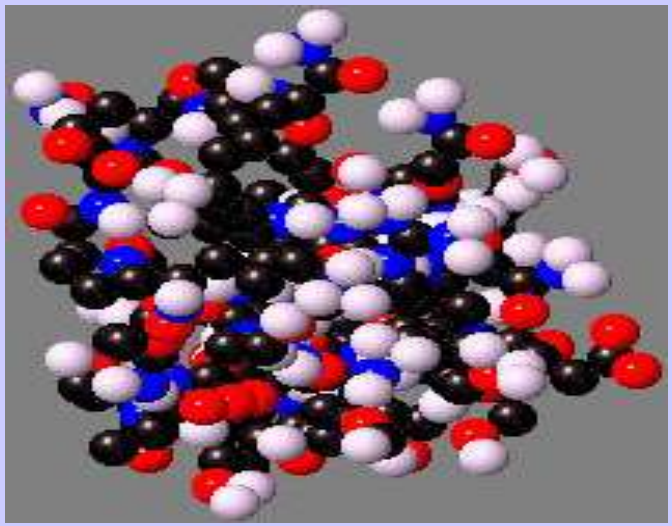
Apprentissage semi-supervisé asymétrique et estimations d'affinités locales dans les protéines



Christophe N. Magnan (Doctorant, LIF, Marseille)

CAP 2005, 03/06/2005

Plan

- Apprentissage semi-supervisé asymétrique
- Identifiabilité
- Expériences sur données artificielles
- Prédiction des ponts disulfures
- Conclusions et perspectives

Apprentissage semi-supervisé asymétrique

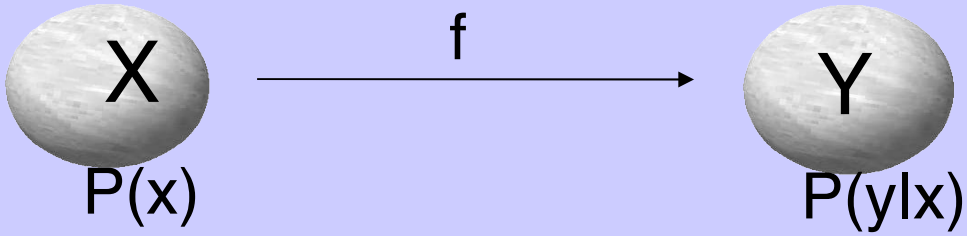
- variante de l'apprentissage semi-supervisé
  $S_{lab}=\{ (x_i,y_i) \}$, $S_{unl}=\{ x_i \}$, $x_i \in X$, $y_i \in Y$
- classification binaire: $Y=\{ 0 , 1 \}$
- données: $S_{pos}=\{(x_i,1)\}$, $S_{unl}=\{ x_i \}$
- S_{pos} distribué selon $P(x \mid y=1)$

Apprentissage semi-supervisé asymétrique

- F. Denis: « PAC learning from positive statistical queries » (ALT 1998)
- F. Denis, F. De Comité, R. Gilleron, F. Letouzey: « Positive and unlabeled examples help learning » (ALT 1999)
- F. Denis, R. Gilleron, A. Laurent, M.Tommasi: « Text classification and cotraining from positive and unlabeled examples » (ICML 2003)
- B. Liu, X. Li: « Learning to classify text using positive and unlabeled data » (IJCAI 2003)

- Problème: identifiabilité?

Identifiabilité



- Données: P(x) et P(x | y=1)
- Définition: une distribution $\mathcal{D}$ est identifiable ssi les données permettent de déterminer cette distribution.
- Objectif: déterminer P(y=1 | x) ∀x
- Mais P(y=1 | x) est-elle déterminée par les données?

Identifiabilité (suite)

$$P(y=1 \mid x) = \frac{P(x \mid y=1).P(y=1)}{P(x)}$$

- Propriété: P(y=1|x) déterminée ssi P(y=1) est déterminé
- Propriété: $\mathcal{D}$ identifiable ssi P(y=1) identifiable
- Les données positives et non étiquetées permettent-elles d'identifier P(y=1)?

Identifiabilité (suite)

- NON sans hypothèse supplémentaire

> => $\forall \lambda \leq \inf \{ P(x)/P(x \mid y=1) \}$, on peut construire une distribution P où $\lambda = P(y=1)$

- Mais pour certaines classes de distributions, P(y=1) identifiable à partir de P(x) et P(x | y=1).

Exemple trivial de distribution identifiable

- Modèles déterministes:

$\forall x \in X, P(y=1|x)=1$ ou $P(y=1|x)=0$

$$P(y=1)= \Sigma_x P(x)P(y=1|x)$$

Distributions du classifieur naïf de Bayes

- $C_{Bayes}(x) = \text{argmax}_y P(y|x) = \text{argmax}_y P(x|y).P(y)$

  => optimal mais incalculable dans la plupart des problèmes réels

- Hypothèse naïve de Bayes : les attributs sont indépendants conditionnellement à chaque classe

- $C_{NaiveBayes}(x) = \text{argmax}_y P(y) \prod P(x^i|y)$

  => nombre de paramètres raisonnable, hypothèse naïve.

Identifiabilité des modèles naïfs de Bayes

- ***Théorème:*** pour les distributions qui respectent l'hypothèse naïve de Bayes, P(y=1) est déterminé par P(x) et P(x | y=1) dès 2 attributs *.

- * ***Condition:*** il existe x^i et x^j pour lesquels $P(x^i|y=1) \neq P(x^i)$ et $P(x^j|y=1) \neq P(x^j)$.

- Soient x^i et x^j tels que $P(x^i|y=1) \neq P(x^i)$ et $P(x^j|y=1) \neq P(x^j)$, alors:

$$P(y=1) = \frac{P(x^i \cap x^j) - P(x^i)P(x^j)}{P(x^i|y=1)P(x^j|y=1) - P(x^i|y=1)P(x^j) - P(x^j|y=1)P(x^i) + P(x^i \cap x^j)}$$

Estimateur consistant

- Estimateur consistant de P(y=1):

$$(*)\ P(y=1)=\frac{\sum_{i,j} \left| P(x^i \cap x^j)-P(x^i)P(x^j) \right|}{\sum_{i,j} \left| P(x^i|y=1)P(x^j|y=1)-P(x^i|y=1)P(x^j)- P(x^j|y=1)P(x^i)+P(x^i \cap x^j) \right|}$$

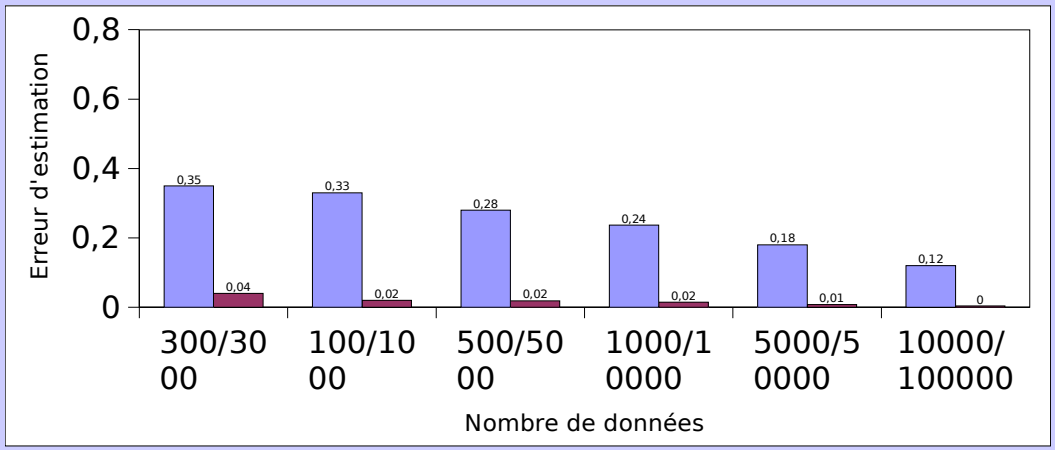
- Algorithme sous-jacent:
  - ***Entrée***: S_{pos}, S_{unl}
    - *1) Estimer P(x) et P(x|y=1)*
    - *2) Calculer P(y=1) avec (*)*
    - *3) Calculer P(y=1|x) ∀x*
  - ***Sortie***: *un modèle* θ

Identifiabilité des modèles naïfs de Bayes à partir de P(.)

- 2001:D. Geiger, D. Heckerman, H. King, C. Meek: « Stratified exponential families: Graphical models and model selection » (The Annals of statistics, 29(2), 505-529)
- Identifiabilité des modèles naïfs de Bayes dès 3 attributs et à permutation de classes près à partir de P(.)
- Convergence très lente, inutilisable en pratique

Erreur d'estimation de P(y=1) en fonction du nombre de données:

$$\left(E\left(\hat{P}(y=1)-P(y=1) \right)^2 \right)^{1/2}$$



Estimations analytiques vs Principe du maximum de vraisemblance

- Vraisemblance:
  - $L(\theta,S_{lab})= \prod P((x_i,y_i) \mid \theta)$, $i \in \{1,...,|S|\}$
  - $L(\theta,S_{unl})= \prod [\ P(y=1)P(x_i \mid y=1,\theta)+(1-P(y=1))P(x_i \mid y=0,\theta)\ ]$
  - $L(\theta,S_{lab},S_{unl})= L(\theta,S_{lab})\ .\ L(\theta,S_{unl})$
- Principe : maximiser L

Méthode E.M. et classifieur naïf de Bayes

- 1999, A. Mc Callum, S. Thrun, T. Mitchell: « Text classification from labeled and unlabeled documents using e.m. »

- Application de E.M. à l'inférence des paramètres des modèles naïfs de Bayes en contexte semi-supervisé

- Classification de textes: gain significatif de performances grâce aux données de S_{unl}

Adaptation au cas asymétrique

- ***Entrée:*** S_{pos}, S_{unl}
  - 1) Estimer P(.) et P(.|y=1)
  - 2) Estimer P(y=1) avec (*)
  - 3) Model initial θ^0
  - 4) $\forall x \in S_{unl}$, calculer $P(y=1|x,\theta^k)$ et $P(y=0|x,\theta^k)$
  - 5) Calculer les paramètres du modèle θ^{k+1}
  - 6) Itérer à l'étape 4) jusqu'à convergence
- ***Sortie:*** un modèle θ
- Pas de résultats théoriques sur la vitesse de convergence

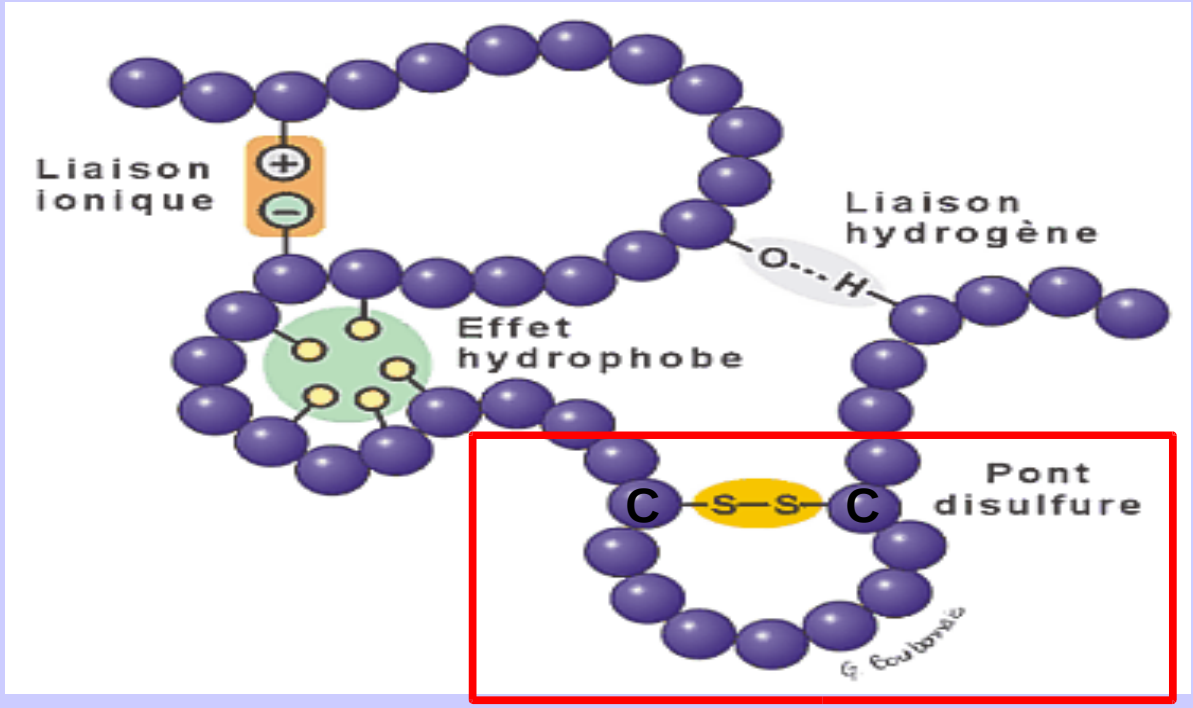
Résultats expérimentaux sur données artificielles

Nombre d'attributs binaires			20
Performances modèle cible			0.035 (0.025)
$\|S\|$	Algorithme	Ens. apprentissage	Performances
$\|S_{lab}\|$=100 $\|S_{unl}\|$=1000	NB semi-sup	$S_{lab} + S_{unl}$	0.048 (0.097) 24,1
	NB asym analytique	$S_{pos} + S_{unl}$	0.082 (0.051)
	NB asym + EM	$S_{pos} + S_{unl}$	0.042 (0.031) 24
	NB unl	S_{unl}	0.204 (0.167)
$\|S_{lab}\|$=500 $\|S_{unl}\|$=5000	NB semi-sup	$S_{lab} + S_{unl}$	0.037 (0.027) 21,6
	NB asym analytique	$S_{pos} + S_{unl}$	0.044 (0.027)
	NB asym + EM	$S_{pos} + S_{unl}$	0.037 (0.027) 23,6
	NB unl	S_{unl}	0.126 (0.124)

Résultats expérimentaux sur données artificielles

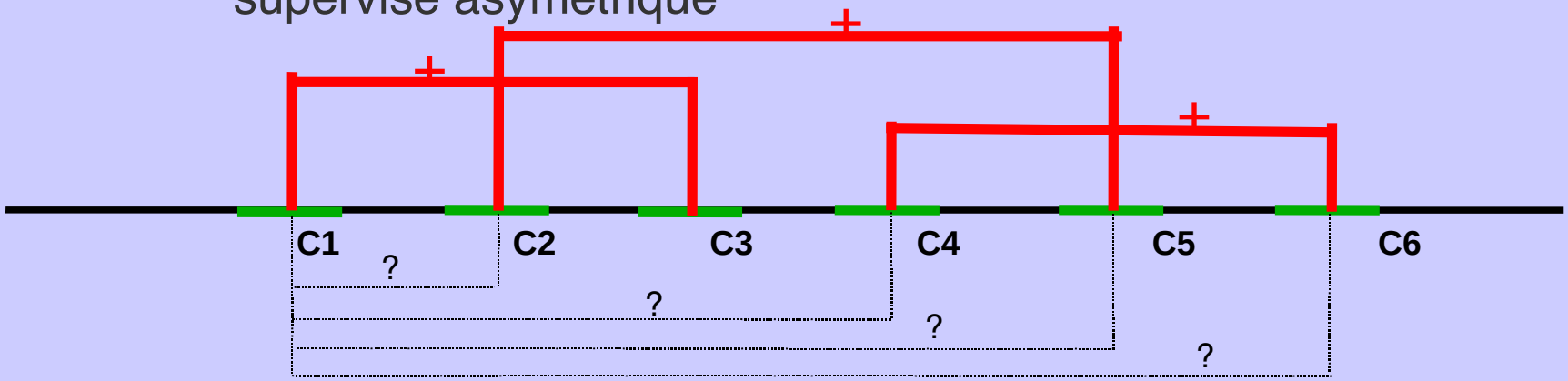
Nombre d'attributs binaires			20
Performances modèle cible			0.035 (0.025)
|S|	Algorithme	Ens. apprentissage	Performances
$\|S_{lab}\|$=100 $\|S_{unl}\|$=1000	NB semi-sup	$S_{lab} + S_{unl}$	0.048 (0.097) 24,1
	NB asym analytique	$S_{pos} + S_{unl}$	0.082 (0.051)
	NB asym + EM	$S_{pos} + S_{unl}$	0.042 (0.031) 24
	NB unl	S_{unl}	0.204 (0.167)
$\|S_{lab}\|$=500 $\|S_{unl}\|$=5000	NB semi-sup	$S_{lab} + S_{unl}$	0.037 (0.027) 21,6
	NB asym analytique	$S_{pos} + S_{unl}$	0.044 (0.027)
	NB asym + EM	$S_{pos} + S_{unl}$	0.037 (0.027) 23,6
	NB unl	S_{unl}	0.126 (0.124)

Application aux ponts disulfures



Motivations

- Etudier la contribution de l'environnement local aux cystéines à l'appariement de celles-ci : notion d'affinité locale
- Comment extraire au mieux cette information
- Etudier le statut des paires de cystéines non appariées:
  - exemples négatifs : apprentissage supervisé
  - exemples de classe non déterminée : apprentissage semi-supervisé asymétrique



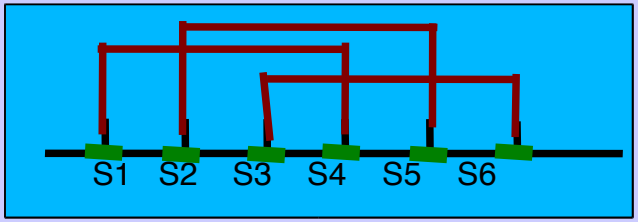
Méthodes utilisées

- Apprentissage supervisé:
  - algorithme naïf de Bayes
- Apprentissage semi-supervisé asymétrique:
  - algorithme naïf de Bayes avec E.M. et estimateur de P(y=1)

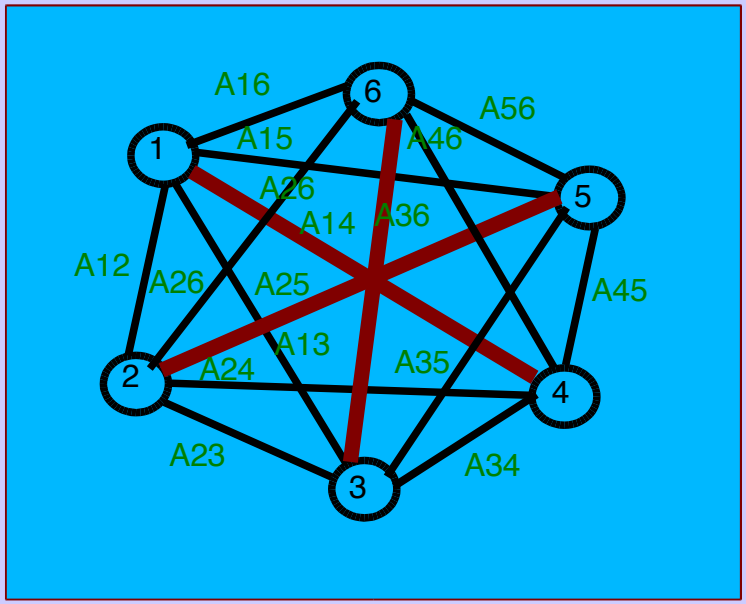
Représentation des données

- Soient f et f' deux segments, centrés sur des cystéines, de rayon n:
  - $f=x_{-n},\ldots,x_{-1},C,x_{1},\ldots,x_{n}$
  - $f'=x'_{-n},\ldots,x'_{-1},C,x'_{1},\ldots,x'_{n}$
- $(f,f') = \{ x_i x'_j \}$, $i,j \in \{ -n,\ldots,n \}$
- Alphabet de taille 231

Protocole de test



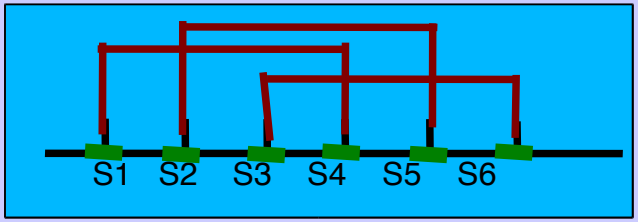
<=>



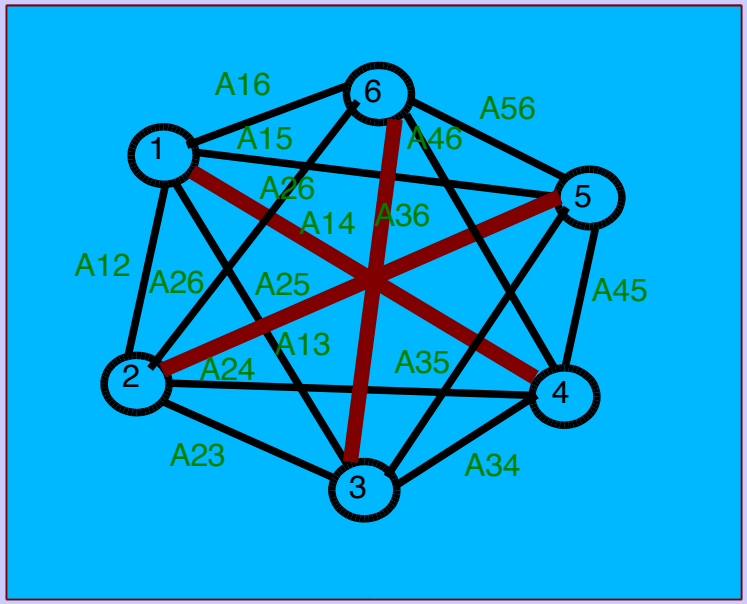
- = portion de 15 résidus (xk) centrée sur une cystéine

$A_{ij} = P(y=1 \mid f_i, f_j)$

Protocole de test



<=>



- = portion de 15 résidus (xk) centrée sur une cystéine

Aij = P(y=1 I fi,fj)

- => couplage parfait de poids maximal

Christophe N. Magnan (Doctorant, LIF, Marseille)
CAP 2005, 03/06/2005

Résultats

- Données issues de la Protein Data Bank (PDB)
  C. Geourjon (IBCP, Lyon), ACI GENOTO3D
- Comparaison des résultats sur le nombre de ponts correctement prédits

Nombre de ponts	2	3
Nombre de protéines	51	50
Nombre de ponts total	102	150
Aléa	34 (33,33%)	30 (20%)
Naïve Bayes supervisé	41 (40,2%)	26,25 (17,5%)
Naïve Bayes asym + EM	**60 (58,8%)**	**50,1 (33,4%)**

Conclusions et perspectives
Apprentissage asymétrique

- Identifiabilité des modèles naïfs de Bayes
- Algorithmes
- Résultats expérimentaux très satisfaisants
- Perspectives:
  - Autres classes de modèles identifiables
  - Méthodes pour identifier ces modèles

Conclusions et perspectives
Ponts disulfures

- Présence probable d'une information locale autour des cystéines
- L'hypothèse de représentation des données semble valide
- Perspectives:
  - Utiliser plus d'informations
  - Intégrer ces résultats dans des méthodes plus sophistiquées
  - Stratégies locales/globales

Références apprentissage semi-supervisé asymétrique

- F. Denis: « PAC learning from positive statistical queries » (ALT 1998)
- F. Denis, F. De Comité, R. Gilleron, F. Letouzey: « Positive and unlabeled examples help learning » (ALT 1999)
- F. Denis, R. Gilleron, A. Laurent, M.Tommasi: « Text classification and cotraining from positive and unlabeled examples » (ICML 2003)
- B. Liu, X. Li: « Learning to classify text using positive and unlabeled data » (IJCAI 2003)

Références sur la prédiction des ponts disulfures

- P. Fariselli, R. Casadio: « Prediction of disulfide connectivity in proteins » (2001)
- P. Fariselli, P. Martelli, R. Casadio: « A neural network-based method for predicting the disulfide connectivity in proteins » (2002)
- A. Vullo, P. Frasconi: « A recursive connectionist approach for predicting disulfide connectivity in proteins » (2003)
- A. Vullo, P. Frasconi: « Disulfide connectivity prediction using recursive neural networks and evolutionary information » (2004)