Apprentissage à partir de données diversement étiquetées pour l'étude du rôle de l'environnement local dans les interactions entre acides aminés

Christophe N. Magnan

Université de Provence - Aix-Marseille I
Laboratoire d'Informatique Fondamentale, UMR CNRS 6166
Ecole Doctorale Mathématiques et Informatique de Marseille (ED 184)

Sous la direction de François Denis (PR, LIF) et Cécile Capponi (MCF, LIF)

12 Décembre 2007

Contexte

Cadre de cette étude

- Modélisation *de novo* de la structure 3D des protéines
- Apprentissage automatique

Contexte

Cadre de cette étude

- Modélisation *de novo* de la structure 3D des protéines
- Apprentissage automatique

A.C.I. Masses de données **Genoto3D**

- **Responsable :** Yann Guermeur (LORIA)
- **Laboratoires :** LORIA, IBCP, IRISA, LIRMM, INRA, LIF
- **Dates :** Avril 2003 - Novembre 2006
- **Cadre :** prédiction de la structure 3D des protéines

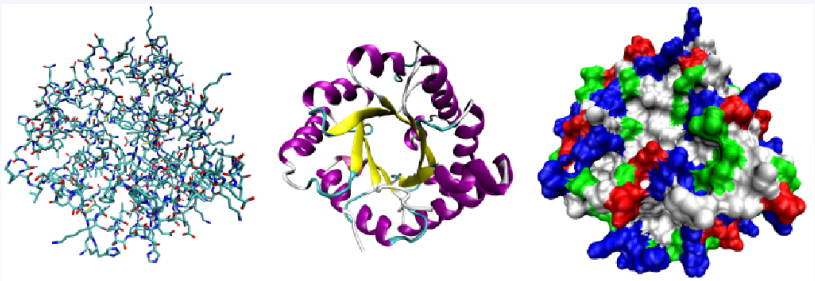
Protéines, fonctions, et conformation spatiale

Les protéines

- Chaînes d'acides aminés (20 acides aminés)
- Responsables de la totalité des activités cellulaires
- Fonctions physiologiques en grande partie conférées par la structure 3D $\Rightarrow$ nécessité de connaître la conformation spatiale

Exemple de séquence primaire (Id. PDB 1TIM)

APRKFFVGGNWKMNGKRKSLGELIHTLDGAKLSADTEVVCGAPSIYLDFARQKLDAKIGVAAQNCYKVPKGAFTGEISPAMI
KDIGAAWVILGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFQETKAIADNVKDWSKVVLAYEPV
WAIGTGKTATPQQAQEVHEKLRGWLKTHVSDAVAVQSRIIYGGSVTGGNCKELASQHDVDGFLVGGASLKPEFVDIINAKH



Quelques chiffres

Séquences protéiques (Swiss-Prot, TrEMBL)

- Extraites des séquençages automatiques de génomes
- $\simeq$ 5 Millions de séquences recensées

Structures 3D (Protein Data Bank)

- Déterminées par R.M.N., Radiocristallographie, ...
- $\simeq$ 47000 structures déterminées
  $\Rightarrow$ nécessité de proposer d'autres techniques de modélisation

Méthodes de novo

Principe

- Prédiction de la structure à partir de la séquence
- Méthodes issues de l'apprentissage automatique
- Appuyées par les travaux d'Anfinsen (Prix Nobel, 1972)

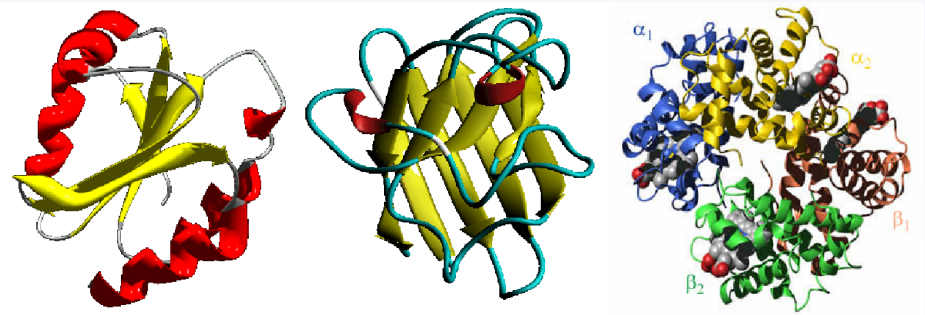
Méthodes de novo

Principe

- Prédiction de la structure à partir de la séquence
- Méthodes issues de l'apprentissage automatique
- Appuyées par les travaux d'Anfinsen (Prix Nobel, 1972)

Prédiction de la structure secondaire

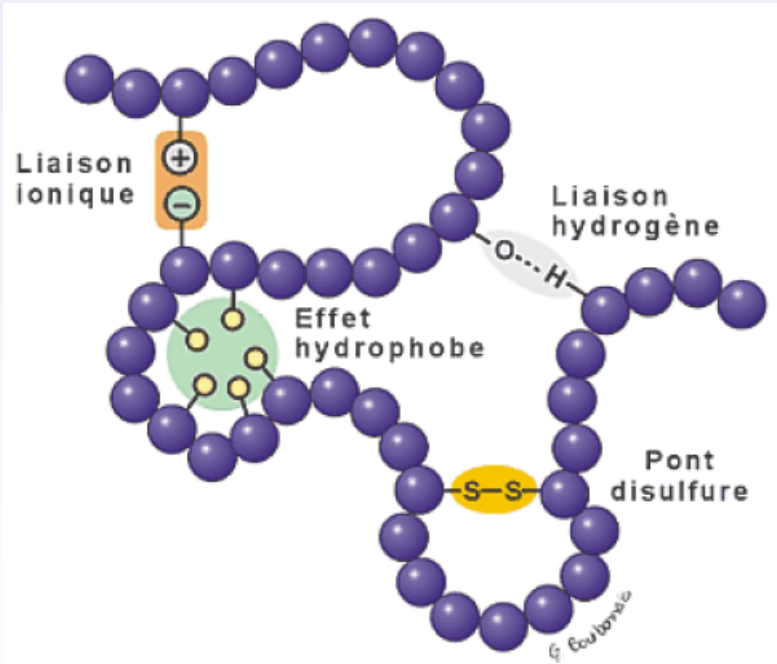
- Formes caractéristiques fréquentes : hélices α, brins β, angles
- Méthodes performantes [Ruan et al., 2005, Cheng et Baldi, 2007]



Contacts entre deux acides aminés distants

Différents types de contacts

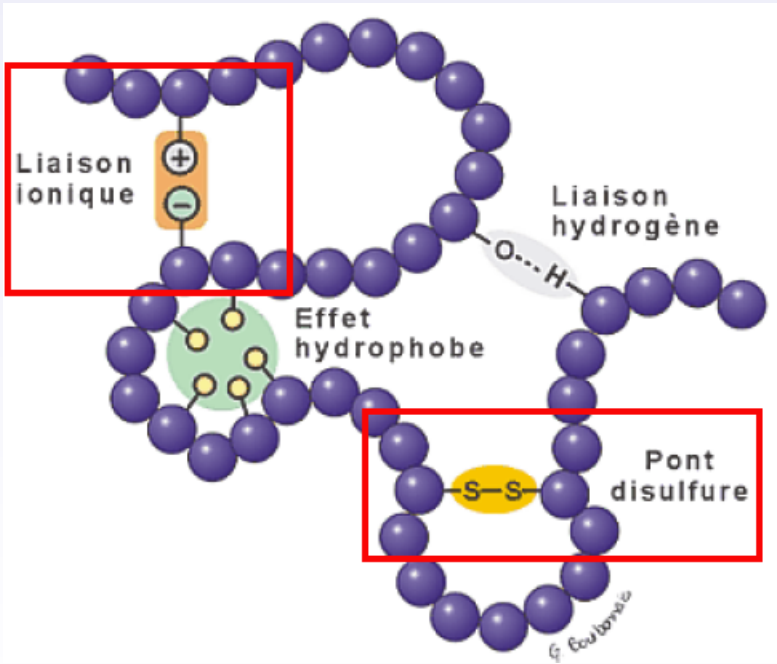
- Ponts disulfures (liaisons covalentes entre deux cystéines)
- Ponts salins (liaisons ioniques)
- Liaisons hydrogène
- ...



Contacts entre deux acides aminés distants

Différents types de contacts

- Ponts disulfures (liaisons covalentes entre deux cystéines)
- Ponts salins (liaisons ioniques)
- Liaisons hydrogène
- ...



Classification supervisée

Description

- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$
- P sur $\mathcal{X} \times \mathcal{Y}$, supposée fixe mais inconnue
- $\mathcal{X}$: espace de description, $\mathcal{Y}$: classes $(= \{+, -\})$

Prédiction des ponts disulfures

- $\mathcal{X}$ = descriptions d'une paire de cystéines oxydées
- $\mathcal{Y}$ = appariée / non appariée

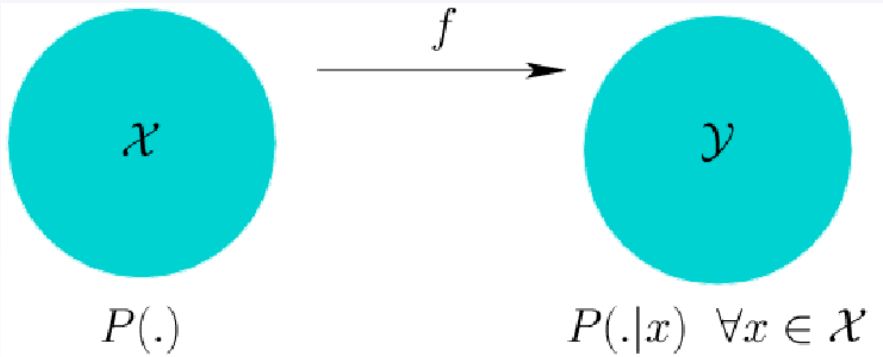
Classification supervisée

Description

- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$
- P sur $\mathcal{X} \times \mathcal{Y}$, supposée fixe mais inconnue
- $\mathcal{X}$: espace de description, $\mathcal{Y}$: classes $(= \{+, -\})$

Objectif

- Calculer, à partir de S, un classifieur $f : \mathcal{X} \to \mathcal{Y}$
- Critère : minimiser $R(f) = P(f(x) \neq y)$



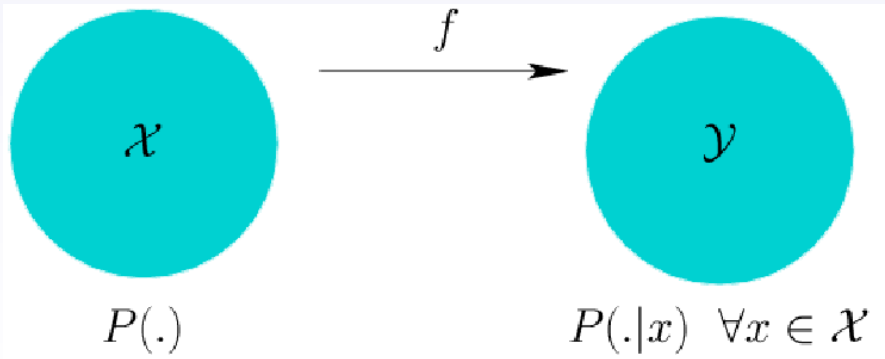
Classification supervisée

Description

- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$
- P sur $\mathcal{X} \times \mathcal{Y}$, supposée fixe mais inconnue
- $\mathcal{X}$: espace de description, $\mathcal{Y}$: classes $(= \{+, -\})$

Solution optimale : la règle de Bayes

- $\forall x \in \mathcal{X}$, $f_{Bayes}(x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P(y|x)$
- P inconnue et S ne permet généralement pas de l'approximer



« Variantes » de la classification supervisée

Classification semi-supervisée

- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$
- $S_{unl} = \{x'_1, ..., x'_{l'}\}$ i.i.d. selon $P(.)$ sur $\mathcal{X}$

« Variantes » de la classification supervisée

Classification semi-supervisée

- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$
- $S_{unl} = \{x'_1, ..., x'_{l'}\}$ i.i.d. selon $P(.)$ sur $\mathcal{X}$

Classification supervisée avec présence de bruit de classification

- $S^\eta = \{(x_1, y_1^\eta), ..., (x_l, y_l^\eta)\}$
- Les classes originales y_i sont corrompues avant d'être observées
- Différents modèles de bruit de classification
- **Ex :** bruit de classification uniforme CN [Angluin et Laird, 1988]

  $y_i^\eta \neq y_i$ avec probabilité $\eta < 0.5$ constante ($\mathcal{Y} = \{+, -\}$)

« Variantes » de la classification supervisée

Classification semi-supervisée

- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$
- $S_{unl} = \{x'_1, ..., x'_{l'}\}$ i.i.d. selon $P(.)$ sur $\mathcal{X}$

Classification supervisée avec présence de bruit de classification

- $S^\eta = \{(x_1, y_1^\eta), ..., (x_l, y_l^\eta)\}$
- Les classes originales y_i sont corrompues avant d'être observées
- Différents modèles de bruit de classification
- **Ex :** bruit de classification uniforme CN [Angluin et Laird, 1988]

  $y_i^\eta \neq y_i$ avec probabilité $\eta < 0.5$ constante ($\mathcal{Y} = \{+, -\}$)

Objectif

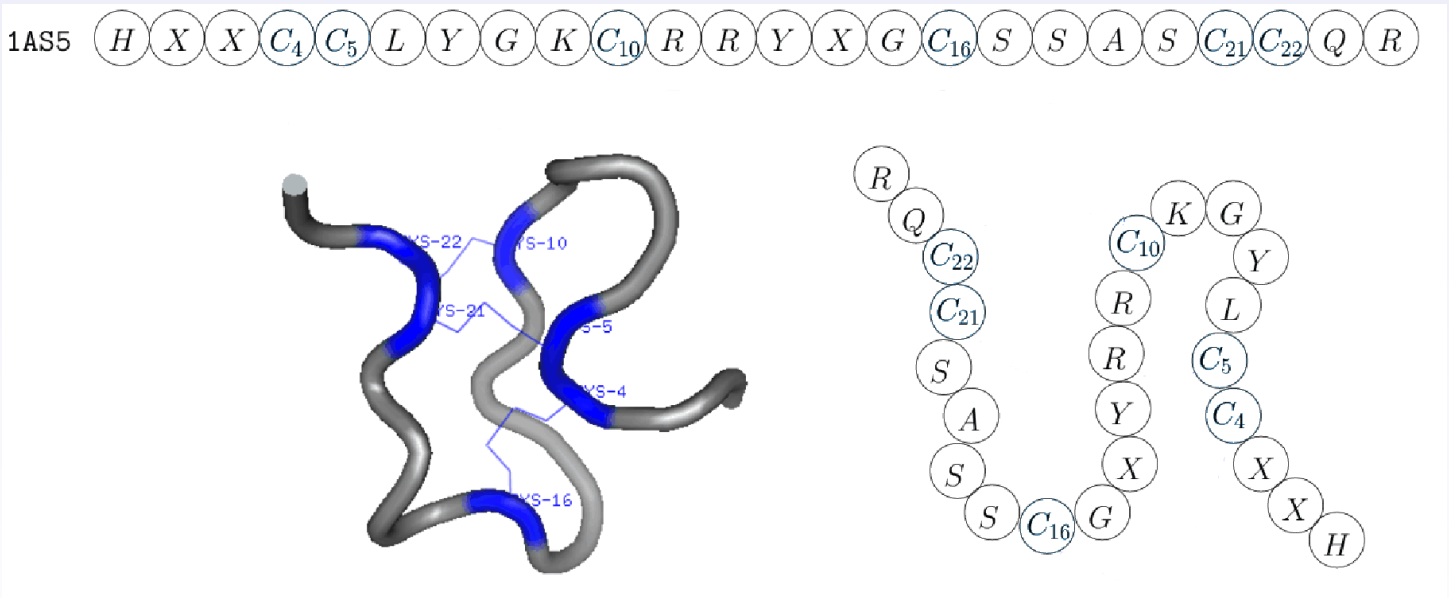
Identique au cas supervisé classique

$\mathcal{I}$

Première étude du problème de la prédiction des ponts disulfures et contribution à l'apprentissage semi-supervisé asymétrique

Prédiction des ponts disulfures

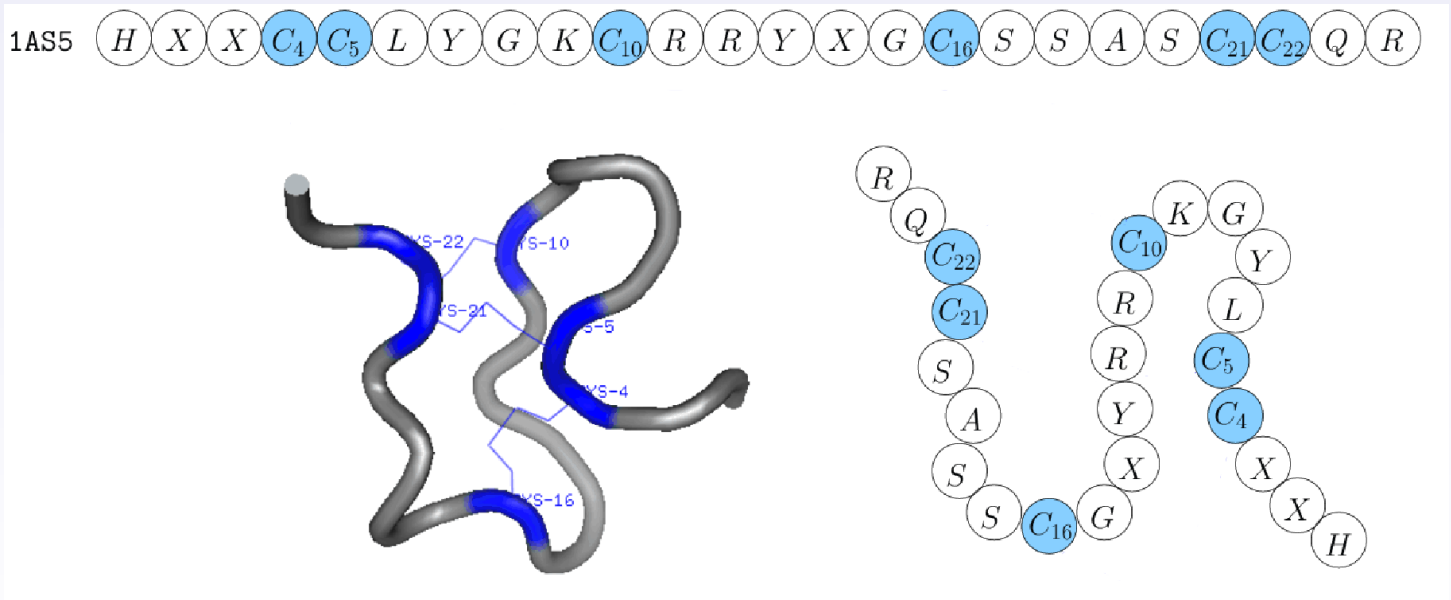
Deux étapes de prédiction :



Prédiction des ponts disulfures

Deux étapes de prédiction :

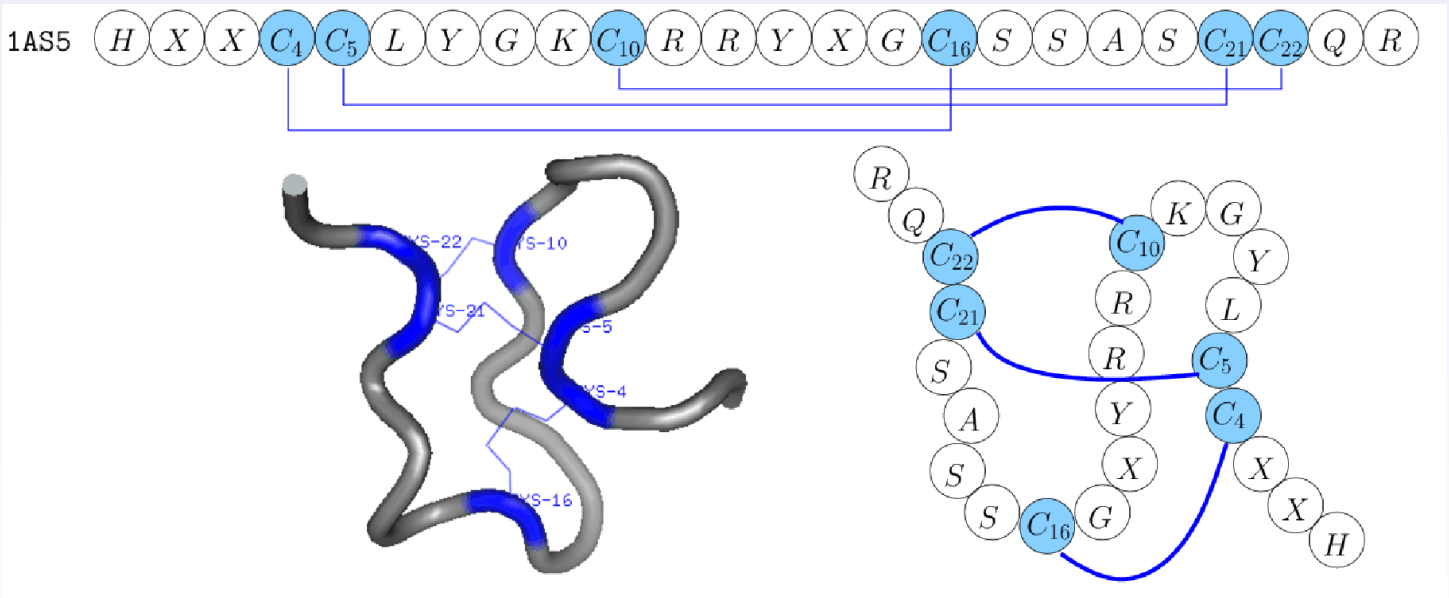
- Quelles cystéines forment un pont ? $\Rightarrow$ traitée efficacement



Prédiction des ponts disulfures

Deux étapes de prédiction :

- Quelles cystéines forment un pont ? ⇒ traitée efficacement
- Les ponts eux-mêmes (appariements) ⇒ problème ouvert



Prédiction des ponts disulfures

Travaux de référence sur le problème

- [Fariselli et al., 2001, 2002, Vullo et Frasconi, 2003, 2004] ...
- De fortes similarités dans tous ces travaux

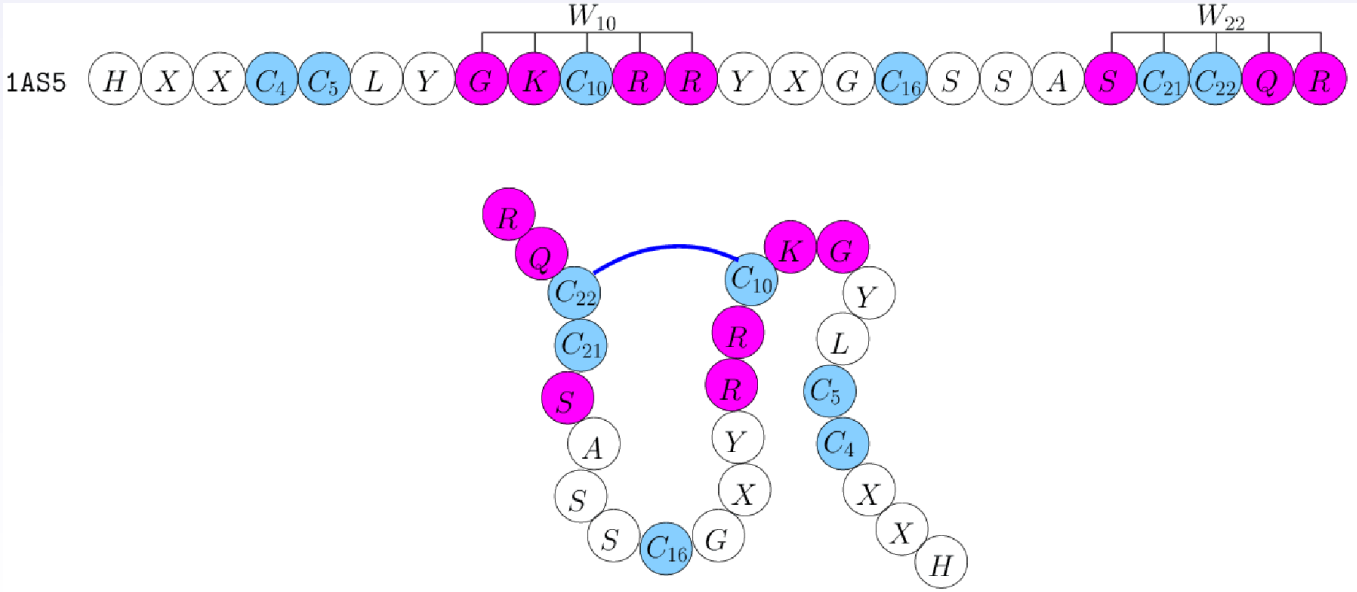
Prédiction des ponts disulfures

Travaux de référence sur le problème

- [Fariselli et al., 2001, 2002, Vullo et Frasconi, 2003, 2004] ...
- De fortes similarités dans tous ces travaux

Le choix de la description des paires de cystéines oxydées

- Principalement : l'environnement local des cystéines sur la séquence
- Une paire (C_i, C_j) décrite par (W_i, W_j)
- (W_i, W_j) codées sous forme vectorielle



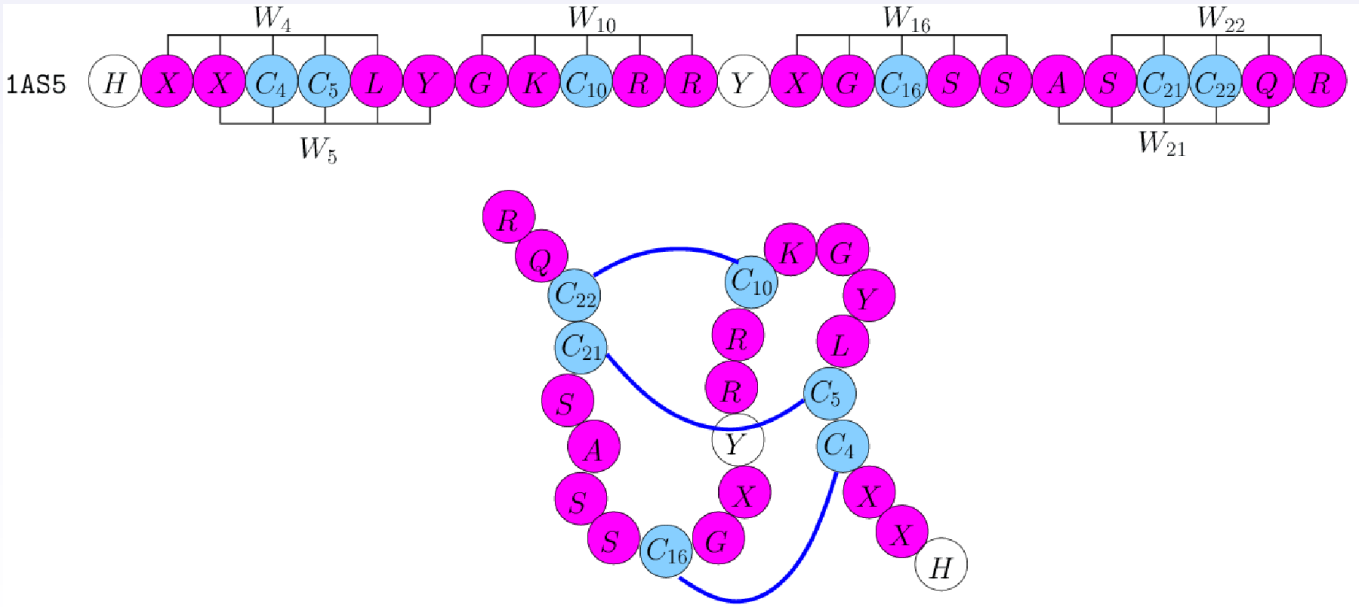
Prédiction des ponts disulfures

Travaux de référence sur le problème

- [Fariselli et al., 2001, 2002, Vullo et Frasconi, 2003, 2004] ...
- De fortes similarités dans tous ces travaux

Notre première étude

- Pourquoi ce choix ? caractériser les paires « compatibles » des autres
- Aucune preuve biologique, certains biologistes sceptiques
- Quelle implication sur le statut des paires non appariées ?



Notion d'affinité entre contextes locaux

Statut des paires observées

- **Les paires appariées :** des exemples positifs de paires compatibles
- **Les paires non appariées ?**
  - Contraintes d'unicité d'appariement et globales sur la structure
  - Des exemples négatifs de paires compatibles ?

Notion d'affinité entre contextes locaux

Statut des paires observées

- **Les paires appariées :** des exemples positifs de paires compatibles
- **Les paires non appariées ?**
  - Contraintes d'unicité d'appariement et globales sur la structure
  - Des exemples négatifs de paires compatibles ?

Notre hypothèse

- **Paires non appariées :** exemples de classe non déterminée
- **Données observées :** positives et non étiquetées
- Apprentissage semi-supervisé asymétrique [Denis, 1998]

Notion d'affinité entre contextes locaux

Statut des paires observées

- **Les paires appariées :** des exemples positifs de paires compatibles
- **Les paires non appariées ?**
  - Contraintes d'unicité d'appariement et globales sur la structure
  - Des exemples négatifs de paires compatibles ?

Notre hypothèse

- **Paires non appariées :** exemples de classe non déterminée
- **Données observées :** positives et non étiquetées
- Apprentissage semi-supervisé asymétrique [Denis, 1998]

Idée

- Comparer les performances de classifieurs appris à partir :
  - de la modélisation supervisée
  - de la modélisation semi-supervisée asymétrique
- peu de résultats exploitables en contexte semi-supervisé asymétrique
  $\Rightarrow$ nécessite une étude de ce contexte

Apprentissage semi-supervisé asymétrique

Cadre général de l'apprentissage statistique

- La distribution P sur $\mathcal{X} \times \mathcal{Y}$ est déterminée par :
  - $P(.)$ sur $\mathcal{X}$
  - $P(.|y = +)$ sur $\mathcal{X}$ (notée $P(.|+)$)
  - $P(y = +)$ (notée $P(+)$)

Apprentissage semi-supervisé asymétrique

- Suppose que l'on dispose :
  - d'exemples positifs, tirés selon $P(.|+)$ sur $\mathcal{X}$
  - d'exemples non étiquetés, tirés selon $P(.)$ sur $\mathcal{X}$
- si $P(.|+)$ et $P(.) \Rightarrow P(+)$ alors $P(.|+)$ et $P(.) \Rightarrow P$ sur $\mathcal{X} \times \mathcal{Y}$

Apprentissage semi-supervisé asymétrique

Cadre général de l'apprentissage statistique

- La distribution P sur $\mathcal{X} \times \mathcal{Y}$ est déterminée par :
  - $P(.)$ sur $\mathcal{X}$
  - $P(.|y = +)$ sur $\mathcal{X}$ (notée $P(.|+)$)
  - $P(y = +)$ (notée $P(+)$)

Apprentissage semi-supervisé asymétrique

- Suppose que l'on dispose :
  - d'exemples positifs, tirés selon $P(.|+)$ sur $\mathcal{X}$
  - d'exemples non étiquetés, tirés selon $P(.)$ sur $\mathcal{X}$
- si $P(.|+)$ et $P(.) \Rightarrow P(+)$ alors $P(.|+)$ et $P(.) \Rightarrow P$ sur $\mathcal{X} \times \mathcal{Y}$

Proposition

Sans information supplémentaire, $P(+)$ n'est pas déterminée par $P(.)$ sur $\mathcal{X}$ et $P(.|+)$ sur $\mathcal{X}$

Apprentissage semi-supervisé asymétrique

Cadre général de l'apprentissage statistique

- La distribution P sur $\mathcal{X} \times \mathcal{Y}$ est déterminée par :
  - $P(.)$ sur $\mathcal{X}$
  - $P(.|y = +)$ sur $\mathcal{X}$ (notée $P(.|+)$)
  - $P(y = +)$ (notée $P(+)$)

Apprentissage semi-supervisé asymétrique

- Suppose que l'on dispose :
  - d'exemples positifs, tirés selon $P(.|+)$ sur $\mathcal{X}$
  - d'exemples non étiquetés, tirés selon $P(.)$ sur $\mathcal{X}$
- si $P(.|+)$ et $P(.) \Rightarrow P(+)$ alors $P(.|+)$ et $P(.) \Rightarrow P$ sur $\mathcal{X} \times \mathcal{Y}$

Proposition

Sans information supplémentaire, $P(+)$ n'est pas déterminée par $P(.)$ sur $\mathcal{X}$ et $P(.|+)$ sur $\mathcal{X}$: c'est un problème mal posé

Hypothèses rendant le problème bien posé

Problèmes déterministes : $\forall x \in \mathcal{X},\ P(+|x) = 1$ *ou* $P(+|x) = 0$

$$P(+) = \sum_{x \in \mathcal{X} / P(x|+) \neq 0} P(x)$$

Hypothèses rendant le problème bien posé

Problèmes déterministes : $\forall x \in \mathcal{X}$, $P(+|x) = 1$ *ou* $P(+|x) = 0$

$$P(+) = \sum_{x \in \mathcal{X} / P(x|+) \neq 0} P(x)$$

Hypothèse « naïve » de Bayes : $P(.|y)$ *distributions produits*

- $P(x|y) = \prod_{i=1}^{m} P(x^i|y)$ $\forall x = (x^1, ..., x^m) \in \mathcal{X} = \prod_{i=1}^{m} \mathcal{X}^i$

Hypothèses rendant le problème bien posé

Problèmes déterministes : $\forall x \in \mathcal{X},\ P(+|x) = 1$ *ou* $P(+|x) = 0$

$$P(+) = \sum_{x \in \mathcal{X} / P(x|+) \neq 0} P(x)$$

Hypothèse « naïve » de Bayes : $P(.|y)$ *distributions produits*

- $P(x|y) = \prod_{i=1}^{m} P(x^i|y)\ \forall x = (x^1, ..., x^m) \in \mathcal{X} = \prod_{i=1}^{m} \mathcal{X}^i$
- $P(.) = \alpha P(.|+) + (1-\alpha) P(.|-)$ avec $\alpha = P(+)$, or les mélanges finis de distributions produits sont identifiables [Yakowitz et Spragins, 1968]

Hypothèses rendant le problème bien posé

Problèmes déterministes : $\forall x \in \mathcal{X}$, $P(+|x) = 1$ ou $P(+|x) = 0$

$$P(+) = \sum_{x \in \mathcal{X} / P(x|+) \neq 0} P(x)$$

Hypothèse « naïve » de Bayes : $P(.|y)$ distributions produits

- $P(x|y) = \prod_{i=1}^{m} P(x^i|y)$ $\forall x = (x^1, ..., x^m) \in \mathcal{X} = \prod_{i=1}^{m} \mathcal{X}^i$
- $P(.) = \alpha P(.|+) + (1 - \alpha) P(.|-)$ avec $\alpha = P(+)$, or les mélanges finis de distributions produits sont identifiables [Yakowitz et Spragins, 1968]
- Nous montrons que $P(.|+)$ et $P(.) \Rightarrow P(+)$
  - sous des conditions plus faibles
  - prise en compte de la distribution $P(.|+)$ sur $\mathcal{X}$

Hypothèses rendant le problème bien posé

Problèmes déterministes : $\forall x \in \mathcal{X},\ P(+|x) = 1$ *ou* $P(+|x) = 0$

$$P(+) = \sum_{x \in \mathcal{X} / P(x|+) \neq 0} P(x)$$

Hypothèse « naïve » de Bayes : $P(.|y)$ *distributions produits*

- $P(x|y) = \prod_{i=1}^{m} P(x^i|y)\ \forall x = (x^1, ..., x^m) \in \mathcal{X} = \prod_{i=1}^{m} \mathcal{X}^i$
- $P(.) = \alpha P(.|+) + (1-\alpha) P(.|-)$ avec $\alpha = P(+)$, or les mélanges finis de distributions produits sont identifiables [Yakowitz et Spragins, 1968]
- Nous montrons que $P(.|+)$ et $P(.) \Rightarrow P(+)$
  - sous des conditions plus faibles
  - prise en compte de la distribution $P(.|+)$ sur $\mathcal{X}$

$$P(+) = \frac{P(x^i{=}k, x^j{=}l) - P(x^i{=}k)P(x^j{=}l)}{P(x^i{=}k|+)P(x^j{=}l|+) - P(x^i{=}k)P(x^j{=}l|+) - P(x^j{=}l)P(x^i{=}k|+) + P(x^i{=}k, x^j{=}l)}$$

[Magnan, CAp 2005, RIA 2006]

Conséquence

Hypothèse naïve de Bayes et classifieur de Bayes

- Le classifieur de Bayes devient le classifieur naïf de Bayes :

$$f_{NB}(x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P(y|x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P(y) \cdot \prod_{i=1}^{m} P(x^i|y)$$

- Classifieur simple mais efficace [Domingos et Pazzani, 1996]

Conséquence

Hypothèse naïve de Bayes et classifieur de Bayes

- Le classifieur de Bayes devient le classifieur naïf de Bayes :

$$f_{NB}(x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P(y|x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P(y) \cdot \prod_{i=1}^{m} P(x^i|y)$$

- Classifieur simple mais efficace [Domingos et Pazzani, 1996]

Classifieur Naïf de Bayes

- Spécifié par les paramètres :
  - $P(+)$
  - $p_{ik} = P(x^i = k|+)$
  - $q_{ik} = P(x^i = k|-)$
- $\theta = \{P(+), p_{ik}, q_{ik}\}$ est appelé *modèle naïf de Bayes*

Conséquence

Hypothèse naïve de Bayes et classifieur de Bayes

- Le classifieur de Bayes devient le classifieur naïf de Bayes :

$$f_{NB}(x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P(y|x) = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P(y) \cdot \prod_{i=1}^{m} P(x^i|y)$$

- Classifieur simple mais efficace [Domingos et Pazzani, 1996]

Classifieur Naïf de Bayes

- Spécifié par les paramètres :
  - $P(+)$
  - $p_{ik} = P(x^i = k|+)$
  - $q_{ik} = P(x^i = k|-)$
- $\theta = \{P(+), p_{ik}, q_{ik}\}$ est appelé *modèle naïf de Bayes*
- Tous les paramètres de ces modèles sont déterminés en contexte semi-supervisé asymétrique

Algorithme NB-SSA

NB-SSA : adaptation semi-supervisée asymétrique de NB

- **Entrée :** S_{pos}, S_{unl}
- **Sortie :** $\hat{\theta} = \{\hat{P}(+),\ \hat{p}_{ik},\ \hat{q}_{ik}\ \}$
- $S_{pos} \rightarrow \hat{p}_{ik}$ $\quad$ $\hat{p}_{ik}$ et $S_{unl} \rightarrow \hat{P}(+)$ $\quad$ $\hat{p}_{ik}$, S_{unl} et $\hat{P}(+) \rightarrow \hat{q}_{ik}$

[Magnan, CAp 2005, RIA 2006]

$$\hat{P}(+) = \frac{\sum_{\substack{i,j \in \{1,\ldots,m\}, i \neq j \\ k \in \mathcal{X}^i, l \in X^j}} \hat{P}(x^i{=}k, x^j{=}l) - \hat{P}(x^i{=}k)\hat{P}(x^j{=}l)}{\sum_{\substack{i,j \in \{1,\ldots,m\}, i \neq j \\ k \in \mathcal{X}^i, l \in X^j}} \hat{p}_{ik}\hat{p}_{jl} - \hat{P}(x^i{=}k)\hat{p}_{jl} - \hat{P}(x^j{=}l)\hat{p}_{ik} + \hat{P}(x^i{=}k, x^j{=}l)}$$

$$\hat{q}_{ik} = \frac{\hat{P}(x^i = k) - \hat{p}_{ik}\hat{P}(+)}{1 - \hat{P}(+)}$$

Maximisation de la Vraisemblance

NB-SSA-EM : NB-SSA + Méthode E.M. [Dempster et al., 1977]

- Vraisemblance : $L(\theta, S) = \prod_{z_i \in S} P(z_i|\theta)$
- EM : méthode itérative pour obtenir un maximum local de $L(\theta, S)$
- **Modèle initial θ_0 :** calculé avec `NB-SSA`
- **Modèles θ_{n+1} $(n \geq 0)$:** obtenus à partir de θ_n, S_{pos} et S_{unl}

[Magnan, CAp 2005, RIA 2006]

Maximisation de la Vraisemblance

NB-SSA-EM : NB-SSA + Méthode E.M. [Dempster et al., 1977]

- Vraisemblance : $L(\theta, S) = \prod_{z_i \in S} P(z_i|\theta)$
- EM : méthode itérative pour obtenir un maximum local de $L(\theta, S)$
- **Modèle initial θ_0 :** calculé avec `NB-SSA`
- **Modèles θ_{n+1} $(n \geq 0)$:** obtenus à partir de θ_n, S_{pos} et S_{unl}

[Magnan, CAp 2005, RIA 2006]

$$P(+) = \frac{l + \sum_{i=1}^{l'} \hat{P}(y'_i = +|x'_i, \theta_n)}{l + l'}$$

$$p_{jk} = \frac{n_{jk} + \sum_{x'_i \in S_{unl} / x'^{j}_i = k} \hat{P}(y'_i = +|x'_i, \theta_n)}{l + \sum_{i=1}^{l'} \hat{P}(y'_i = +|x'_i, \theta_n)} \qquad q_{jk} = \frac{\sum_{x'_i \in S_{unl} / x'^{j}_i = k} \hat{P}(y'_i = -|x'_i, \theta_n)}{\sum_{i=1}^{l'} \hat{P}(y'_i = -|x'_i, \theta_n)}$$

Expériences sur données ponts disulfures

Description

- Protocole identique à celui des travaux du domaine
- Jeu de données préparé par C. Geourjon (IBCP, Lyon)
- Nombreuses séries (rayon des fenêtres, codage)
- Comparaison avec un choix aléatoire des ponts

Description du jeu de données	
Nb de ponts	Nb de protéines
2	77
3	64
4	33
5	24

Expériences sur données ponts disulfures

Description

- Protocole identique à celui des travaux du domaine
- Jeu de données préparé par C. Geourjon (IBCP, Lyon)
- Nombreuses séries (rayon des fenêtres, codage)
- Comparaison avec un choix aléatoire des ponts

Nombre de ponts par protéine	Choix aléatoire des ponts	Performances avec NB	Performances avec NBSSA-EM
2	33.3%	40.2%	**58.8%**
3	20%	17.5%	**33.4%**
4	14.3%	12.7%	**16.3%**
5	11.1%	5.8%	**13.2%**

Expériences sur données ponts disulfures

Description

- Protocole identique à celui des travaux du domaine
- Jeu de données préparé par C. Geourjon (IBCP, Lyon)
- Nombreuses séries (rayon des fenêtres, codage)
- Comparaison avec un choix aléatoire des ponts

Conclusions

- Résultats encourageants, mais...
- $\neq$ non validées par un test statistique avec grande confiance
- Ne permettent pas de valider notre hypothèse
- Ne montrent pas qu'une information ait été détectée
  $\Rightarrow$ la question de l'existence d'une information locale impliquée dans la formation de ponts doit être étudiée

II

La question de l'existence d'une information locale impliquée dans les appariements d'acides aminés

Motivations

Répondre aux questions

- Les environnements locaux d'acides aminés appariés jouent-ils un rôle dans la formation de ponts ?
- Existe-il des affinités + ou - fortes entre ces contextes locaux ?
- Comment le prouver ?

Motivations

Répondre aux questions

- Les environnements locaux d'acides aminés appariés jouent-ils un rôle dans la formation de ponts ?
- Existe-il des affinités + ou - fortes entre ces contextes locaux ?
- Comment le prouver ?

Intérêt

Savoir si cette information est pertinente pour prédire les ponts

Une première approche

Notion d'affinité entre paires d'environnements locaux

- $\Omega_r = \Sigma^{2r+1} = \{\text{segments de protéines de taille } 2r+1\}$
- $P(B(w, w')|w, w', n)$: probabilité que w et w' soient appariés sachant que ce sont des environnements locaux d'a.a. d'une protéine avec n ponts

Une première approche

Notion d'affinité entre paires d'environnements locaux

- $\Omega_r = \Sigma^{2r+1} = \{\text{segments de protéines de taille } 2r+1\}$
- $P(B(w,w')|w,w',n)$: probabilité que w et w' soient appariés sachant que ce sont des environnements locaux d'a.a. d'une protéine avec n ponts

Une première approche de l'information locale

- Considérons des protéines à n ponts :

$$P(B(w,w')|w,w',n) = \frac{1}{2n-1} \Leftrightarrow \text{pas d'information locale}$$

Une première approche

Notion d'affinité entre paires d'environnements locaux

- $\Omega_r = \Sigma^{2r+1} = \{\text{segments de protéines de taille } 2r+1\}$
- $P(B(w,w')|w,w',n)$: probabilité que w et w' soient appariés sachant que ce sont des environnements locaux d'a.a. d'une protéine avec n ponts

Une première approche de l'information locale

- Considérons des protéines à n ponts :

$$P(B(w,w')|w,w',n) = \frac{1}{2n-1} \Leftrightarrow \text{pas d'information locale}$$

- Méthode statistique simple pour décider si une information locale existe
- Estimation directe de ces probabilités impossible :
  avec $r = 3 \rightarrow |\{(w,w'), w, w' \in \Omega_r\}| = 20^{12} \simeq 4 \cdot 10^{15}$!
- Seules quelques centaines d'exemples disponibles...

Une approche simplificatrice et raisonnable

Idée

- Il n'est pas nécessaire de connaître $P(B(w, w')|w, w', n)$
- On veut montrer que certaines paires ont une propension plus forte à se retrouver appariées par leurs acides aminés centraux que d'autres

Une approche simplificatrice et raisonnable

Idée

- Il n'est pas nécessaire de connaître $P(B(w, w')|w, w', n)$
- On veut montrer que certaines paires ont une propension plus forte à se retrouver appariées par leurs acides aminés centraux que d'autres

Solution proposée : une fonction d'affinité

- Supposer l'existence d'une fonction $g : \Omega_r^2 \rightarrow \mathcal{Y}$ avec $|\mathcal{Y}|$ petit

Une approche simplificatrice et raisonnable

Idée

- Il n'est pas nécessaire de connaître $P(B(w, w')|w, w', n)$
- On veut montrer que certaines paires ont une propension plus forte à se retrouver appariées par leurs acides aminés centraux que d'autres

Solution proposée : une fonction d'affinité

- Supposer l'existence d'une fonction $g : \Omega_r^2 \rightarrow \mathcal{Y}$ avec $|\mathcal{Y}|$ petit
- g « regroupe » les paires (w, w') similaires :

$$g(w_1, w_2) = g(w_1', w_2') \Rightarrow P(B(w_1, w_2)|w_1, w_2, n) \simeq P(B(w_1', w_2')|w_1', w_2', n)$$

Une approche simplificatrice et raisonnable

Idée

- Il n'est pas nécessaire de connaître $P(B(w, w')|w, w', n)$
- On veut montrer que certaines paires ont une propension plus forte à se retrouver appariées par leurs acides aminés centraux que d'autres

Solution proposée : une fonction d'affinité

- Supposer l'existence d'une fonction $g : \Omega_r^2 \rightarrow \mathcal{Y}$ avec $|\mathcal{Y}|$ petit
- g « regroupe » les paires (w, w') similaires :

$$g(w_1, w_2) = g(w_1', w_2') \Rightarrow P(B(w_1, w_2)|w_1, w_2, n) \simeq P(B(w_1', w_2')|w_1', w_2', n)$$

- en niveaux d'affinité :

$$y < y' \Rightarrow P(B(w_1, w_2)|g(w_1, w_2) = y) < P(B(w_1', w_2')|g(w_1', w_2') = y')$$

Une approche simplificatrice et raisonnable

Idée

- Il n'est pas nécessaire de connaître $P(B(w, w')|w, w', n)$
- On veut montrer que certaines paires ont une propension plus forte à se retrouver appariées par leurs acides aminés centraux que d'autres

Solution proposée : une fonction d'affinité

- Supposer l'existence d'une fonction $g : \Omega_r^2 \rightarrow \mathcal{Y}$ avec $|\mathcal{Y}|$ petit
- g « regroupe » les paires (w, w') similaires :

$$g(w_1, w_2) = g(w_1', w_2') \Rightarrow P(B(w_1, w_2)|w_1, w_2, n) \simeq P(B(w_1', w_2')|w_1', w_2', n)$$

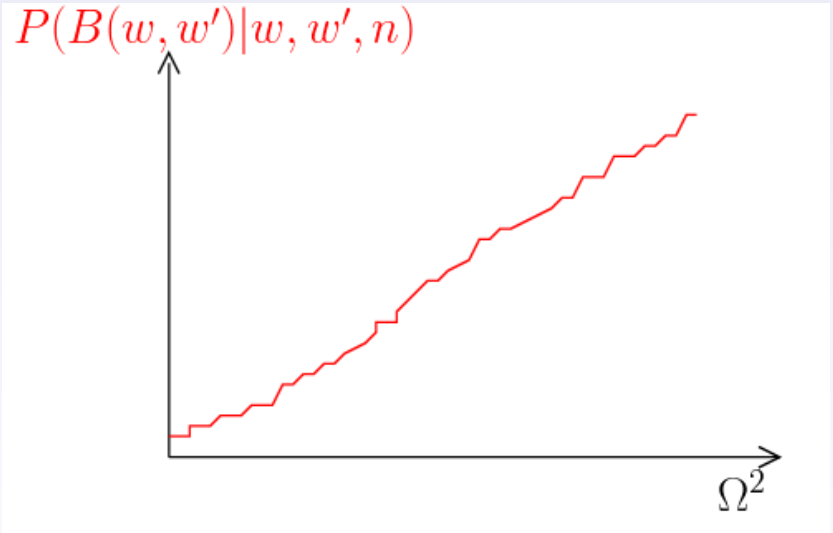
- en niveaux d'affinité :

$$y < y' \Rightarrow P(B(w_1, w_2)|g(w_1, w_2) = y) < P(B(w_1', w_2')|g(w_1', w_2') = y')$$

- Dans ce cas, seules les $P(B(w, w')|g(w, w') = y, n)$ sont à étudier
- L'équivalence donnée dans la première approche reste valable

Cas le plus simple : $\mathcal{Y} = \{0, 1\}$

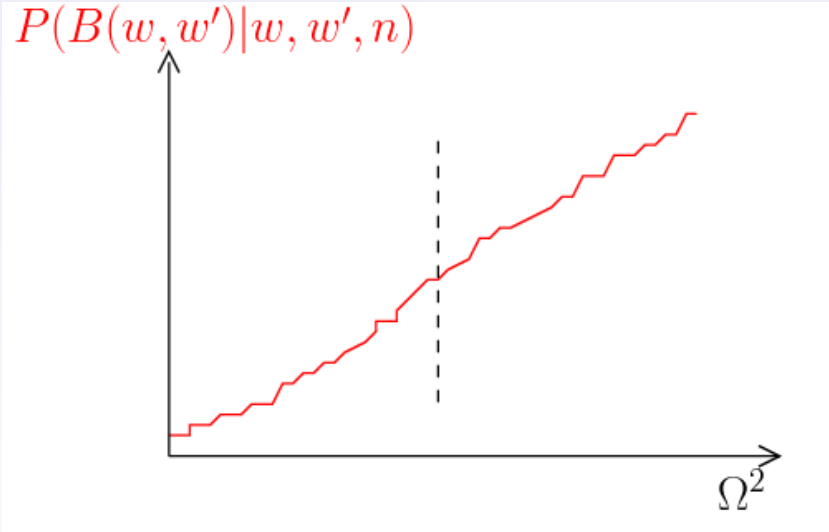
Conséquences



Cas le plus simple : $\mathcal{Y} = \{0, 1\}$

Conséquences

- Les paires (w, w') sont partitionnées en deux classes

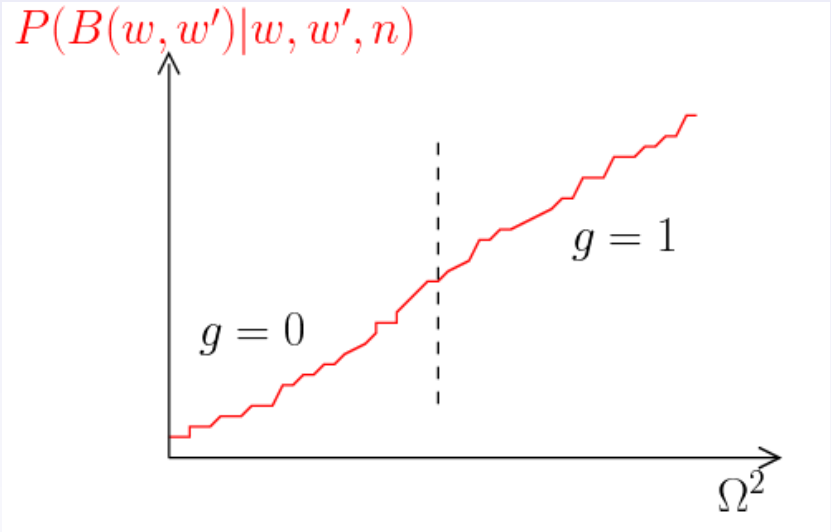


$$g(w_1, w_2) = g(w_1', w_2') \Rightarrow P(B(w_1, w_2)|w_1, w_2, n) \simeq P(B(w_1', w_2')|w_1', w_2', n)$$

Cas le plus simple : $\mathcal{Y} = \{0, 1\}$

Conséquences

- Les paires (w, w') sont partitionnées en deux classes
- Chaque classe correspond à un niveau d'affinité

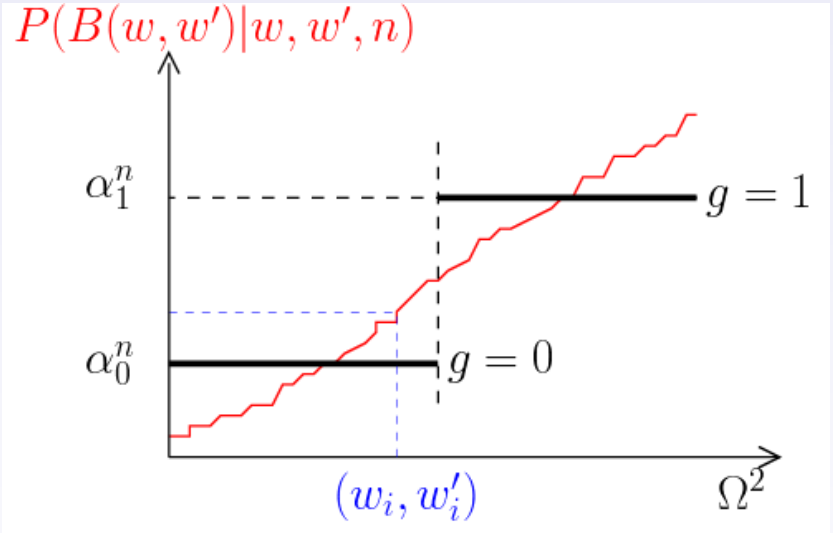


$$y < y' \Rightarrow P(B(w_1, w_2)|g(w_1, w_2) = y) < P(B(w_1', w_2')|g(w_1', w_2') = y')$$

Cas le plus simple : $\mathcal{Y} = \{0, 1\}$

Conséquences

- Les paires (w, w') sont partitionnées en deux classes
- Chaque classe correspond à un niveau d'affinité

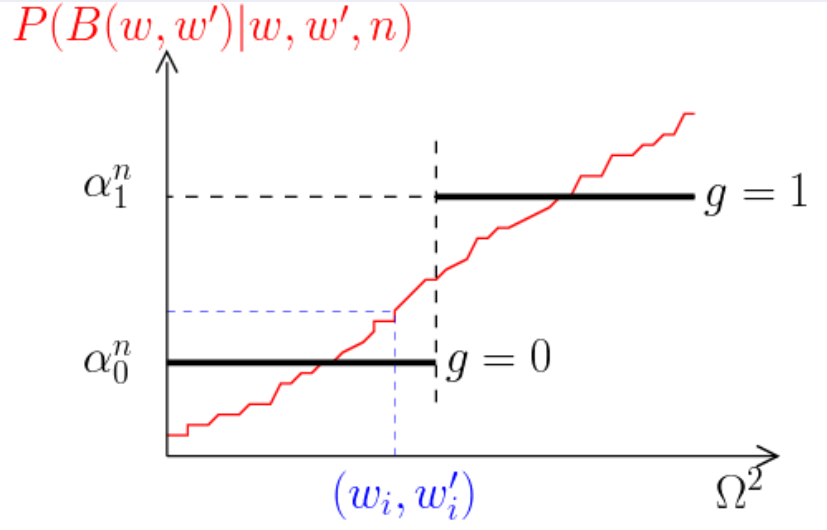


$$P(B(w,w')|g(w,w'),n) = \begin{cases} \alpha_1^n \text{ if } g(w,w') = 1 \\ \alpha_0^n \text{ if } g(w,w') = 0 \end{cases}$$

Cas le plus simple : $\mathcal{Y} = \{0, 1\}$

Conséquences

- Les paires (w, w') sont partitionnées en deux classes
- Chaque classe correspond à un niveau d'affinité

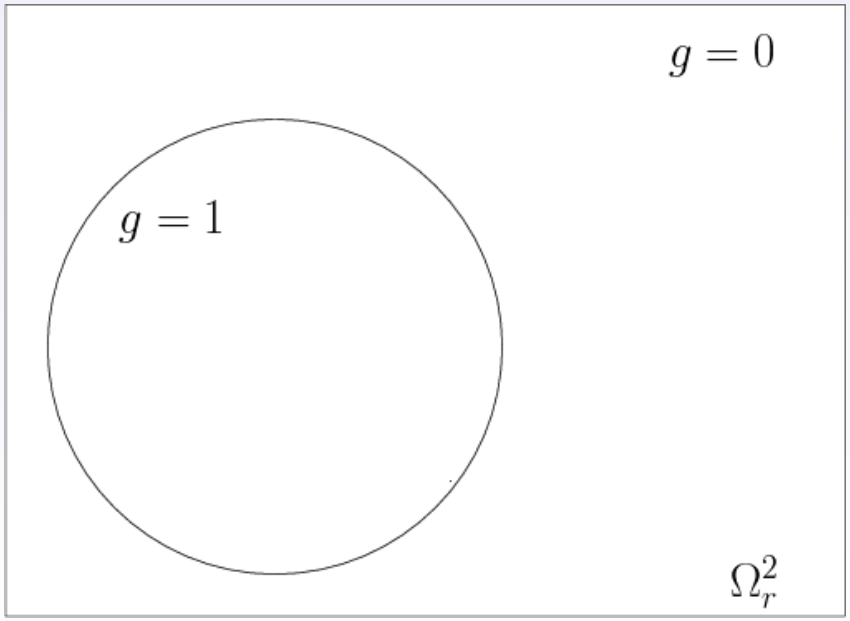


Si une information locale existe : on doit avoir $\alpha_1^n > \alpha_0^n$
Dans le cas contraire : $\alpha_1^n = \alpha_0^n = \frac{1}{2n-1}$

Propriétés induites par g

Une fonction accessible, mais pas directement

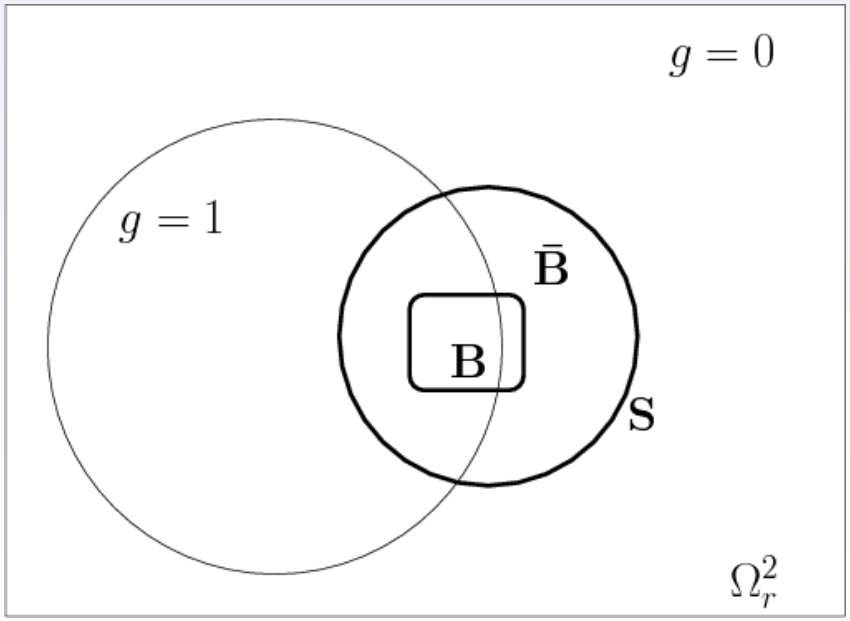
- Chaque paire (w, w') n'appartient qu'à une seule classe attribuée par g



Propriétés induites par g

Une fonction accessible, mais pas directement

- Chaque paire (w, w') n'appartient qu'à une seule classe attribuée par g
- Les observations dont on dipose (paires appariées et non appariées) en fournissent alors une vue indirecte

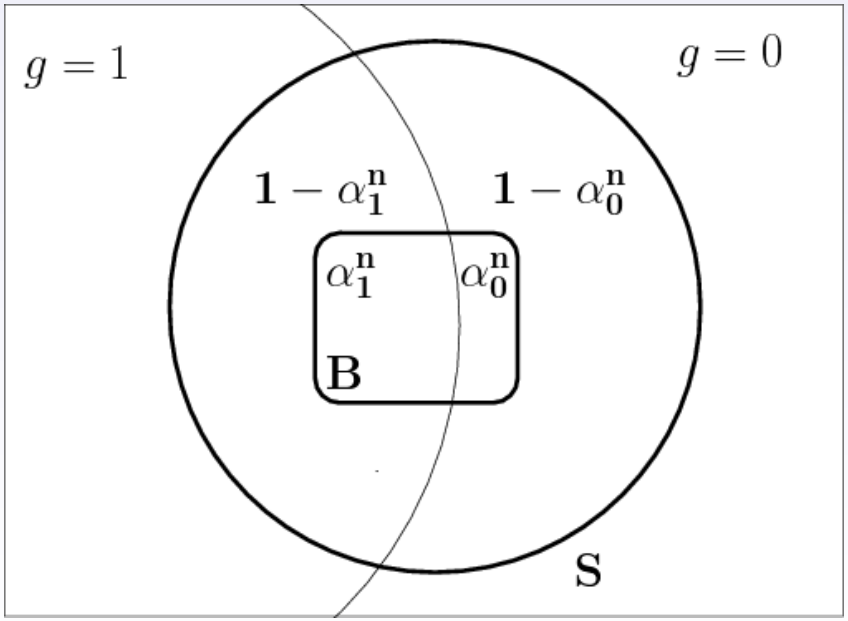


Propriétés induites par g

Une fonction accessible, mais pas directement

- Chaque paire (w, w') n'appartient qu'à une seule classe attribuée par g
- Les observations dont on dipose (paires appariées et non appariées) en fournissent alors une vue indirecte

$$P(B(w,w')|g(w,w'),n) = \begin{cases} \alpha_1^n \text{ if } g(w,w') = 1 \\ \alpha_0^n \text{ if } g(w,w') = 0 \end{cases}$$

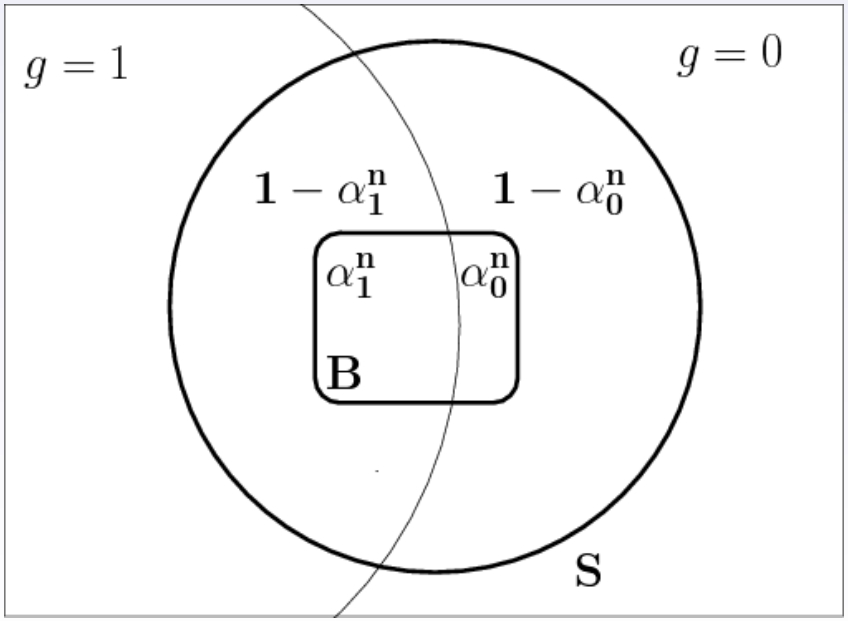


Propriétés induites par g

Une fonction accessible, mais pas directement

- Chaque paire (w, w') n'appartient qu'à une seule classe attribuée par g
- Les observations dont on dipose (paires appariées et non appariées) en fournissent alors une vue indirecte

$$\begin{cases} g = 1 \text{ correspond aux ponts observés avec un bruit } \eta^+ = 1 - \alpha_1^n \\ g = 0 \text{ correspond aux paires non appariées avec bruit } \eta^- = \alpha_0^n \end{cases}$$



Prouver l'existence d'une information locale

Modèle de bruit de classification induit

- Généralisation du modèle de bruit de classification CN
- Le taux de bruit est conditionnel à chaque classe
- Bruit de classification conditionnel à chaque classe (CCCN)

Prouver l'existence d'une information locale

Modèle de bruit de classification induit

- Généralisation du modèle de bruit de classification CN
- Le taux de bruit est conditionnel à chaque classe
- Bruit de classification conditionnel à chaque classe (CCCN)

Détecter la présence d'une information locale

Si une information locale existe
Si elle est représentable par une fonction apprenable en contexte CCCN
Alors, on doit pouvoir la détecter, l'extraire et l'évaluer

Apprentissage avec bruit CCCN

Littérature sur ce modèle de bruit

- Une seule brève référence à ce modèle dans [Blum et Mitchell, 1998]
- Aucun résultat et aucune connaissance sur ce modèle de bruit

Apprentissage avec bruit CCCN

Littérature sur ce modèle de bruit

- Une seule brève référence à ce modèle dans [Blum et Mitchell, 1998]
- Aucun résultat et aucune connaissance sur ce modèle de bruit

Le double intérêt d'une étude de contexte d'apprentissage

- Elle permettrait l'application du protocole proposé
- Contribution intéressante au domaine de l'apprentissage

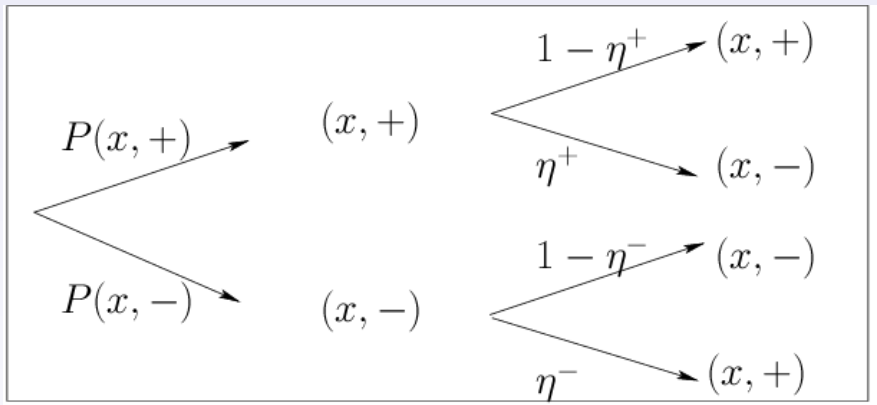
III

Etude de l'apprentissage supervisé en présence de bruit de classification CCCN

Définition

Classification supervisée avec bruit CCCN

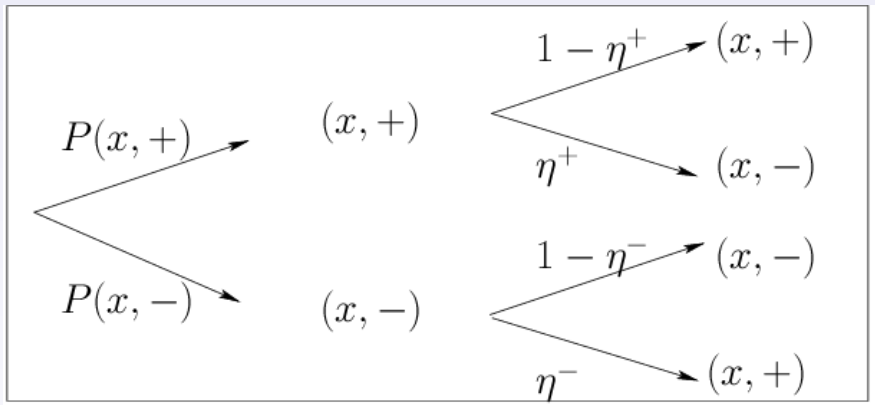
- $S^\eta = \{(x_1, y_1^\eta), ..., (x_l, y_l^\eta)\}$
- Les exemples de S^η sont distribués selon la distribution P^η
- $\eta^+ + \eta^- < 1$ pour lever toute ambiguïté



Définition

Classification supervisée avec bruit CCCN

- $S^\eta = \{(x_1, y_1^\eta), ..., (x_l, y_l^\eta)\}$
- Les exemples de S^η sont distribués selon la distribution P^η
- $\eta^+ + \eta^- < 1$ pour lever toute ambiguïté

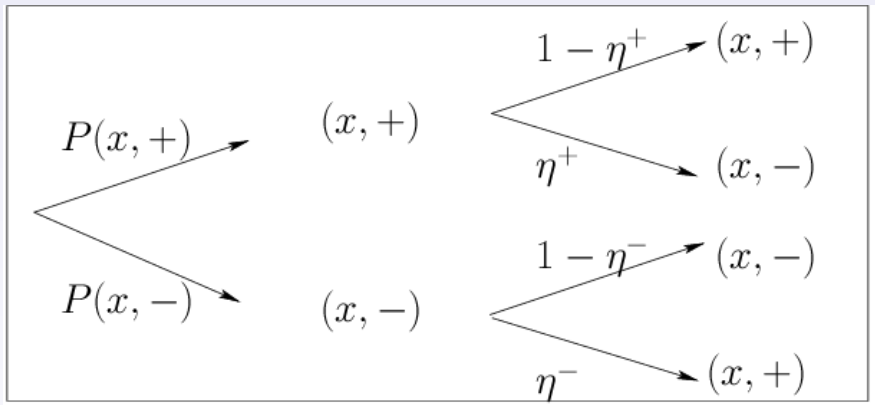


$$\begin{cases} P^\eta(x,+) = (1-\eta^+)\cdot P(x,+) + \eta^- \cdot P(x,-) \\ P^\eta(x,-) = \eta^+ \cdot P(x,+) + (1-\eta^-)\cdot P(x,-) \end{cases}$$

Définition

Classification supervisée avec bruit CCCN

- $S^\eta = \{(x_1, y_1^\eta), ..., (x_l, y_l^\eta)\}$
- Les exemples de S^η sont distribués selon la distribution P^η
- $\eta^+ + \eta^- < 1$ pour lever toute ambiguïté



$$\begin{cases} P^\eta(x,+) = (1-\eta^+)\cdot P(x,+) + \eta^- \cdot P(x,-) \\ P^\eta(x,-) = \eta^+ \cdot P(x,+) + (1-\eta^-)\cdot P(x,-) \end{cases}$$

Objectif

Calculer f qui minimise $R(f)$ relativement à la distribution originale P

Classification supervisée avec bruit CCCN

Etude du problème sans hypothèse sur P

- On cherche à minimiser :

$$R(f) = \frac{R^{\eta}(f) - \eta^{+} \cdot p - \eta^{-} \cdot (1-p)}{1 - \eta^{-} - \eta^{+}}$$

- avec $R^{\eta}(f) = P^{\eta}(f(x) \neq y)$ et $p = P(f(x) = +)$

Classification supervisée avec bruit CCCN

Etude du problème sans hypothèse sur P

- On cherche à minimiser :

$$R(f) = \frac{R^\eta(f) - \eta^+ \cdot p - \eta^- \cdot (1-p)}{1 - \eta^- - \eta^+}$$

- avec $R^\eta(f) = P^\eta(f(x) \neq y)$ et $p = P(f(x) = +)$
- Ne correspond pas forcément à minimiser $R^\eta(f)$

Classification supervisée avec bruit CCCN

Etude du problème sans hypothèse sur P

- On cherche à minimiser :

$$R(f) = \frac{R^\eta(f) - \eta^+ \cdot p - \eta^- \cdot (1-p)}{1 - \eta^- - \eta^+}$$

- avec $R^\eta(f) = P^\eta(f(x) \neq y)$ et $p = P(f(x) = +)$
- Ne correspond pas forcément à minimiser $R^\eta(f)$
- Différence importante avec le modèle CN ($\eta^+ = \eta^- = \eta < 0.5$)

$$R(f) = \frac{R^\eta(f) - \eta}{1 - 2\eta}$$

Classification supervisée avec bruit CCCN

Etude du problème sans hypothèse sur P

- On cherche à minimiser :

$$R(f) = \frac{R^\eta(f) - \eta^+ \cdot p - \eta^- \cdot (1-p)}{1 - \eta^- - \eta^+}$$

- avec $R^\eta(f) = P^\eta(f(x) \neq y)$ et $p = P(f(x) = +)$
- Ne correspond pas forcément à minimiser $R^\eta(f)$
- Différence importante avec le modèle CN ($\eta^+ = \eta^- = \eta < 0.5$)

$$R(f) = \frac{R^\eta(f) - \eta}{1 - 2\eta}$$

- Peut être impossible si les η^+ et η^- sont inconnus

Classification supervisée avec bruit CCCN

Etude du problème sans hypothèse sur P

- On cherche à minimiser :

$$R(f) = \frac{R^\eta(f) - \eta^+ \cdot p - \eta^- \cdot (1-p)}{1 - \eta^- - \eta^+}$$

- avec $R^\eta(f) = P^\eta(f(x) \neq y)$ et $p = P(f(x) = +)$
- Ne correspond pas forcément à minimiser $R^\eta(f)$
- Différence importante avec le modèle CN ($\eta^+ = \eta^- = \eta < 0.5$)

$$R(f) = \frac{R^\eta(f) - \eta}{1 - 2\eta}$$

- Peut être impossible si les η^+ et η^- sont inconnus

Théorème

Le problème est mal posé

[Denis, Magnan et Ralaivola, CAp 2006, ICML 2006]

Une première condition d'identifiabilité de P

Toutefois

- Dans certains cas, la distribution $P^{\eta} \Rightarrow P$
- Dans ces cas, le problème devient bien posé

Une première condition d'identifiabilité de P

Toutefois

- Dans certains cas, la distribution $P^\eta \Rightarrow P$
- Dans ces cas, le problème devient bien posé

Proposition

Soient $\mathcal{P}$ un ensemble de distributions sur $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Q} = \{P(.|y)|y = - \text{ ou } y = +, P \in \mathcal{P}\}$. Si les 2-mélanges de $\mathcal{Q}$ sont identifiables, alors $\mathcal{P}$ est identifiable avec bruit CCCN

Une première condition d'identifiabilité de P

Toutefois

- Dans certains cas, la distribution $P^\eta \Rightarrow P$
- Dans ces cas, le problème devient bien posé

Proposition

Soient $\mathcal{P}$ un ensemble de distributions sur $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Q} = \{P(.|y)|y = - \text{ ou } y = +, P \in \mathcal{P}\}$. Si les 2-mélanges de $\mathcal{Q}$ sont identifiables, alors $\mathcal{P}$ est identifiable avec bruit CCCN

Corollaires

- Mélanges de distributions produits identifiables

Une première condition d'identifiabilité de P

Toutefois

- Dans certains cas, la distribution $P^\eta \Rightarrow P$
- Dans ces cas, le problème devient bien posé

Proposition

Soient $\mathcal{P}$ un ensemble de distributions sur $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Q} = \{P(.|y) | y = - \text{ ou } y = +, P \in \mathcal{P}\}$. Si les 2-mélanges de $\mathcal{Q}$ sont identifiables, alors $\mathcal{P}$ est identifiable avec bruit CCCN

Corollaires

- Mélanges de distributions produits identifiables
- Classifieurs naïfs de Bayes déterminés par P^η

[Denis, Magnan et Ralaivola, CAp 2006, ICML 2006]

Contributions supplémentaires

Mélanges de distributions produits

- Expressions analytiques des coefficients de mélange
- Elles permettent :
  - de retrouver l'expression analytique de $P(+)$ obtenue en contexte semi-supervisé asymétrique
  - d'obtenir une expression analytique des paramètres des classifieurs naïfs de Bayes

Contributions supplémentaires

Mélanges de distributions produits

- Expressions analytiques des coefficients de mélange
- Elles permettent :
  - de retrouver l'expression analytique de $P(+)$ obtenue en contexte semi-supervisé asymétrique
  - d'obtenir une expression analytique des paramètres des classifieurs naïfs de Bayes

NB-CCCN Algorithme Naïf de Bayes en contexte CCCN

- Estimateurs des paramètres déduits des formules précédentes
- **Sortie :** $\hat{\theta}$, une estimation consistante de θ

Contributions supplémentaires

Mélanges de distributions produits

- Expressions analytiques des coefficients de mélange
- Elles permettent :
  - de retrouver l'expression analytique de $P(+)$ obtenue en contexte semi-supervisé asymétrique
  - d'obtenir une expression analytique des paramètres des classifieurs naïfs de Bayes

NB-CCCN Algorithme Naïf de Bayes en contexte CCCN

- Estimateurs des paramètres déduits des formules précédentes
- **Sortie :** $\hat{\theta}$, une estimation consistante de θ

NB-CCCN-EM NB-CCCN + Méthode E.M.

- **Valeurs manquantes :** quelles données sont corrompues ?
- Modèle initial θ_0 obtenu avec `NB-CCCN`
- Formules de calcul des paramètres de θ_n

Le problème d'évaluation des modèles inférés

Comment évaluer les classifieurs calculés ?

- Données d'apprentissage sont corrompues par du bruit CCCN
- Des données test devraient également l'être
- Or, dans ce cas, elles n'attestent pas de la qualité des modèles
- Conséquence directe de l'inconsistance du principe ERM
  $\Rightarrow$ question laissée ouverte par nos travaux

Le problème d'évaluation des modèles inférés

Comment évaluer les classifieurs calculés ?

- Données d'apprentissage sont corrompues par du bruit CCCN
- Des données test devraient également l'être
- Or, dans ce cas, elles n'attestent pas de la qualité des modèles
- Conséquence directe de l'inconsistance du principe ERM
  $\Rightarrow$ question laissée ouverte par nos travaux

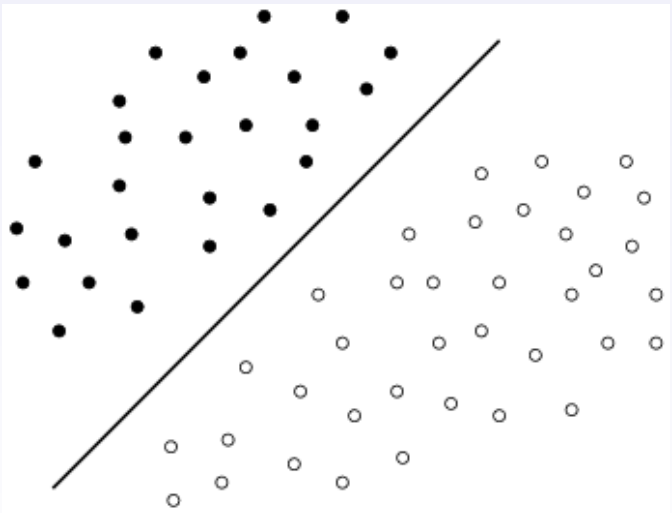
Expériences sur données artificielles et UCI

- Evaluation sur des données non corrompues
- Le bruit CCCN peut efficacement être éliminé
- Les algorithmes permettent d'obtenir de bonnes estimations des modèles naïfs de Bayes

Classe des séparateurs linéaires

Description sommaire

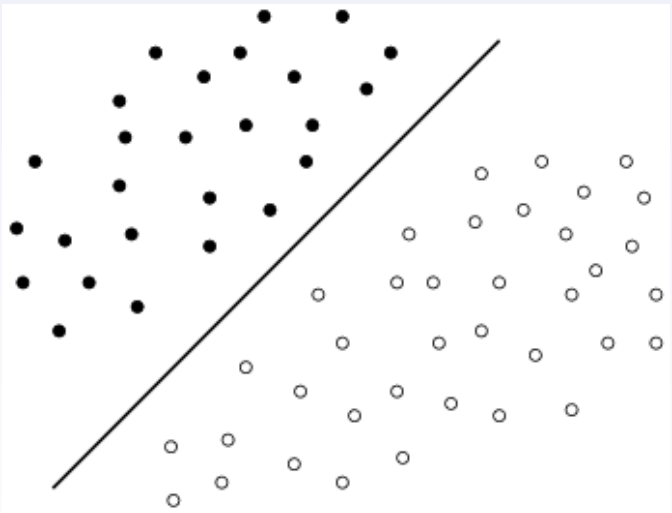
- $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Y} = \{+, -\}$
- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$
- Un hyperplan H sépare S si les données positives et négatives de S se trouvent de part et d'autre de H

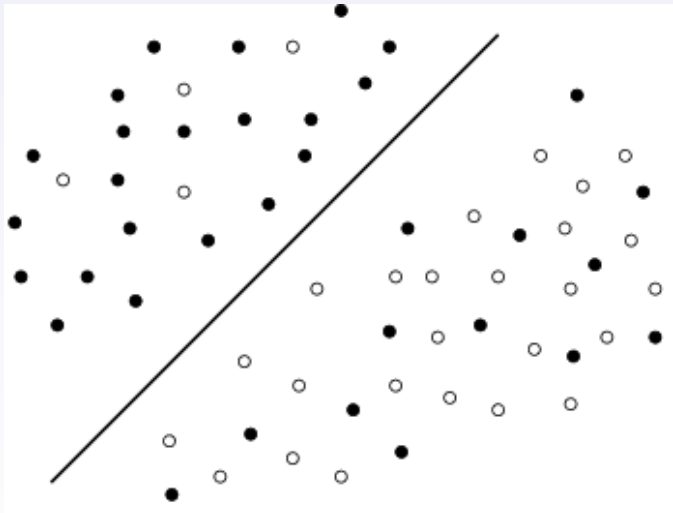


Classe des séparateurs linéaires

Description sommaire

- $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Y} = \{+, -\}$
- $S = \{(x_1, y_1), ..., (x_l, y_l)\}$
- Un hyperplan H sépare S si les données positives et négatives de S se trouvent de part et d'autre de H
- Peut-on apprendre ces séparateurs en présence de bruit CCCN ?





Classe des séparateurs linéaires

Algorithme du perceptron

Entrée : S, séparable par un hyperplan w^*
$w = \overrightarrow{0}$
Tant que $\exists (x, y) \in S$ tel que $w \cdot yx < 0$ **faire**
  $x_{upd} = yx$ tel que $w \cdot yx < 0$
  $w = w + x_{upd}$
Fin Tant que
Sortie : w, qui sépare S

Propriétés

- Exponentiel dans le pire des cas
- Généralement efficace et performant en pratique
- Même face aux séparateurs linéaires optimaux [Graepel et al., 2001]

Résultats obtenus dans le cadre CCCN

Bruits η^+ et η^- connus

- On peut estimer la somme des exemples mal classés par le w courant
- C'est un bon vecteur de mise à jour du perceptron [Blum et al., 1996]

Résultats obtenus dans le cadre CCCN

Bruits η^+ et η^- connus

- On peut estimer la somme des exemples mal classés par le w courant
- C'est un bon vecteur de mise à jour du perceptron [Blum et al., 1996]

Bruits η^+ et η^- inconnus

- Tester différentes valeurs pour η^+ et η^- $(\eta^+ + \eta^- < 1)$
- Apprendre une hypothèse avec chacun de ces taux de bruit
- Comment sélectionner une bonne hypothèse ?

Résultats obtenus dans le cadre CCCN

Bruits η^+ et η^- connus

- On peut estimer la somme des exemples mal classés par le w courant
- C'est un bon vecteur de mise à jour du perceptron [Blum et al., 1996]

Bruits η^+ et η^- inconnus

- Tester différentes valeurs pour η^+ et η^- $(\eta^+ + \eta^- < 1)$
- Apprendre une hypothèse avec chacun de ces taux de bruit
- Comment sélectionner une bonne hypothèse ?

Nous montrons ...

Qu'il existe un critère de sélection consistant pour les séparateurs linéaires

Résultats obtenus dans le cadre CCCN

Bruits η^+ et η^- connus

- On peut estimer la somme des exemples mal classés par le w courant
- C'est un bon vecteur de mise à jour du perceptron [Blum et al., 1996]

Bruits η^+ et η^- inconnus

- Tester différentes valeurs pour η^+ et η^- $(\eta^+ + \eta^- < 1)$
- Apprendre une hypothèse avec chacun de ces taux de bruit
- Comment sélectionner une bonne hypothèse ?

Nous montrons ...

Qu'il existe un critère de sélection consistant pour les séparateurs linéaires

Importance de ce résultat

- On obtient un algorithme du perceptron en contexte CCCN
- Premier cas pour lequel nous avons pu établir un tel critère d'induction

IV

Première expérimentation du protocole de détection d'affinités locales

Expérimentations sur des données biologiques

Rappels

- Environnements locaux impliqués dans la formation de ponts ?
- Nous avons proposé de chercher une fonction d'affinité
- Algorithmes tolérants au bruit CCCN nécessaires pour les apprendre

Expérimentations sur des données biologiques

Rappels

- Environnements locaux impliqués dans la formation de ponts ?
- Nous avons proposé de chercher une fonction d'affinité
- Algorithmes tolérants au bruit CCCN nécessaires pour les apprendre

Algorithme utilisé pour chercher une fonction d'affinité

- L'algorithme du perceptron CCCN

Expérimentations sur des données biologiques

Rappels

- Environnements locaux impliqués dans la formation de ponts ?
- Nous avons proposé de chercher une fonction d'affinité
- Algorithmes tolérants au bruit CCCN nécessaires pour les apprendre

Algorithme utilisé pour chercher une fonction d'affinité

- L'algorithme du perceptron CCCN

Jeu de données Ponts Salins

- Développé par C. Geourjon (IBCP, Lyon)
- 1688 protéines - 7594 ponts salins

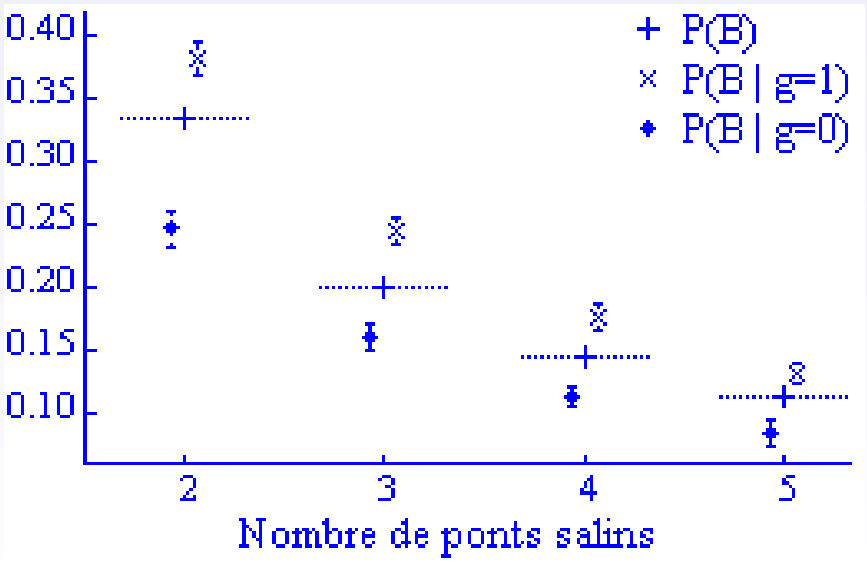
Jeu de données Ponts Disulfures

- Développé par J. Cheng (UCI, Californie)
- 1018 protéines - 2541 ponts disulfures

Résultats sur les données ponts salins

Une information locale clairement détectée

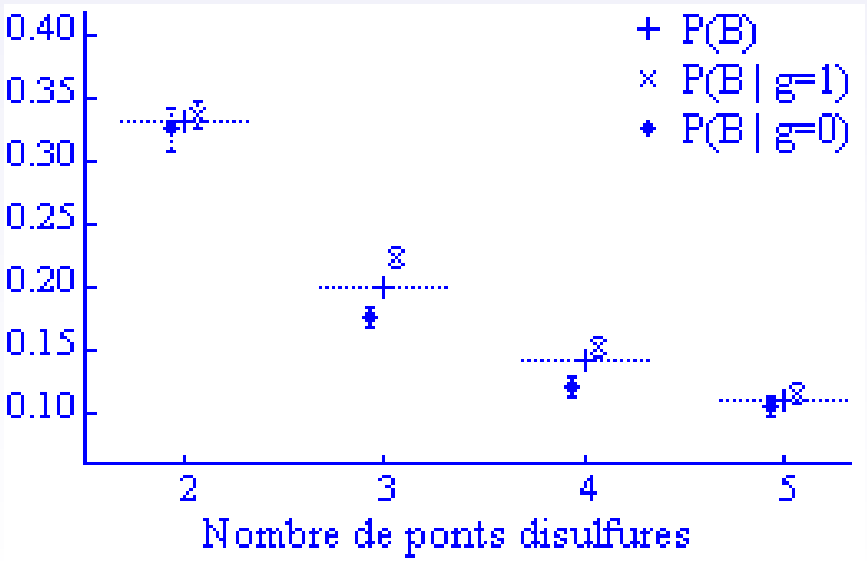
- Les fonctions g apprises montrent qu'un signal clair est détecté
- Ces fonctions sont toujours telles que $P(B|g=1) > P(B|g=0)$
- Il y a une information locale impliquée dans la formation de ces ponts
- Non corrélée à la structure secondaire



Résultats sur les données ponts disulfures

Un signal trop faible pour conclure

- Quel que soit le nombre de ponts, $P(B|g=1) \simeq P(B|g=0)$
- De nombreuses explications possibles :
  - Réalité biologique
  - Classe de séparateurs trop peu expressive
  - Manque de données
  - ....
- Ces expériences ne permettent pas de dire lesquelles sont les + probables



Conclusions

Prédiction des contacts entre acides aminés

- Première étude sur l'existence d'une information locale
- Un protocole de détection de l'affinité locale
- Montre l'existence de cette information pour les ponts salins
- Le cas des ponts disulfures reste ouvert

Conclusions

Prédiction des contacts entre acides aminés

- Première étude sur l'existence d'une information locale
- Un protocole de détection de l'affinité locale
- Montre l'existence de cette information pour les ponts salins
- Le cas des ponts disulfures reste ouvert

Apprentissage automatique

- Etude du contexte semi-supervisé asymétrique
- Introduction de l'apprentissage avec bruit CCCN
- Résultats d'identifiabilité dans ce contexte
- Des algorithmes d'apprentissage efficaces dans ce contexte

Conclusions

Prédiction des contacts entre acides aminés

- Première étude sur l'existence d'une information locale
- Un protocole de détection de l'affinité locale
- Montre l'existence de cette information pour les ponts salins
- Le cas des ponts disulfures reste ouvert

Apprentissage automatique

- Etude du contexte semi-supervisé asymétrique
- Introduction de l'apprentissage avec bruit CCCN
- Résultats d'identifiabilité dans ce contexte
- Des algorithmes d'apprentissage efficaces dans ce contexte

Une plateforme logicielle Java : NoTALAP

- Implémentation du protocole de détection d'affinités locales

Perspectives

Prédiction des contacts entre acides aminés

- Ne pas séparer les protéines par nombre de ponts
- Intégrer la fonction d'affinité g dans la prédiction des contacts

Perspectives

Prédiction des contacts entre acides aminés

- Ne pas séparer les protéines par nombre de ponts
- Intégrer la fonction d'affinité g dans la prédiction des contacts

Apprentissage automatique

- D'autres classes de fonctions apprenables en contexte CCCN
- Méthodes à noyaux ? $\longrightarrow$ noyaux de paires de fenêtres

- Comment montrer que des données sont CCCN ?
- Un critère inductif consistant pour toute méthode ?
- Comment évaluer les classifieurs sur des données bruitées ?

Publications

Conférences internationales

- **[ICML 2006]** François Denis, Christophe N. Magnan, Liva Ralaivola
  "Efficient Learning of NB Classifiers under Class-Conditional Classification Noise"
- **[ICML 2006]** Liva Ralaivola, François Denis, Christophe N. Magnan *"CN=CPCN"*
- **[BIBM 2007]** Christophe N. Magnan, Cécile Capponi, François Denis
  "A Protocol to Detect Local Affinities Involved in Proteins Distant Interactions"

Conférences nationales

- **[CAp 2005]** Christophe N. Magnan *"Apprentissage semi-supervisé asymétrique et estimation d'affinités locales dans les protéines"*
- **[CAp 2006]** François Denis, Christophe N. Magnan, Liva Ralaivola
  "Apprentissage de classifieurs naïfs de Bayes à partir de données soumises à un bruit de classification conditionnel à chaque classe"
- **[CAp 2006]** Liva Ralaivola, François Denis, Christophe N. Magnan
  "Bruits de classification constant et constant par morceaux : égalité des ensembles de classes de concepts apprenables"
- **[CAp 2007]** Christophe N. Magnan, Cécile Capponi, François Denis
  "Un protocole de détection d'affinités locales dans les protéines"

Revue Nationale

- **[Revue R.I.A., Volume 20, Num 6, 11/2006]** Christophe N. Magnan
  "Asymmetrical Semi-Supervised Learning and Prediction of Disulfide Connectivity in Proteins"