithme d'apprentissage
- U et L : des ensembles d'exemples non étiquetés et étiquetés
- n : le nombre d'exemples d'apprentissage souhaité
- T : échantillon d'apprentissage (avec $|T| < n$)
- Une fonction **Utile** : $\mathcal{X} \times \mathcal{H} \rightarrow \mathbb{R}$ qui estime l'utilité d'un exemple x pour l'apprentissage d'une hypothèse

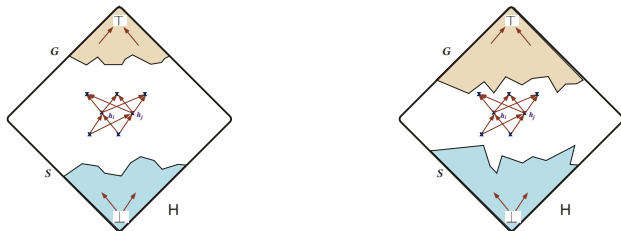
tant que $|T| < n$ faire

- (A) **Apprendre** : Explorer \mathcal{H} grâce à \mathcal{L} et T (et éventuellement U)
- (B) **Rechercher l'exemple** $q = \text{ArgMax}_{u \in U} \text{Utile}(u, \mathcal{H})$
- (C) Retirer q de U et **demander son étiquette** $f(q)$ à l'oracle
- (D) Ajouter q à L et ajouter $(q, f(q))$ à T

fin

Réduction de l'espace des versions

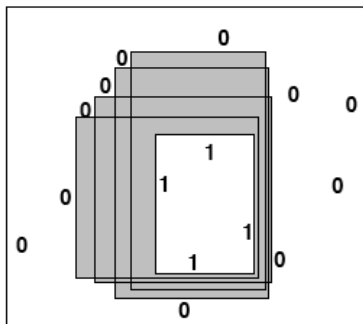
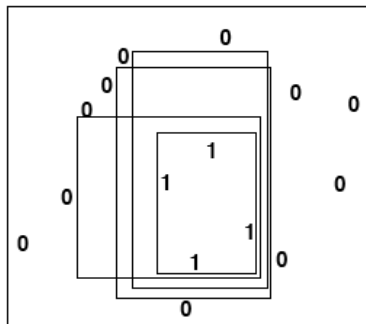
Le principe général



Comment réduire au plus vite l'EV ?

Comment réduire au maximum l'EV ?

Région d'incertitude entre le S_set et le G_set



Problème : comment déterminer la région d'incertitude pour y sélectionner les nouvelles requêtes ?

[CAL94] David Cohn, Les Atlas & Richard Ladner. *Improving generalization with active learning*. Machine Learning, 15: 201-221, 1994

Région d'incertitude

- 1 La région d'incertitude est recalculée après chaque exemple (ou après chaque petit ensemble d'exemples)
- 2 et les nouveaux exemples sont tirés dans cette région.

Région d'incertitude

Détermination avec SG-net

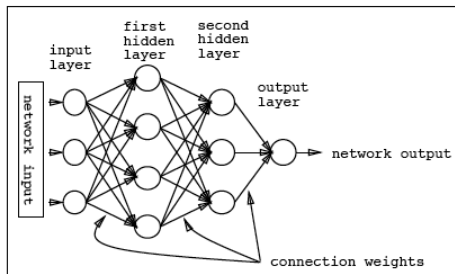


Figure 4: A simple feedforward neural network. Each node computes the weighted sum of its inputs, passes that sum through a sigmoidal “squashing” function, and passes the result on as its output.

[CAL94] David Cohn, Les Atlas & Richard Ladner. *Improving generalization with active learning*. Machine Learning, 15: 201-221, 1994

Région d'incertitude

Détermination avec SG-net

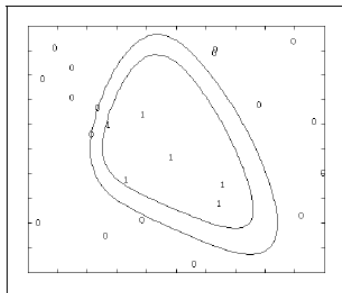


Figure 6: A naïve approach to representing the region of uncertainty: we can use the network's transition area between 0 and 1 to represent the part of the domain where the network is "uncertain."

[CAL94] [David Cohn, Les Atlas & Richard Ladner. *Improving generalization with active learning*. Machine Learning, 15: 201-221, 1994](#)

Région d'incertitude

Détermination avec SG-net

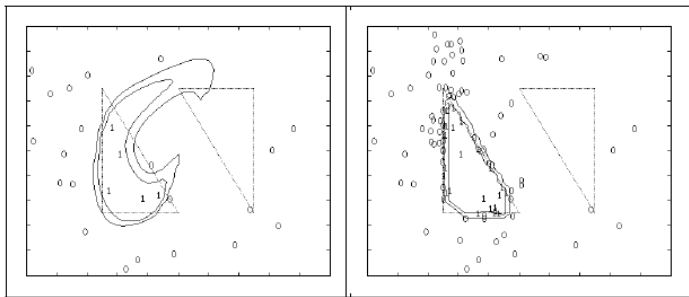


Figure 7: A pathological example of naïve network querying. In (a) on left, an initial random sample has failed to detect the second, disjoint region of the target concept. In (b) on right, after 10 successive iterations then, the naïve querying algorithm has ignored that region and concentrated on the region where is *has* seen examples. The dotted line denotes the true boundary of the unknown target concept.

[CAL94] David Cohn, Les Atlas & Richard Ladner. *Improving generalization with active learning*. Machine Learning, 15: 201-221, 1994

Région d'incertitude

Détermination avec SG-net

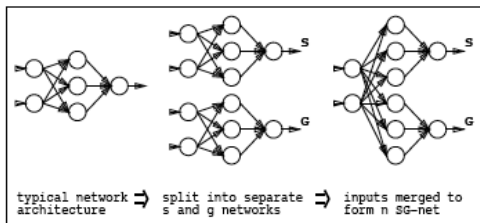


Figure 9: Construction of an SG-network equivalent to the original.

- Construire un S-net maximalement spécifique (en cherchant à classer comme négatifs les exemples non étiquetés)
- Construire un G-net maximalement général (en cherchant à classer comme positifs les exemples non étiquetés)

Région d'incertitude

Détermination avec SG-net

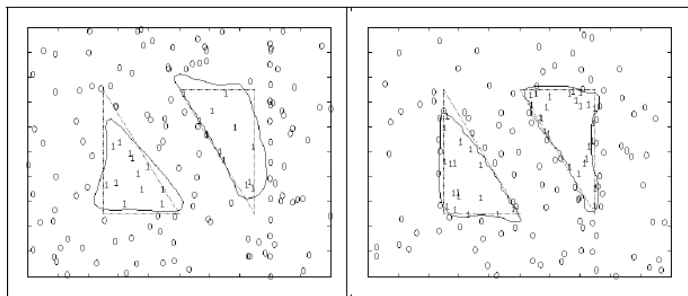


Figure 10: The triangle learner problem. When learned by 150 random examples in (a) on left, and when learned by 150 examples drawn in 15 passes of selective sampling in (b) on right. The dotted line denotes the true boundary of the unknown target concept.

[CAL94]

David Cohn, Les Atlas & Richard Ladner. *Improving generalization with active learning*. Machine Learning, 15: 201-221, 1994

Active Learning with SVM

SIMPLE MARGIN

SIMPLE MARGIN

- 1 Sélectionner l'exemple le plus proche de la séparatrice : $|\mathbf{w} \cdot \Phi(\mathbf{x})|$ minimal.

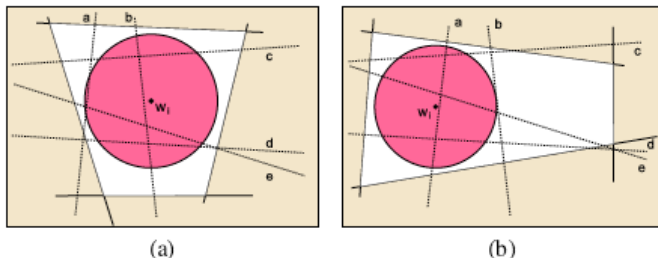


Figure 3.4: (a) Simple Margin will query b. (b) Simple Margin will query a.

Suppose que l'EV est symétrique et que w_i est placé au centre.

Active Learning with SVM

MAXMIN MARGIN

MAXMIN MARGIN

- 1 Pour chaque exemple candidat : calculer la marge m^+ si il était étiqueté $+$ et sa marge m^- si il était étiqueté $-$
- 2 Sélectionner l'exemple pour lequel m^+ et m^- sont les plus proches

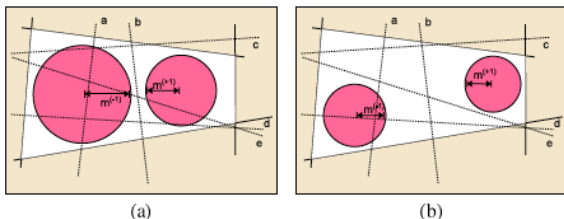


Figure 3.5: (a) MaxMin Margin will query b. The two SVMs with margins m^- and m^+ for b are shown. (b) MaxRatio Margin will query e. The two SVMs with margins m^- and m^+ for e are shown.

Suppose que l'EV est symétrique.

Active Learning with SVM

RATIO MARGIN

RATIO MARGIN

- 1 Pour chaque exemple candidat : calculer la marge m^+ si il était étiqueté + et sa marge m^- si il était étiqueté -
- 2 Sélectionner l'exemple pour lequel $\min\left(\frac{m^-}{m^+}, \frac{m^+}{m^-}\right)$ est maximal

Réduction d'incertitude

Uncertainty sampling

Cadre probabiliste dans lequel les hypothèses produisent des **prédictions** accompagnées d'un **degré de probabilité** ou de **confiance**

- 1 Les exemples étiquetés sont utilisés pour apprendre une première hypothèse
- 2 Les exemples non étiquetés pour lesquels la prédiction est accompagnée d'un degré ~ 0.5 (les plus incertains) sont candidats pour une requête

Idéalement, chaque exemple testé élimine presque la moitié de l'E.V.

[LG94]

D. Lewis and W. Gale. *A sequential algorithm for training text classifiers*. ACM-SIGIR-94, pp.3-12.

Réduction d'incertitude

Uncertainty sampling

Mesure d'incertitude fondée sur ...

- 1 ... la **probabilité de la classe prédite**

$$Incertain(\mathbf{x}) = \frac{1}{\text{ArgMax}_{y \in \mathcal{Y}} \hat{p}(y|\mathbf{x})}$$

- 2 ... la **proximité à la frontière de décision**

Active Learning with statistical models

- 1 À chaque étape, l'espérance de la variance est calculée en ajoutant un exemple candidat à l'ensemble d'apprentissage
- 2 Les exemples pour lesquels la variance est la plus forte sont candidats pour une requête

-
- [CGJ96] [David Cohn, Zoubin Ghahramani and Michael Jordan](#) *Active learning with statistical models* JAIR, 4 (1996), 129-145.
- [STP01] [Maytal Saar-Tsechansky and Foster Provost](#) *Active learning for class probability estimation and ranking*. Proc. of 17th Intl. Joint Conf. on Artificial Intelligence (IJCAI-2001) (pp. 911–920).

Réduction d'incertitude

Analyse critique

Points positifs

- Intuitif
- Facile à mettre en œuvre
 - L'incertitude de prédiction peut être calculée sur de nombreux systèmes d'apprentissage
- Peu coûteux (calcul de $|U|$ prédictions)

Points négatifs

- Problème quand données non séparables
 - Données bruitées
 - Fonction cible trop complexe
- Tend à explorer les zones de « mélange », et à ignorer le reste de \mathcal{X}

Réduction d'incertitude

Uncertainty sampling

Utilisée pour :

- Apprentissage de régression logistique [LG94, LC94]
- HMM à états partiellement cachés [SW01]
- SVM [SC00, CCS00]
- Programmation Logique Inductive [TCM99]

-
- [LG94] D. Lewis and W. Gale. *A sequential algorithm for training text classifiers*. ACM-SIGIR-94, pp.3-12.
- [LC94] D. Lewis and J. Catlett. *Heterogeneous uncertainty sampling for supervised learning*. Proc. of ICML-94, pp.148-156.
- [SW01] T. Scheffer and S. Wrobel. *Active learning of partially hidden Markov models*. Proc. of ECML/PKDD-2001, Workshop on « Active Learning Database Sampling, Experimental Design: Views on Instance Selection ».
- [SC00] G. Schohn and D. Cohn. *Less is more: Active Learning with Support Vector Machines*. Proc. of ICML-00, pp.839-846.
- [CCS00] C. Campbell, N. Cristianini, A. Smola. *Query learning with large margin classifiers*. Proc. of ICML-00, pp.111-118.
- [TCM99] C. Thompson, E. Califf, R. Mooney. *Active learning for natural language parsing and information extraction*. Proc. of ICML-99, pp.406-414.

Échantillonnage par comité de modèles

Query by committee

Vise à réduire l'espace des versions

Plusieurs hypothèses sont apprises en parallèle sur les mêmes données.

- 1 On suppose que les hypothèses apprises sont représentatives de l'espace des versions
- 2 Le désaccord au sein du comité lors de la prédiction de l'étiquette de points $x \in U$ permet d'estimer la capacité des exemples à réduire l'E.V.

Idéalement, chaque exemple testé élimine presque la moitié de l'E.V.

Rq.: Le *query by bagging* est également possible (apprentissage des hypothèses sur des sous-échantillons différents) [NH98]

[SOS92] H. Seung & M. Opper and H. Sompolinsky. *Query by committee*. COLT'92, pp.287-294, 1992.

[NH98] A. Naoki and M. Hiroshi. *Query learning strategies using boosting and bagging*. ICML'98, pp.1-9, 1998

Échantillonnage par comité de modèles

Mesures de désaccord

Mesures de désaccord ...

- 1 ... basée sur l'entropie
- 2 ... par comptage des mauvaises prédictions
- 3 ... basée sur la divergence de Kullback-Leibler

Échantillonnage par comité de modèles

Applications

- Perceptrons [FSST97]
- Classifieurs naïf de Bayes [McCN98]
- WINNOW [LT97]
- Extension à l'apprentissage bayésien [Mit97]

-
- [FSST97] Y. Freund, S. Seung, E. Shamir and N. Tishby. *Selective sampling using the query by committee algorithm*. Machine Learning journal, 28, pp.133-168.
- [McCN98] A. McCallum and K. Nigam. *Employing EM in pool-based active learning for text classification*. Proc. of ICML-98, Workshop on "Learning for text categorization".
- [LT97] R. Liere and P. Tadepalli. *Active learning with committees for text categorization*. Proc. of AAAI-97, pp.591-596.
- [Mit97] T. Mitchell. *Machine Learning* McGraw-Hill, 1997.

Méthodes par comité

Analyse critique

Points positifs

- Intuitif
- Facile à mettre en œuvre
- Coût raisonnable (calcul de $k \cdot |U|$ prédictions si k hypothèses dans le comité)

Points négatifs

- Il faut des hypothèses cohérentes (E.V. non vide)
- Il faut constituer un comité assez varié et représentatif de l'E.V.
- Il faut choisir une mesure de désaccord
- Approche heuristique sans garantie

Minimisation de l'espérance d'erreur

Expected-error minimization

Sélectionner l'exemple non étiqueté minimisant l'espérance d'erreur de l'hypothèse sur l'ensemble de test.

MAIS possible seulement pour des classes d'hypothèses extrêmement simples
[CGJ96]

Méthodes heuristiques pour estimer l'erreur

- 1 Choisir une fonction de perte utilisée pour estimer le futur taux d'erreur (e.g. [RMcM01])
- 2 Chaque exemple x non étiqueté est considéré
- 3 L'apprenant estime la réduction du taux d'erreur pour chaque étiquette possible de x
- 4 L'exemple conduisant à la plus grande réduction du taux d'erreur est sélectionné.

[RMcM01] N. Roy and A. McCallum. *Toward optimal active learning through sampling estimation of error reduction*. ICML-01, pp.441-448.

Minimisation de l'espérance d'erreur

Expected-error minimization

- Estimation empirique pour le classifieur naïf de Bayes [RMcC01]
- Estimation empirique pour l'apprentissage de paramètres dans les réseaux bayésiens [TK00]
- Estimation empirique pour l'apprentissage par plus proches voisins [LMR99]

-
- [RMcM01] **N. Roy and A. McCallum.** *Toward optimal active learning through sampling estimation of error reduction.* ICML-01, pp.441-448.
- [TK00] **S. Tong and D. Koller.** *Active learning for parameter estimation in Bayesian networks.* Proc. of NIPS-00, pp.647-653.
- [LMR99] **M. Lindenbaum, S. Markovitch and D. Rusakov.** *Selective sampling for nearest neighbor classifiers.* Proc. of AAAI-99, pp.366-371.