Application de l'apprentissage statistique en bioinformatique

Yann Guermeur

LORIA - CNRS

http://www.loria.fr/~guermeur

20 décembre 2007

Plan

Théorie statistique de l'apprentissage

- Origines de la théorie
- Taux de convergence du risque empirique
- Principes inductifs ERM et SRM

Modèles de l'apprentissage numérique

- Modèles connexionnistes de base
- Machines à noyau - SVM
- Modélisation stochastique
- Systèmes hybrides

Applications en bioinformatique

- Biologie structurale prédictive
- Détermination des zones codantes de l'ADN
- Détection des homologies éloignées

Rappels d'analyse - Différents types de convergence

Convergence d'une suite

$(u_n)_{n\in\mathbb{N}^*}$: suite de réels

(u_n) *converge* vers a quand n tend vers l'infini $\Longleftrightarrow$

$$\forall \epsilon \in \mathbb{R}_+^*,\ \exists n_0 \in \mathbb{N}^* \ : \forall n \geq n_0,\ |u_n - a| \leq \epsilon$$

Convergences d'une suite de fonctions

$(f_n)_{n\in\mathbb{N}^*}$: suite de fonctions de E dans $\mathbb{R}$

(f_n) *converge simplement* sur E vers f $\Longleftrightarrow$

$$\forall \epsilon \in \mathbb{R}_+^*,\ \forall x \in E,\ \exists n_0 \in \mathbb{N}^* \ : \forall n \geq n_0,\ |f_n(x) - f(x)| \leq \epsilon$$

(f_n) *converge uniformément* sur E vers f $\Longleftrightarrow$

$$\forall \epsilon \in \mathbb{R}_+^*,\ \exists n_0 \in \mathbb{N}^* \ : \forall n \geq n_0,\ \forall x \in E,\ |f_n(x) - f(x)| \leq \epsilon$$

Rappels de probabilités - Différents types de convergence

$(\Omega, \mathcal{B}, P)$: espace probabilisé

$(X_n)_{n\in\mathbb{N}^*}$: suite de variables aléatoires réelles (v.a.r.) définies sur $(\Omega, \mathcal{B}, P)$

Convergence en probabilité

(X_n) converge vers X *en probabilité* (ou en mesure ou stochastiquement) si

$$\forall \epsilon > 0, \quad P\left\{|X_n - X| \geq \epsilon\right\} \xrightarrow[n\longrightarrow\infty]{} 0 \qquad \left((X_n) \xrightarrow{P} X\right)$$

Convergence presque sûre

(X_n) converge vers X *presque sûrement* (ou en/avec une probabilité 1) s'il existe un ensemble P-négligeable $E \in \mathcal{B}$ /

$$\forall \omega \notin E, \quad X_n(\omega) - X(\omega) \xrightarrow[n\longrightarrow\infty]{} 0 \qquad \left((X_n) \xrightarrow{p.s.} X\right)$$

i.e.

$$P\left(\left\{\omega \in \Omega \quad : \quad \lim_{n\longrightarrow\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

Rappels de probabilités - Différents types de convergence

Convergence en loi

(X_n) converge *en loi* (ou en distribution) vers X si la suite des fonctions de répartition F_{X_n} converge vers la fonction de répartition F_X de X en tout point de continuité de F_X

$$(X_n) \xrightarrow{\mathcal{L}} X \text{ ou } (X_n) \xrightarrow{d} X$$

(X_n) converge en loi vers $X \Longleftrightarrow$

$$\lim_{n \to \infty} \mathbb{E}\left(f(X_n)\right) = \mathbb{E}\left(f(X)\right)$$

pour toute fonction **continue et bornée** f de $\mathbb{R}$ dans $\mathbb{R}$

Convergence presque sûre $\Longrightarrow$ convergence en probabilité $\Longrightarrow$ convergence en loi

$$\left((X_n) \xrightarrow{p.s.} X\right) : \quad \forall \epsilon > 0, \quad \lim_{n \longrightarrow \infty} P\left\{\sup_{m \geq n} |X_m - X| \geq \epsilon\right\} = 0 \Longrightarrow$$

$$\lim_{n \longrightarrow \infty} P\left\{|X_n - X| \geq \epsilon\right\} = 0 \quad : \left((X_n) \xrightarrow{P} X\right)$$

Rappels de probabilités - Comportement asymptotique des suites de v.a.r.

Loi faible des grands nombres (Bernoulli, 1680 $\rightarrow$)

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi qu'une variable X vérifiant $\mathbb{E}\left(|X|\right) < \infty$ et $\mathbb{E}(X) = \mu$. Alors

$$\forall \epsilon > 0, \quad P\left\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right\} \xrightarrow[n\longrightarrow\infty]{} 0$$

i.e., en notant $\bar{X}_n = \frac{X_1+X_2+\ldots+X_n}{n}$,

$$\bar{X}_n \xrightarrow{P} \mu$$

Loi forte des grands nombres (Laplace, Poisson, XIX-ième $\rightarrow$)

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi qu'une variable X vérifiant $\mathbb{E}\left(|X|\right) < \infty$ et $\mathbb{E}(X) = \mu$. Alors **avec une probabilité 1**

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \rightarrow \mu$$

i.e.

$$\bar{X}_n \xrightarrow{p.s.} \mu$$

Les hypothèses des deux lois sont les mêmes !

Rappels de probabilités - Comportement asymptotique des suites de v.a.r.

Différence entre loi faible et loi forte des grands nombres

La **loi faible** assure que pour n suffisamment grand, $\frac{X_1+X_2+\ldots+X_n}{n}$ diffère peu de μ avec une forte probabilité. Elle n'exclut pas la situation où de larges écarts peuvent se produire pour une infinité d'événements, pourvu que leur probabilité totale soit suffisamment faible. La **loi forte** exclut cette situation.

Elle assure en particulier qu' **avec une probabilité 1** et pour toute valeur de ϵ,

$$\left|\bar{X}_n - \mu\right|$$

ne sera supérieure à ϵ qu'un nombre fini de fois.

Lemme de Borel-Cantelli

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a.r. définie sur un espace probabilisé $(\Omega, \mathcal{B}, P)$.
(1) Si pour tout $\epsilon > 0$, $\sum_{n\in\mathbb{N}^*} P\left\{|X_n - X| \geq \epsilon\right\} < \infty$, alors $\left((X_n) \xrightarrow{p.s.} X\right)$
(2) Si les v.a.r. X_n sont *mutuellement indépendantes*, alors $\left((X_n) \xrightarrow{p.s.} 0\right)$ s.s.i. $\sum_{n\in\mathbb{N}^*} P\left\{|X_n| \geq \epsilon\right\} < \infty$ pour tout $\epsilon > 0$.

Rappels de probabilités - Vitesse de convergence - Inégalités de concentration

Inégalité de Bienaymé-Tchebychev

Soit X une v.a.r. d'espérance μ et de variance σ^2 finies. Alors

$$\forall \epsilon > 0, \;\; P\left\{|X - \mu| \geq \epsilon\right\} \leq \frac{\sigma^2}{\epsilon^2}$$

Borne de Hoeffding (1963)

Soit $(X_i)_{1 \leq i \leq n}$ une suite de n variables aléatoires indépendantes, d'espérance $\mu = 0$ et prenant respectivement leurs valeurs dans les intervalles $[a_i, b_i]$.

$$\forall \epsilon > 0, \;\; P\left\{X_1 + \ldots + X_i + \ldots + X_n \geq \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Rappels de probabilités - Convergence uniforme des mesures empiriques

Lois fortes des grands nombres uniformes

Théorème de Glivenko-Cantelli (1933)

Soit X une variable aléatoire réelle sur $(\Omega, \mathcal{B}, P)$, de fonction de répartition F_X. Soit $(X_i)_{1 \leq i \leq n}$ une suite de n variables aléatoires indépendantes et de même loi que X (n-échantillon). Notons

$$F_{X,n}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$$

$F_{X,n}$ est la fonction de répartition *empirique* de l'échantillon.

La convergence de $(F_{X,n})_{n \in \mathbb{N}^*}$ vers F_X est **presque sûrement uniforme**, i.e.

$$\sup_x |F_{X,n}(x) - F_X(x)| \xrightarrow{p.s.} 0$$

(la convergence presque sûre résulte de l'application du lemme de Borel-Cantelli)

Apprentissage statistique - hypothèses fondamentales

- **Caractérisation du problème**

1. Etude du lien associant des objets $x \in \mathcal{X}$ à leurs étiquettes $y \in \mathcal{Y} = \{1, \ldots, Q\}$
2. Hypothèse : existence d'un couple aléatoire (X, Y) à valeurs dans $\mathcal{X} \times \mathcal{Y}$, distribué suivant une mesure de probabilité P
3. Problème : la loi jointe de (X, Y) est inconnue

- **Ce dont on dispose**

1. $s_m = \{(X_i, Y_i) : 1 \leq i \leq m\}$: m-échantillon constitué de copies de (X, Y) indépendantes
2. $\mathcal{F}$, famille de fonctions f, de $\mathcal{X}$ dans $\mathcal{Y}$, ou $\mathcal{G}$, famille de fonctions g, de $\mathcal{X}$ dans $\mathbb{R}^Q$

- **Ce que l'on cherche à faire**

1. ℓ, fonction de perte : $\ell\left(y, f(x)\right) = \mathbb{1}_{\{f(x) \neq y\}}$ ou $\ell\left(y, g(x)\right) = \mathbb{1}_{\{g_y(x) \leq \max_{k \neq y} g_k(x)\}}$
2. Risque : espérance de la fonction de perte
3. Sélection d'une fonction f^* ou g^* minimisant sur sa famille le risque

$$R(f) = \mathbb{E}\mathbb{1}_{\{f(X) \neq Y\}} = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\{f(x) \neq y\}} dP\left(x, y\right) = P(f(X) \neq Y)$$

Apprentissage statistique - Principe inductif ERM

Apprentissage à partir des données

Problème : la fonctionnelle du risque dépend de P qui est inconnue

$\Longrightarrow$ elle ne peut être minimisée directement

Stratégie : pour optimiser les performances, on essaie d'obtenir l'information pertinente à partir de ce qui est disponible, i.e. on s'appuie sur l'ensemble d'observations s_m

$\Longrightarrow$ on se place dans le cadre de la *statistique inférentielle*

Statistique inférentielle

• Objectif

Disposant d'observations sur un échantillon de taille m, on désire en déduire les propriétés de la population dont il est issu

• Conditions

L'échantillon doit avoir été tiré suivant des règles rigoureuses afin d'être “représentatif”

En pratique, on fait l'hypothèse que les observations disponibles sont issues d'un *échantillonnage aléatoire simple* (tirages équiprobables et indépendants les uns des autres)

Apprentissage statistique - Principe inductif ERM

Utilisation de la statistique inférentielle pour tenter de minimiser le risque

Le *principe inductif de minimisation empirique du risque* (ERM) consiste à substituer au risque son estimation empirique sur s_m

$$R_m(h) = \frac{1}{m} \sum_{i=1}^{m} \ell\left(Y_i, h\left(X_i\right)\right)$$

Problème de consistance du principe inductif ERM

Soit h_m^* une fonction de $\mathcal{H}$ minimisant le risque empirique sur s_m. Le principe inductif ERM est dit *consistant* pour $\mathcal{H}$ et P si les deux convergences suivantes ont lieu

$$\begin{cases} R(h_m^*) \xrightarrow[m \longrightarrow \infty]{P} \inf_{h \in \mathcal{H}} R(h) \\ R_m(h_m^*) \xrightarrow[m \longrightarrow \infty]{P} \inf_{h \in \mathcal{H}} R(h) \end{cases}$$

On parle alors de *convergence forte* du processus d'apprentissage

Remarque : pour parler de consistance, certains auteurs imposent que les deux suites de risques convergent vers le risque minimal, correspondant à la règle de décision de Bayes

Apprentissage statistique - Principe inductif ERM

Consistance stricte de la méthode ERM

On dit que la méthode de minimisation empirique du risque est *strictement* consistance, ou consistante de manière *non triviale*, si la condition suivante est vérifiée

Pour tout c réel, en posant $\mathcal{H}_c = \{h \in \mathcal{H} / R(h) \geq c\}$, et en désignant par $h^*_{m,c}$ une fonction minimisant le risque empirique sur $\mathcal{H}_c$, on a :

$$R_m(h^*_{m,c}) \xrightarrow[m \longrightarrow \infty]{P} \inf_{h \in \mathcal{H}_c} R(h)$$

Lemme

Si la méthode ERM est strictement consistante, alors on dispose également du second résultat de convergence en probabilité, i.e.

$$R(h^*_m) \xrightarrow[m \longrightarrow \infty]{P} \inf_{h \in \mathcal{H}} R(h)$$

CNS de consistance du principe inductif ERM
Cas général

Théorème fondamental de la théorie de l'apprentissage

(Vapnik et Chervonenkis, 1991)

Supposons que, pour toute probabilité P, il existe deux constantes a et A satisfaisant : $\forall h \in \mathcal{H}$, $a \leq R(h) \leq A$. Alors la condition de consistance stricte est équivalente à la convergence en probabilité suivante

$$\forall \epsilon > 0, \quad P^m \left\{ \sup_{h \in \mathcal{H}} \left(R(h) - R_m(h) \right) \geq \epsilon \right\} \xrightarrow[m \longrightarrow \infty]{} 0$$

Cette condition est très forte, car il s'agit d'une convergence uniforme, ou convergence dans le pire des cas, sur $\mathcal{H}$!!!

CNS de consistance du principe inductif ERM
Cas du calcul des dichotomies

Fonction de croissance

Soit $\mathcal{H}|_{\mathcal{X}_n}$ la restriction de $\mathcal{H}$ à un ensemble $\mathcal{X}_n$ de n points de $\mathcal{X}$ et $\Pi_{\mathcal{H}}(\mathcal{X}_n)$ son cardinal.

La fonction de croissance de $\mathcal{H}$ est la fonction de $\mathbb{N}^*$ dans $\mathbb{N}$ définie par :

$$\forall n \in \mathbb{N}^*, \ \Pi_{\mathcal{H}}(n) = \max_{\mathcal{X}_n \in \mathcal{X}^n} \Pi_{\mathcal{H}}(\mathcal{X}_n)$$

Par définition, $\Pi_{\mathcal{H}}(n) \leq 2^n$. On dit que $\mathcal{X}_n$ est *pulvérisé* par $\mathcal{H}$ si $\Pi_{\mathcal{H}}(\mathcal{X}_n) = 2^n$.

Dimension de Vapnik-Chervonenkis (Vapnik et Chervonenkis, 1971)

Pour un espace de description $\mathcal{X}$ donné, la dimension VC de $\mathcal{H}$ est le plus grand entier d vérifiant $\Pi_{\mathcal{H}}(d) = 2^d$, si un tel entier existe, l'infini sinon. Si la dimension VC est finie, elle correspond donc à la taille du plus grand ensemble pulvérisé.

Extensions dédiées au cas multi-classe et/ou à la régression

Dimension de Natarajan, dimension graphique, dimension "fat-shattering"...

CNS de consistance du principe inductif ERM
Cas du calcul des dichotomies

Lemme de Sauer-Shelah (1972)

Si $\mathcal{H}$ possède une dimension VC finie d, alors, pour n supérieur à d, sa fonction de croissance est bornée de la manière suivante :

$$\Pi_{\mathcal{H}}(n) < \left(\frac{en}{d}\right)^d$$

i.e. la fonction de croissance croît **exponentiellement** avec n, pour $n \leq d$, après quoi elle croît **polynomialement**.

CNS de consistance (Vapnik et Chervonenkis, 1971 → 1991)

Pour le calcul des dichotomies, une **condition nécessaire et suffisante** de consistance stricte de la méthode ERM est le fait que $\mathcal{H}$ possède une **dimension VC finie**

Taux de convergence du risque empirique

Borne sur le risque pour des fonctions de coût indicatrices

En reprenant le schéma de la preuve du théorème de Glivenko-Cantelli **(Pollard, 1984)**, on établit des bornes du type :

$$P^m \left\{ \sup_{h \in \mathcal{H}} \frac{R(h) - R_m(h)}{\sqrt{R(h)}} \geq \epsilon \right\} < 4\Pi_{\mathcal{H}}(2m) \exp\left(-\frac{1}{4}\epsilon^2 m\right)$$

Discussion

(1) compte tenu du lemme de Sauer, il apparaît de manière immédiate qu'une valeur finie de la VC dimension est une condition suffisante de convergence ;

(2) il existe des bornes fondées sur d'autres mesures de capacité que la dimension VC ou la fonction de croissance, comme les *nombres de couverture*.

Réseaux de neurones - les origines

Neurone formel (Mc Culloch et Pitts, 1943)

Mc Culloch étudie la strychnisation du cerveau des chimpanzés. Il se convaint que les circuits d'influx nerveux sont analogues à la propagation logique d'un raisonnement.

Avec l'aide de Pitts, il va présenter le *neurone formel*.

- il est binaire : il est soit actif soit inactif ;
- chaque neurone possède un *seuil de déclenchement* fixé ;
- le neurone reçoit des entrées de *synapses excitatrices*, ayant des poids identiques ;
- le neurone peut aussi recevoir des entrées de *synapses inhibitrices*, dont l'action est absolue : si la synapse est active, le neurone ne peut pas être actif ;
- un quantum de temps est nécessaire à l'intégration des entrées synaptiques, basé sur le délai observé physiologiquement.

Fonctionnement

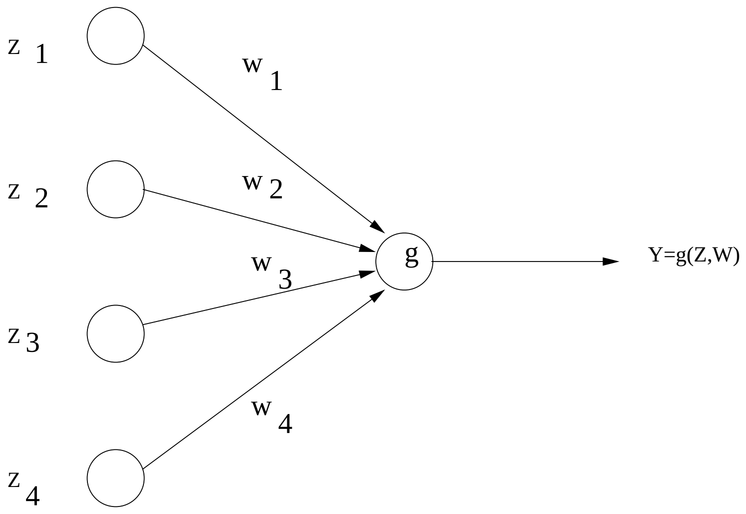
Si aucune synapse inhibitrice n'est active, le neurone ajoute ses entrées synaptiques et vérifie que la somme atteint ou excède son seuil. On dit alors que le neurone devient *actif*, sinon le neurone est *inactif*.

De l'inspiration biologique au modèle de calcul

Unités constitutives des réseaux

Automates caractérisés par :

- un état interne S
- des signaux d'entrée z_i, $(1 \leq i \leq p)$ et un *seuil* ou *biais* θ
- une *fonction de transfert* ou *fonction d'activation* g



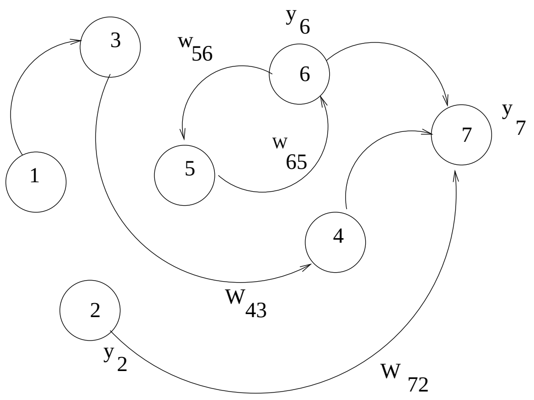
Fonction d'activation usuelle : *sigmoïde*

$$y = \left(1 + \exp\left(-(w^T z - \theta)\right)\right)^{-1}$$

Réseaux de neurones

Définition

Un réseau de neurones est un graphe orienté constitué de neurones interconnectés.



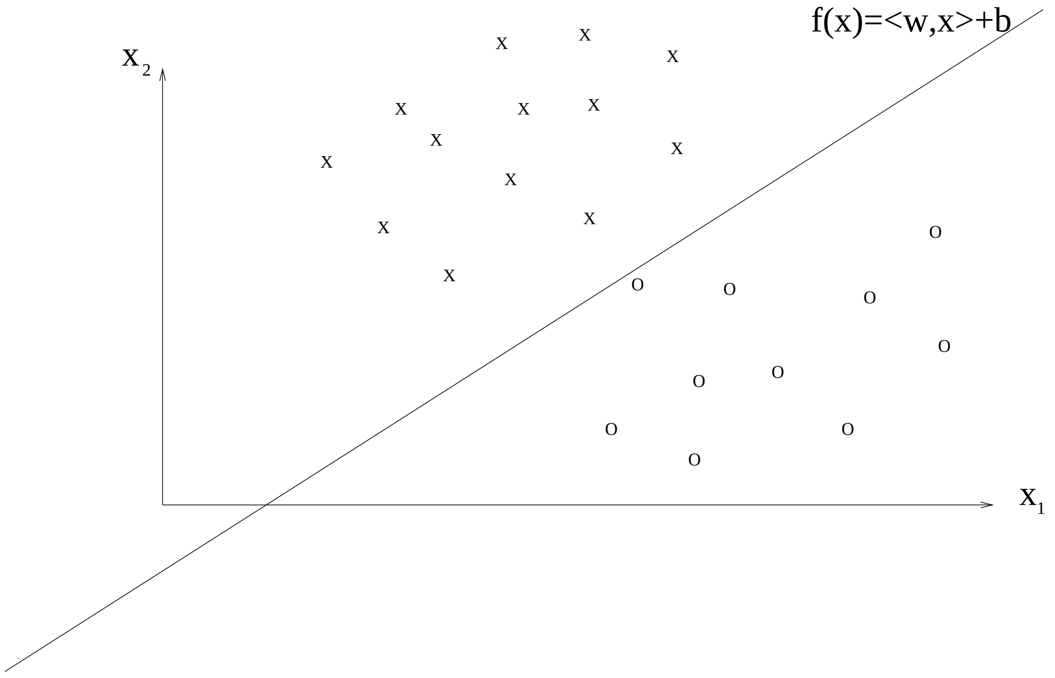
Paramètres du réseau

- *architecture* (topologie, nature des fonctions d'activation...) : définit la classe $\mathcal{H}$ des fonctions calculables par le réseau
- valeurs des poids et des biais : caractérisent une fonction h particulière
- *hyperparamètres* : incorporés en particulier dans la fonction objectif

Le perceptron - Définition

(Rosenblatt, 1956 → 1960)

- Modèle de discrimination linéaire (la frontière de décision est un *hyperplan*).
- Il se distingue du modèle de l'*analyse discriminante de Fisher* (1936) par son algorithme d'apprentissage, qui est *itératif*.



Le perceptron - Algorithme d'apprentissage

Etant donnés un ensemble d'apprentissage linéairement séparable
$s_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \{-1, 1\})^n$ et un *pas d'apprentissage* $\eta \in \mathbb{R}_+$

`Initialisation`

$w_0 \leftarrow 0\,;\ b_0 \leftarrow 0\,;\ k \leftarrow 0\,;\ R \leftarrow \max_{1\leq i\leq n} \|x_i\|$

`Répéter`

    `Pour` i `allant de` 1 `à` n

        `Si` $y_i(w_k^T x_i + b_k) \leq 0$ `alors`

            $w_{k+1} \leftarrow w_k + \eta y_i x_i$

            $b_{k+1} \leftarrow b_k + \eta y_i R^2$

            $k \leftarrow k + 1$

        `Fin si`

    `Fin pour`

`Jusqu'à ce qu'aucune erreur ne soit observée dans la boucle pour`

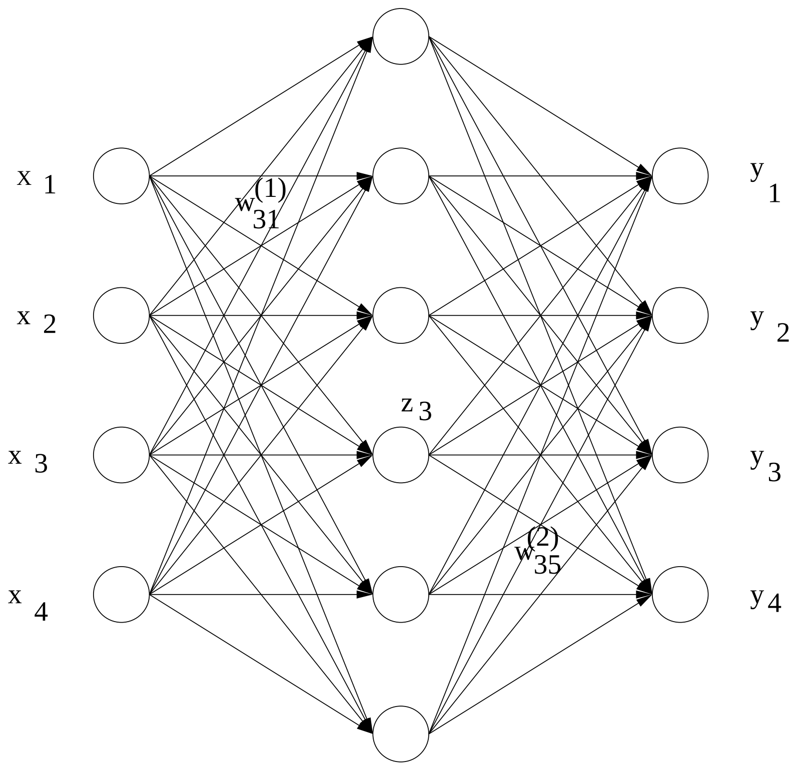
Résultats de convergence (Novikoff, 1960 $\rightarrow$)

Perceptron multi-couche (PMC) - Définition

Réseau à *propagation avant* (sans cycle), dont les unités sont organisées en *couches*.

On distingue les *couches externes* (couches d'entrée et de sortie) et les *couches cachées*.

Les entrées sont présentées sur la première couche et propagées de couche en couche.



Algorithme d'apprentissage du PMC (Rumelhart et *al.*, 1986)

Mise en œuvre du principe inductif ERM

Problème : en discrimination, il est difficile de minimiser directement le taux d'erreur empirique (fonction discontinue . . .)

$\Longrightarrow$ utilisation d'un critère asymptotiquement équivalent, tel que le problème d'optimisation correspondant puisse être résolu (approximativement) en mettant en œuvre un algorithme de programmation mathématique

– application d'un principe fondamental de l'inférence : approximation de l'estimateur du maximum de vraisemblance, de l'estimateur des moindres carrés...
– fonction objectif : coût quadratique, entropie croisée...
– apprentissage en ligne et apprentissage hors ligne
– utilisation de stabilisateurs : régularisation, apprentissage bayésien...

Algorithme de base : rétro-propagation du gradient

Gradient total et gradient stochastique

Mise en œuvre de la rétro-propagation du gradient

Principe (gradient stochastique)

• Recherche d'un *minimum local* de la fonction objectif J par descente de gradient

• A un instant donné de l'apprentissage, la fonction objectif est une fonctionnelle de W, ensemble des paramètres (poids) $w_{jk}^{(l)}$ du réseau et de l'exemple présenté (x_i, y_i)

• La mise à jour du poids $w_{jk}^{(l)}$ s'effectue suivant la formule :

$$\Delta w_{jk}^{(l)} = -\epsilon \frac{\partial J\left(W, (x_i, y_i)\right)}{\partial w_{jk}^{(l)}}$$

où ϵ est le *pas de gradient*

Illustration de l'algorithme de rétro-propagation du gradient

Architecture du réseau

- N_e unités sur la couche d'entrée, N_c unités sur la couche cachée, N_s unités sur la couche de sortie
- Couches successives entièrement interconnectées
- Fonction d'activation des unités de la couche cachée : sigmoïde

$$y_k^{(1)} = \left(1 + \exp\left(-\sum_j w_{kj}^{(1)} x_j^{(0)}\right)\right)^{-1} = \left(1 + \exp\left(-x_k^{(1)}\right)\right)^{-1}$$

- Fonction d'activation des unités de la couche de sortie : exponentielle normalisée

$$y_l^{(2)} = \frac{\exp(\sum_k w_{lk}^{(2)} y_k^{(1)})}{\sum_m \exp(\sum_k w_{mk}^{(2)} y_k^{(1)})} = \frac{\exp(x_l^{(2)})}{\sum_m \exp(x_m^{(2)})}$$

Fonction objectif : entropie croisée

$$J(W,(x,y)) = -\sum_l y_l \ln(y_l^{(2)})$$

PMC et théorie de l'apprentissage - Discussion

Les PMC, un défi au bon sens (Bottou, 1997) ? !

• Dispositif statistiquement **grossièrement sur-paramétré**

• L'optimisation d'une fonction objectif si complexe au moyen d'une procédure de descente de gradient est **numériquement une utopie**

De l'utilité d'un apprentissage "inefficace"

L'inefficacité de la descente en gradient crée une **structure implicite** sur l'ensemble des fonctions réalisables par le réseau

• Initialisation des poids avec des valeurs faibles : au début de l'apprentissage, le modèle est **quasi-linéaire**

• Faible vitesse de convergence de la descente : les non-linéarités interviennent de manière **progressive**

• Arrêt de la procédure avant convergence : *early stopping*

$\Longrightarrow$ principe inductif SRM

Principe inductif SRM

Origines (Vapnik et Chervonenkis, 1974)

La mise en œuvre du principe inductif ERM n'est pas satisfaisante dans le cas des petits échantillons ($m \leq 20\ d$), i.e. lorsque la *confiance VC* n'est pas adaptée

$\Longrightarrow$ Vapnik a introduit le principe de minimisation structurelle du risque (SRM) pour pallier cette insuffisance

Formulation du principe

Recherche du compromis optimal entre la taille de l'ensemble d'apprentissage, la performance sur cet ensemble d'une fonction choisie dans une famille donnée et la valeur qui caractérise la capacité de cette famille

Mise en œuvre

• Minimisation de la borne sur le risque (*risque garanti*)

• Procédure d'optimisation : inverse de celle du principe inductif ERM. On minimise l'intervalle de confiance à risque empirique constant

$\Longrightarrow$ la procédure s'appuie sur une *structure* définie *a priori*

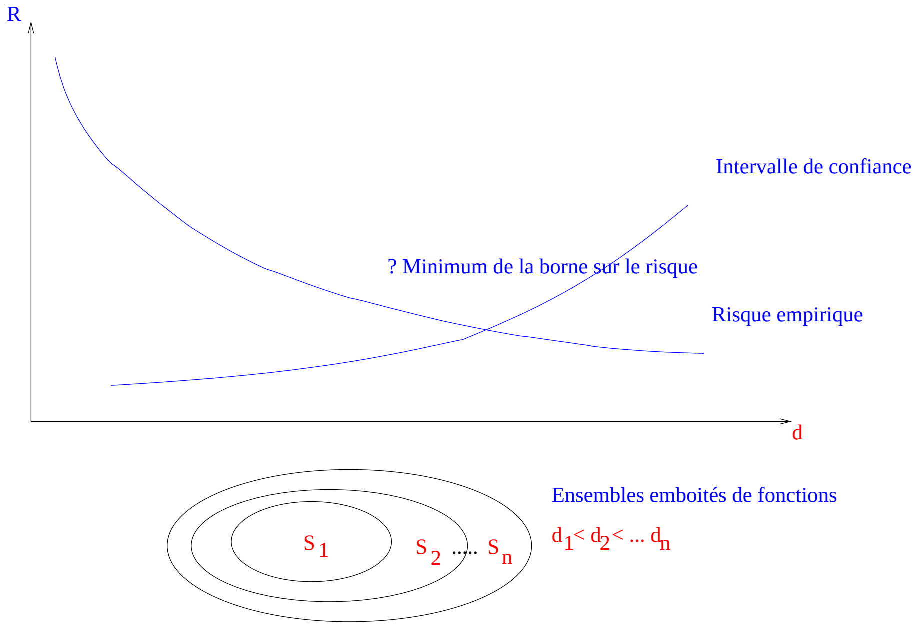
Principe inductif SRM

1) Définition d'une structure

- P étant donnée, on considère la fonction de coût ℓ comme un opérateur sur $\mathcal{H}$
- On définit sur S, image de $\mathcal{H}$ par ℓ, une structure vérifiant

$$S_1 \subset S_2 \subset \ldots \subset S_n \subset \ldots \subset S \qquad d_1 < d_2 < \ldots < d_n < \ldots < d$$

2) Recherche du minimum du risque garanti sur la structure



Consistance du principe inductif SRM

Notations

Soit d_k la dimension VC de S_k et B_k un réel vérifiant, pour toute fonction $h \in \mathcal{H}$ telle que $\ell(., h(.)) \in S_k$ et tout couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$0 \leq \ell\,(y, h(x)) \leq B_k$$

Théorème (Lugosi et Zeger, 1995-1996)

Si la structure (coefficients B_k) vérifie certaines conditions simples alors pour toute distribution sur les données, la mise en œuvre du principe inductif SRM permet une convergence **presque sûre** vers la meilleure solution possible $\Longrightarrow$ la méthode SRM est **universellement fortement consistante.**

Taux de convergence $(\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_k \subset \ldots \subset \mathcal{H})$:

$$\mathbb{E}\left(R_m(h^*_{m,k})\right) - \inf_{h \in \mathcal{H}_k} R(h) \leq c\sqrt{\frac{\text{VC-dim}(\mathcal{H}_k)\ln(m)}{m}}$$

Machines à noyau

Noyau défini positif

Un *noyau défini positif* (n.d.p.) sur l'ensemble $\mathcal{X}$ est une fonction $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ symétrique satisfaisant

$$\forall n \in \mathbb{N}^*, \ \ \forall \left(x_i\right)_{1 \leq i \leq n} \in \mathcal{X}^n, \ \ \forall \left(a_i\right)_{1 \leq i \leq n} \in \mathbb{R}^n, \ \ \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa(x_i, x_j) \geq 0.$$

Machine à noyau

Une machine à noyau est une machine utilisant un n.d.p. comme *mesure de similarité* entre exemples. La classe de fonctions qu'elle calcule appartient à un espace de Hilbert à noyau reproduisant.

Espace de Hilbert à noyau reproduisant

Soit $\mathcal{X}$ un espace quelconque et $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ un espace de Hilbert de fonctions sur $\mathcal{X}$ ($\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$).

Noyau reproduisant (Aronszajn, 1950)

Une fonction $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ est appelée un *noyau reproduisant* (noté n.r.) si et seulement si :

(1) $\mathcal{H}$ contient toutes les fonctions de la forme

$$\forall x \in \mathcal{X}, \quad \kappa_x : t \mapsto \kappa(x, t)$$

(2) $\forall x \in \mathcal{X}, \;\; \forall h \in \mathcal{H}, \;\; h(x) = \langle h, \kappa_x \rangle_{\mathcal{H}}$

Espace de Hilbert à noyau reproduisant

Si un n.r. existe, $\mathcal{H}$ est appelé un *espace de Hilbert à noyau reproduisant* (e.h.n.r. ou RKHS).

Résultats de base sur les RKHS

1. Si un noyau reproduisant existe, il est unique
2. Un noyau reproduisant existe si et seulement si $\forall x \in \mathcal{X}$, la fonctionnelle $h \mapsto h(x)$ (de $\mathcal{H}$ dans $\mathbb{R}$) est continue
3. un noyau reproduisant est un noyau défini positif
4. Si κ est un noyau défini positif, alors il existe un RKHS ayant κ pour noyau reproduisant
5. Si κ est un noyau reproduisant, il vérifie la *propriété reproduisante* :

$$\forall\,(x, x') \in \mathcal{X}^2, \;\; \langle \kappa_x, \kappa_{x'} \rangle_{\mathcal{H}} = \kappa(x, x')$$

RKHS et espace de représentation

De κ vers Φ

Soit κ un noyau défini positif sur un espace $\mathcal{X}$ quelconque. D'après le point (4) du transparent précédent, on peut lui associer un RKHS $\mathcal{H}$. Soit Φ la fonction de $\mathcal{X}$ dans $\mathcal{H}$ définie par $\Phi(x) = \kappa_x$. Par application du point (5) on a :

$$\forall (x, x') \in \mathcal{X}^2, \quad \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

Remarque

Un RKHS possède de nombreuses autres propriétés utiles pour développer des algorithmes d'apprentissage puissants.

Machines à vecteurs support (SVM)

(Boser, Guyon et Vapnik, 1992, Cortes et Vapnik, 1995)

Définition

Système discriminant (à l'origine), développé en application du principe inductif SRM, autour de l'*hyperplan de marge maximale* **(Vapnik, 1982)**

Principe

Deux étapes

1. les données sont projetées dans un espace de Hilbert de grande dimension (voire de dimension infinie), l'*espace de représentation*, au moyen d'une transformation Φ ;
2. dans cet espace, la discrimination est effectuée par l'hyperplan de marge maximale.

Hyperplan optimal

On se place dans le cas du calcul d'une **dichotomie**, sous l'hypothèse que les données sont **linéairement séparables**

Borne sur le risque des modèles affines

$$R(h) \leq R_n^\gamma(h) + \sqrt{\frac{c}{n}\left(\frac{R^2\Lambda^2}{\gamma^2}\ln(n)^2 + \ln\left(\frac{1}{\delta}\right)\right)}$$

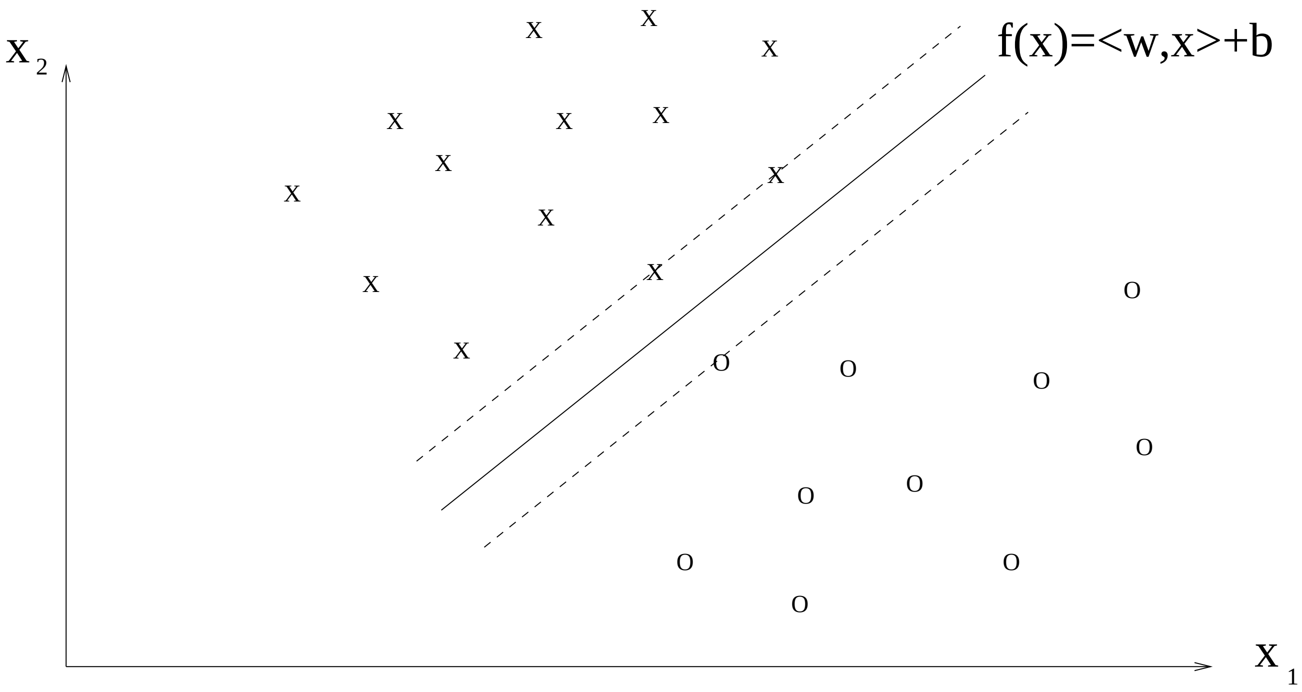
Hyperplans canoniques

Le couple (w, b) caractéristique d'un hyperplan est défini à une constante multiplicative près. Afin d'introduire une normalisation, on considère, pour un ensemble de points s_n donné, la représentation canonique (w_0, b_0) définie de la manière suivante :

$$\min_{(x,y)\in s_n} |w_0^T x + b_0| = 1$$

Hyperplan optimal

Hyperplan *canonique*, de norme minimale, réalisant la dichotomie $\Longrightarrow$ **sa distance aux points d’apprentissage est maximale**



Détermination de l'hyperplan de marge maximale

Problème de programmation quadratique (Fletcher, 1987)

Minimiser

$$J(w) = \|w\|_2^2$$

sous les contraintes :

$$\forall i \in \{1, \ldots, n\}, \quad y_i(w^T x_i + b) \geq 1$$

Expression du *Lagrangien* correspondant :

$$L(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(w^T x_i + b) - 1 \right)$$

Solutions

- Caractérisées par les conditions de Kuhn-Tucker
- Pour des raisons à la fois théoriques et pratiques, on ne résoud pas le problème primal mais son dual de Wolfe

Caractérisation de l'hyperplan de marge maximale

Expression de la fonction calculée

$$w^T x + b = \sum_{i=1}^{n} y_i \alpha_i x_i^T x + b$$

• L'hyperplan s'exprime comme un développement sur les données de la base associées à un **multiplicateur de Lagrange non nul**

Ces données sont appelées *vecteurs support*

Interprétation géométrique des vecteurs support

Vecteurs déterminant la frontière de discrimination

Propriétés statistiques de l'hyperplan de marge maximale

Le nombre de vecteurs support fournit une borne sur le risque (erreur en généralisation)

Ensembles non séparables et SVM non-linéaires

Cas des ensembles non séparables ("soft margin")

Introduction de *variables d'écart* dans les contraintes :

$$\forall i \in \{1, \ldots, m\}, \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Fonction objectif modifiée en conséquence :

$$J(w, b) = \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \xi_i$$

SVM non-linéaires

Des SVM non-linéaires peuvent être construites en utilisant différents types de noyaux (définis positifs).

- Exemples de noyaux définis positifs :

$$\kappa(x, x') = \left(x^T x' + 1\right)^d \qquad \kappa_\sigma(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

SVM non-linéaires et "kernel trick"

Comment travailler dans un espace de représentation de dimension infinie

1. La mise en œuvre d'une SVM ne fait intervenir les données qu'au travers des produits scalaires $x^T x'$ ou $\langle \Phi(x), \Phi(x') \rangle$ (dans l'espace de représentation)
2. $\Longrightarrow$ en choisissant un noyau défini positif κ, $\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$ permet d'opérer dans un espace de représentation (pas nécessairement unique) de manière implicite.

Expression analytique de la solution

– En apprentissage, toute l'information sur s_n est contenue dans le Hessien $H = (h_{ij})_{1 \leq i,j \leq m}$ de la fonction objectif du problème dual défini par :

$$\forall (i, j), \;\; 1 \leq i \leq j \leq m, \;\; h_{ij} = y_i y_j \kappa(x_i, x_j)$$

– En test, l'expression de la fonction calculée est donnée par :

$$\langle w, \Phi(x) \rangle + b = \sum_{i=1}^{m} y_i \alpha_i \kappa\left(x_i, x\right) + b$$

– On n'a plus d'expression analytique pour w, mais l'on peut toujours calculer $\|w\|_2^2$!

SVM non-linéaire

(1) Φ : transformation de l'espace des entrées dans un espace de représentation $E_{\Phi(\mathcal{X})}$
(2) Discrimination linéaire de marge maximale dans $E_{\Phi(\mathcal{X})}$

$$f_{w,b}(x) = signe\left(\langle w, \Phi(x)\rangle + b\right)$$

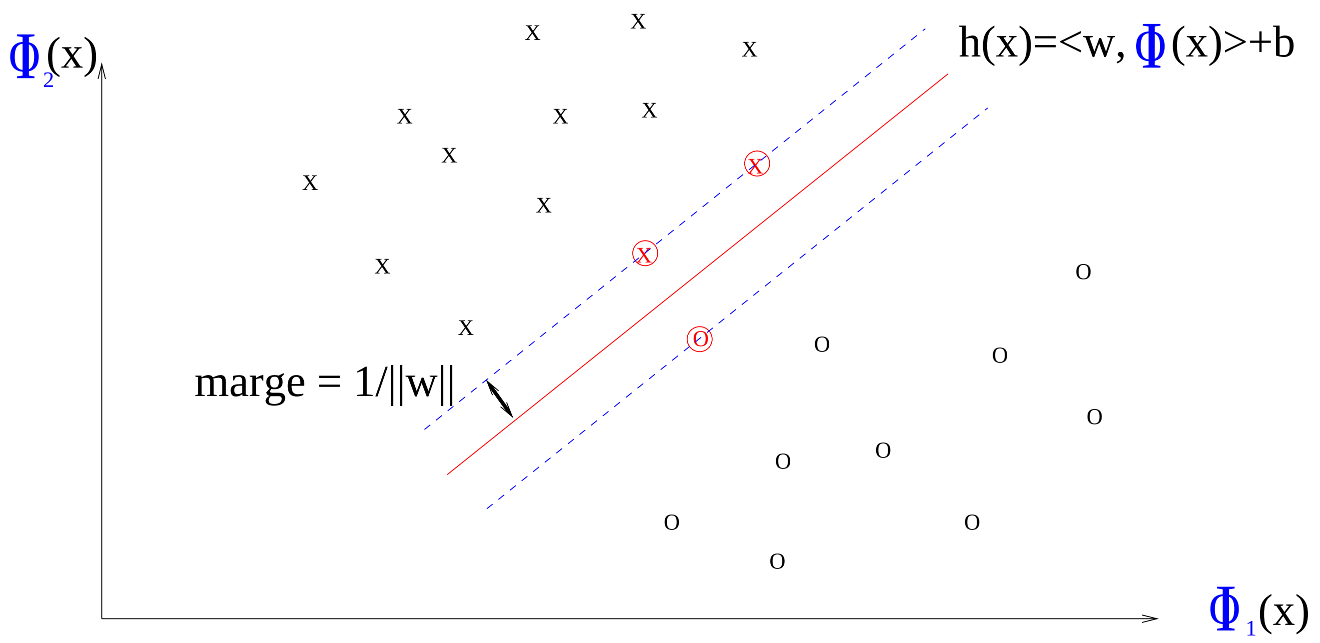


FIG. 1 – Frontière de décision linéaire dans l'espace de représentation

SVM multi-classes (M-SVM)

Architecture (commune à toutes les machines)

Fonctions $h = (h_k)_{1 \leq k \leq Q}$ calculées par l'architecture :

$$\forall x \in \mathcal{X},\ \forall k \in \{1, \ldots, Q\}\,,\ h_k(x) = \tilde{h}_k(x) + b_k = \langle w_k, \Phi(x) \rangle + b_k$$

où

- Φ est soit une transformation dans un espace de représentation soit l'identité
- $\tilde{\mathcal{H}} = \left\{ \tilde{h} = \left(\tilde{h}_k \right) \right\}$ est le produit de Q RKHS
- κ, le noyau associé à $\tilde{\mathcal{H}}$, s'exprime en fonction de Φ comme :

$$\forall (x, x') \in \mathcal{X}^2,\ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

Algorithme d'apprentissage des M-SVM : principe général

Soit $s_m = \{(x_i, y_i)) \ : 1 \leq i \leq m\}$ l'ensemble d'apprentissage

Formulation primale (avec *marges molles*)

Problème 1

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \|\tilde{h}\|^2_{\tilde{\mathcal{H}}} + C \sum_{i=1}^{m} \ell\left(y_i, h(x_i)\right) \right\}$$

Formulation duale

Un *théorème de représentation* établit que l'apprentissage (résolution du Problème 1) revient à déterminer les valeurs des coefficients β_{ik} dans :

$$\forall k \in \{1, \ldots, Q\}, \ h_k(x) = \sum_{i=1}^{m} \beta_{ik} \kappa(x_i, x) + b_k$$

Algorithmes d'apprentissage des M-SVM (formulation primale)

Problème 2 (M-SVM1 (Vapnik et Blanz, 1998, Weston et Watkins, 1998,...))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.c. \begin{cases} \langle w_{y_i} - w_k, \Phi(x_i) \rangle + b_{y_i} - b_k \geq 1 - \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \end{cases}$$

Problème 3 (M-SVM2 (Guermeur, 2002))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} t^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.c. \begin{cases} \|w_k\|^2 \leq t^2, \ (1 \leq k \leq Q) \\ \textit{Contraintes du Problème 2} \end{cases}$$

Algorithmes d'apprentissage des M-SVM (formulation primale)

Problème 4 (M-SVM3 (Crammer et Singer, 2001))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \xi_i \right\}$$

$$s.c.\ \langle w_{y_i} - w_k, \Phi(x_i) \rangle + b_{y_i} - b_k + \delta_{y_i,k} \geq 1 - \xi_i,\ (1 \leq i \leq m), (1 \leq k \leq Q)$$

Problème 5 (M-SVM4 (Lee *et al.*, 2001))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.c. \begin{cases} \sum_{k=1}^{Q} w_k = 0, \quad \sum_{k=1}^{Q} b_k = 0 & \\ \langle w_k, \Phi(x_i) \rangle + b_k \leq -1/(Q-1) + \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \end{cases}$$

Comportement d'une M-SVM (M-SVM1)

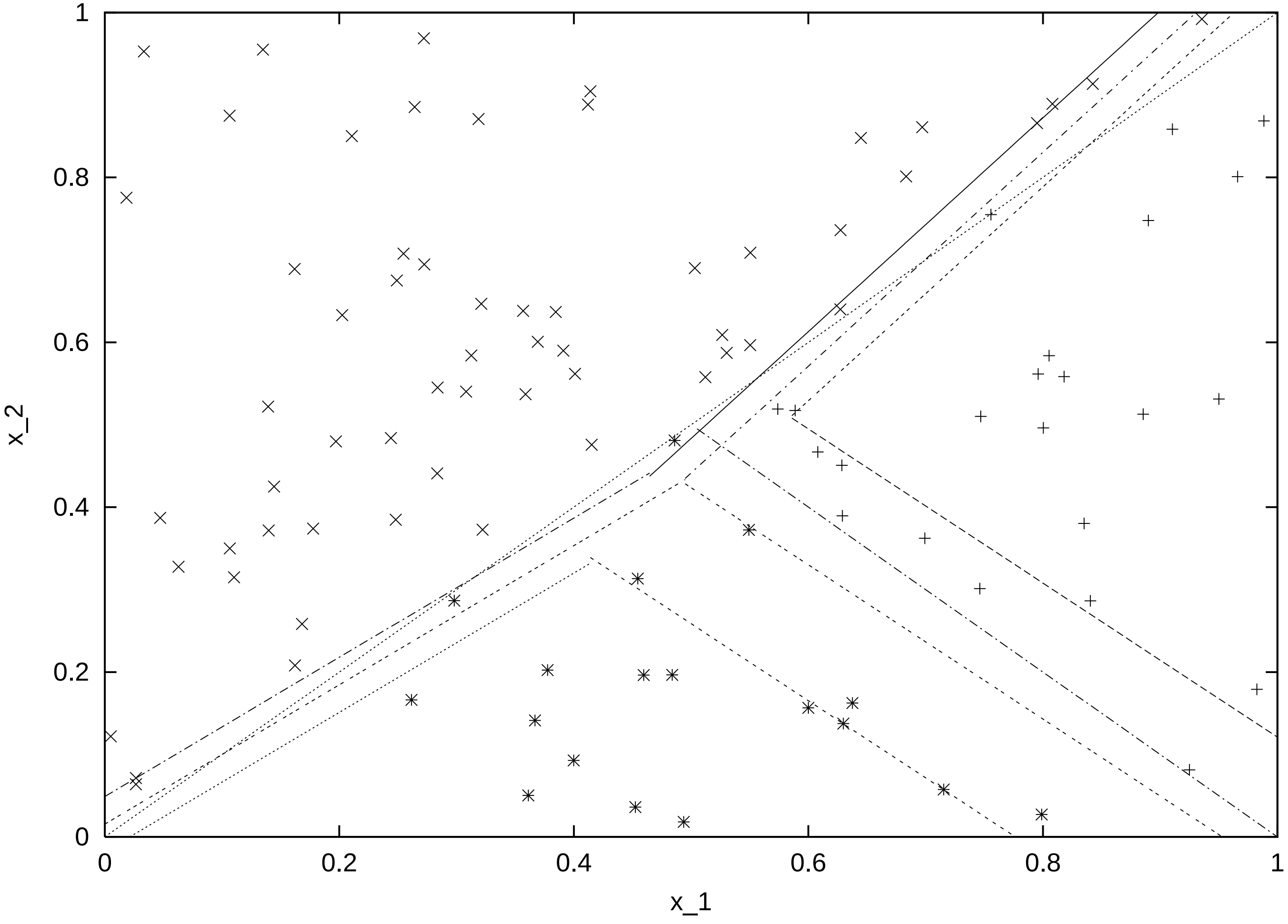


FIG. 2 – 3 catégories linéairement séparables en 2D

Comportement d'une M-SVM (M-SVM1)

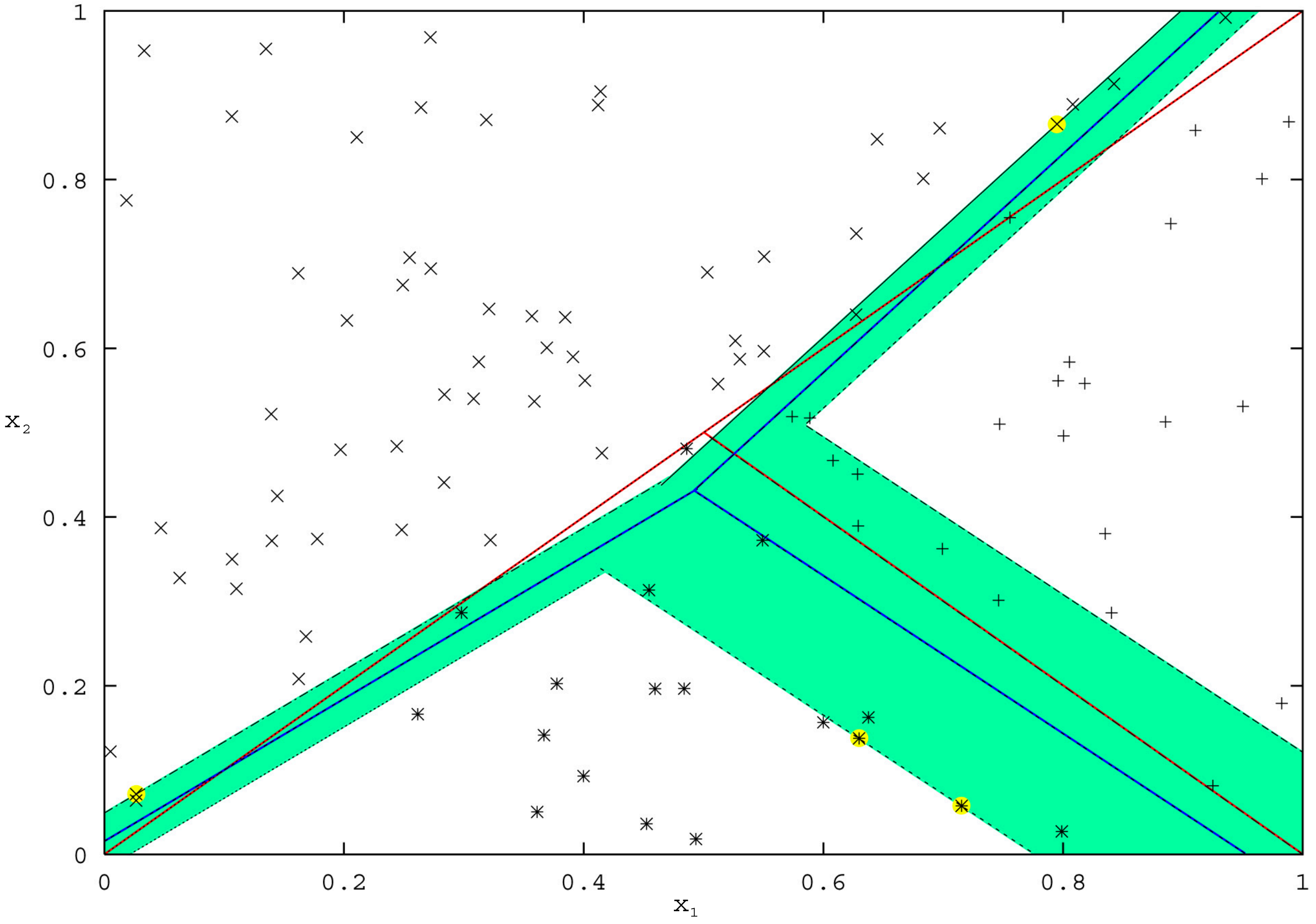


FIG. 3 – Hyperplans de la M-SVM linéaire et vecteurs support

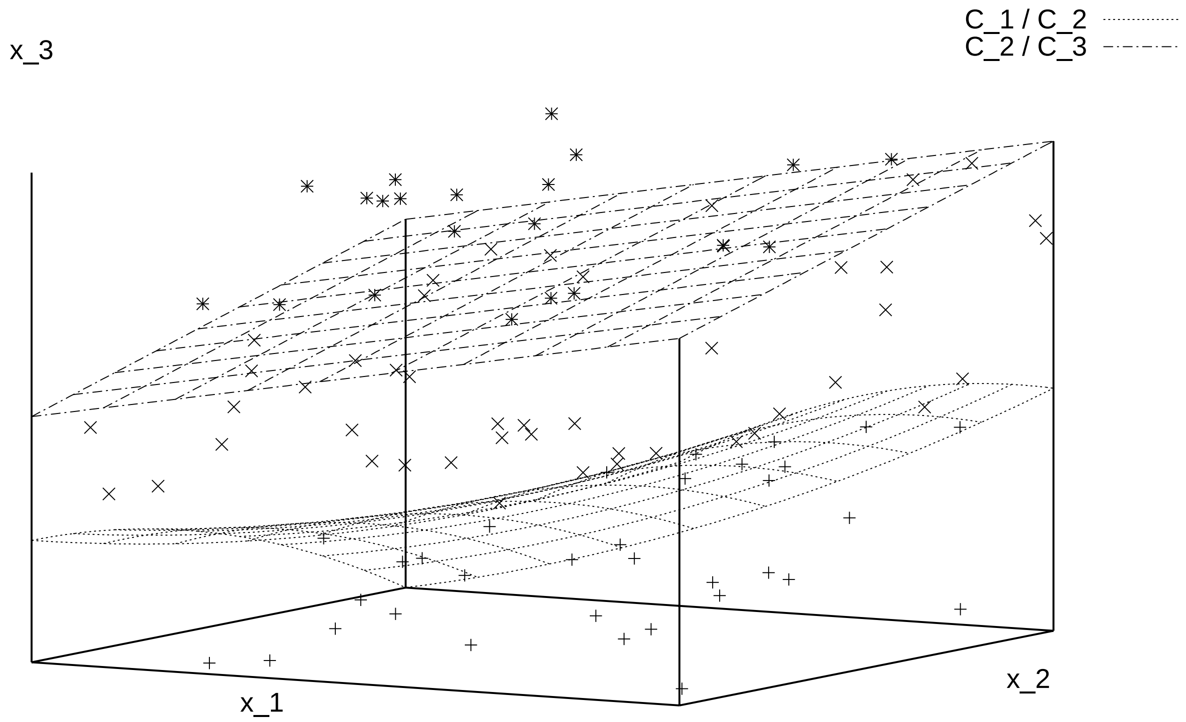


FIG. 4 – 3 catégories non linéairement séparables en 3D

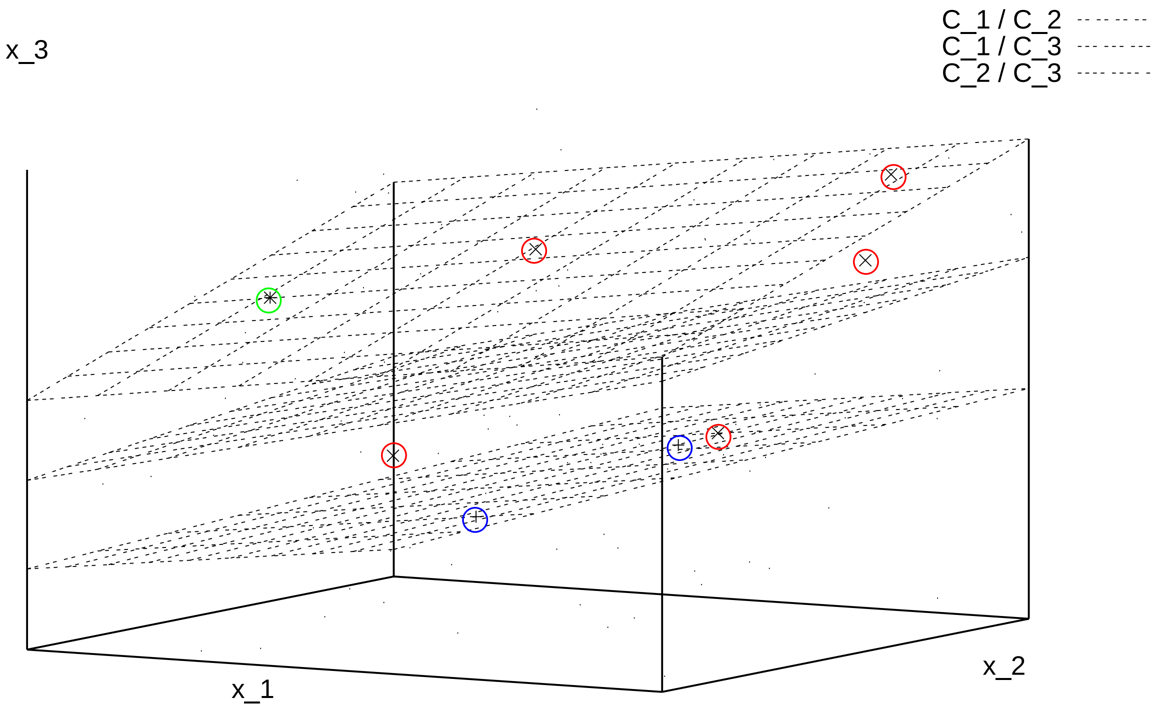
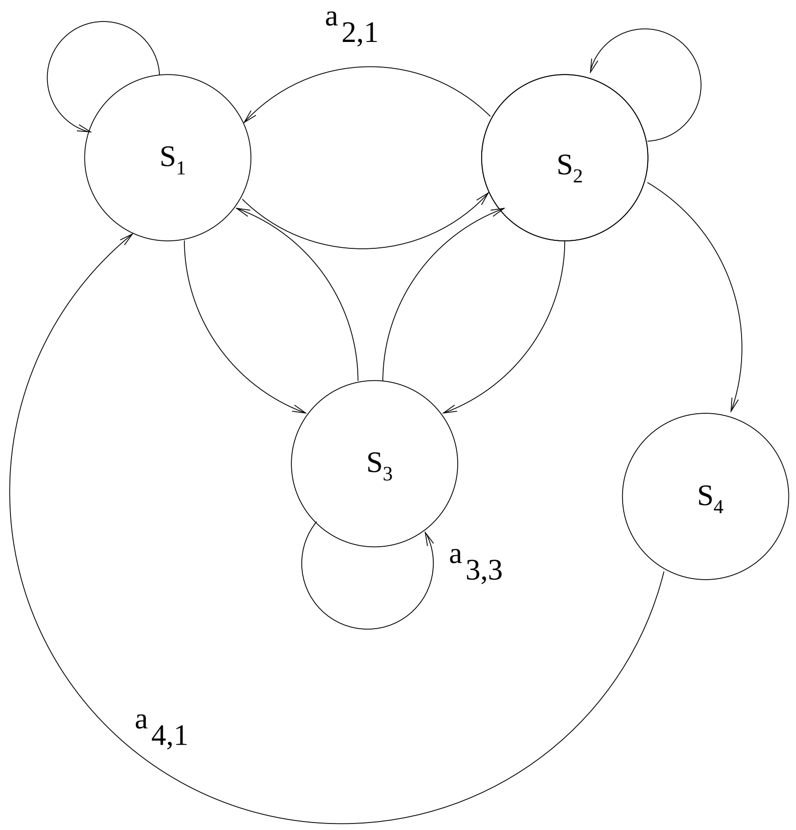


FIG. 5 – Hyperplans de la M-SVM linéaire et vecteurs support

Modèles de Markov cachés (HMM)

Caractérisation

Modèles probabilistes qui constituent une extension des *chaînes de Markov*. Ils font intervenir **deux** *processus stochastiques* encapsulés.



Modèles de Markov cachés (HMM)

Architecture et paramètres d'un modèle

- $S = \{S_1, \ldots, S_i, \ldots, S_N\}$: ensemble d'états
- $V = \{v_1, \ldots, v_k, \ldots, v_M\}$: ensemble de symboles

Soit (q_t, O_t) le couple état-observation à l'"instant" t. On note :

$$a_{ij} = P\{q_{t+1} = S_j \mid q_t = S_i\}, \quad (1 \leq i, j \leq N)$$

$$b_i(k) = P\{O_t = v_k \mid q_t = S_i\}, \quad (1 \leq i \leq N), \quad (1 \leq k \leq M)$$

$$\pi_i = P\{q_1 = S_i\}, \quad (1 \leq i \leq N)$$

$\lambda = \{A, B, \pi\}$: ensemble des paramètres

Trois problèmes fondamentaux

Problème 1

Etant donnés une séquence d'observations $O = (O_t)_{(1 \leq t \leq T)}$ et un modèle $\lambda = \{A, B, \pi\}$, comment calculer $P(O \mid \lambda)$?

Problème 2

Etant donnés une séquence d'observations $O = (O_t)_{(1 \leq t \leq T)}$ et un modèle $\lambda = \{A, B, \pi\}$, comment déterminer la séquence d'états rendant le mieux compte des données ?

Problème 3

Etant donnés une séquence d'observations $O = (O_t)_{(1 \leq t \leq T)}$ et un modèle $\lambda = \{A, B, \pi\}$, comment optimiser les paramètres afin de maximiser $P(O \mid \lambda)$?

Calcul de $P(O \mid \lambda)$

Procédure "avant-arrière"

- Variable "avant" : $\alpha_t(i) = P\left\{O_1, O_2, \ldots, O_t,\ q_t = S_i \middle|\ \lambda\right\}$

`Initialisation`

$$\alpha_1(i) = \pi_i b_i(O_1) \ \ (1 \leq i \leq N)$$

`Induction`

$$\alpha_{t+1}(j) = \left[\sum_i \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \ \ (1 \leq t \leq T-1) \ \ (1 \leq j \leq N)$$

`Terminaison`

$$P(O \mid \lambda) = \sum_i \alpha_T(i)$$

- Variable "arrière" : $\beta_t(i) = P\left\{O_{t+1}, O_{t+2}, \ldots, O_T,\ q_t = S_i \middle|\ \lambda\right\}$

Meilleure séquence d'états - Algorithme de Viterbi

- $\delta_t = \max_{q_1,q_2,\ldots,q_{t-1}} P\{q_1, q_2, \ldots, q_t = S_i,\ O_1, O_2, \ldots, O_t \mid \lambda\}$

$$\delta_{t+1}(j) = \left[\max_i (\delta_t(i) a_{ij})\right] b_j(O_{t+1})$$

`Initialisation`

$$\delta_1(i) = \pi_i b_i(O_1), \quad (1 \le i \le N)$$

$$\psi_1(i) = \emptyset$$

`Procédure récursive`

$$\delta_t(j) = \max_i (\delta_{t-1}(i) a_{ij})\, b_j(O_t), \quad (2 \le t \le T), \quad (1 \le j \le N)$$

$$\psi_t(j) = \arg\max_i [\delta_{t-1}(i) a_{ij}] \quad (2 \le t \le T), \quad (1 \le j \le N)$$

`Terminaison`

$$P^* = \max_i \delta_T(i)$$

$$q_T^* = \arg\max_i \delta_T(i)$$

`Recherche ‘‘arrière’’ de la meilleure séquence d’états`

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \ldots, 1$$

Apprentissage des paramètres

Maximisation de $P(O \mid \lambda)$

Il n'existe **pas d'algorithme fournissant un maximum global**

Algorithme de Baum-Welch

Mise en œuvre du principe de maximum de vraisemblance :

algorithme *EM* $\Longrightarrow$ *optimum local*

$$\Longleftrightarrow$$

Maximisation par rapport à $\bar{\lambda}$ de la fonction auxiliaire de Baum

$$J(\lambda, \bar{\lambda}) = \sum_{Q} P(Q \mid O, \lambda) \log \left\{ P(O, Q \mid \bar{\lambda}) \right\}$$

$$\bar{\lambda}^* = \arg\max_{\bar{\lambda}} J(\lambda, \bar{\lambda}) \Longrightarrow P(O \mid \bar{\lambda}^*) \geq P(O \mid \lambda)$$

Propriétés de la procédure de réestimation

Satisfaction des contraintes stochastiques que doivent vérifier les paramètres

$$\sum_i \bar{\pi}_i = 1$$

$$\sum_j \bar{a}_{ij} = 1 \ \ (1 \leq i \leq N)$$

$$\sum_k \bar{b}_i(k) = 1 \ \ (1 \leq i \leq N)$$

Estimation des paramètres et optimisation sous contraintes

(Nouvelle) utilisation des multiplicateurs de Lagrange !

De la statistique à la biologie

Biologie : une source de beaux problèmes pour les statisticiens et les informaticiens

• L'évolution est-elle fondée sur un ensemble de processus stochastiques ?

• Possibilité d'évaluer les modèles dans tout le spectre des régimes, des petits échantillons jusqu'aux comportements asymptotiques

Modèles statistiques et état de l'art

Les modèles statistiques fournissent actuellement les meilleures performances en :
– biologie structurale prédictive ;
– identification des régions codantes de l'ADN génomique ;
– phylogénie...

Prédiction de la structure et apprentissage automatique Problème d'apprentissage sur de grandes quantités de données

Contexte biologique : Exploitation fonctionnelle des informations provenant des grands programmes de séquençage des génomes : passe par la connaissance de la structure 3D des protéines. C'est cette structure 3D qui est responsable de la fonction biologique.

1. Arrivée massive de séquences protéiques (croissance exponentielle des bases)

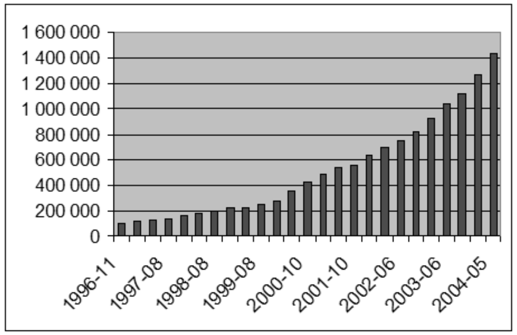


FIG. 6 – Croissance de la banque internationale TREMBL de 1996 à 2004

2. Détermination expérimentale de la structure 3D : tâche très lourde... lorsqu'elle est réalisable

$\Longrightarrow$ Nécessité de passer d'une approche biochimique à une approche prédictive

Problème central en biologie permettant d'aborder
l'essentiel des grandes questions ouvertes en traitement de données séquentielles

Différents niveaux d'organisation structurale des protéines

- **Séquence ou structure primaire (1 536 117 séquences connues)**

`MEEKLKKAKIIFVVGGPGSGKGTQCEKIVQKYGYTHLSTC...`

- **Structure secondaire**

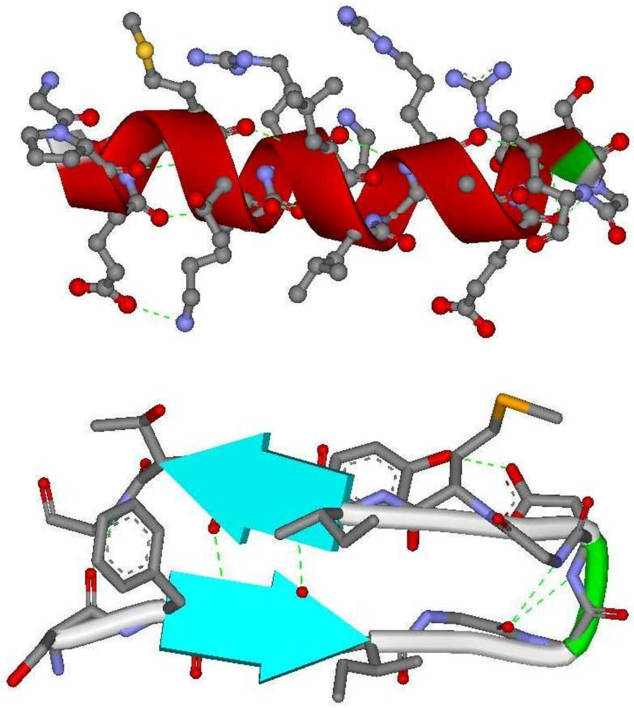
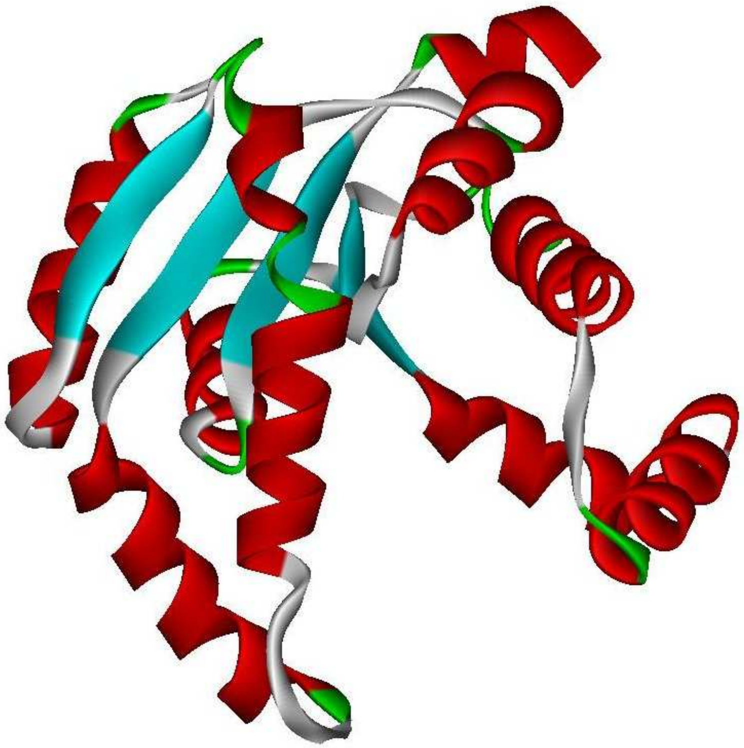


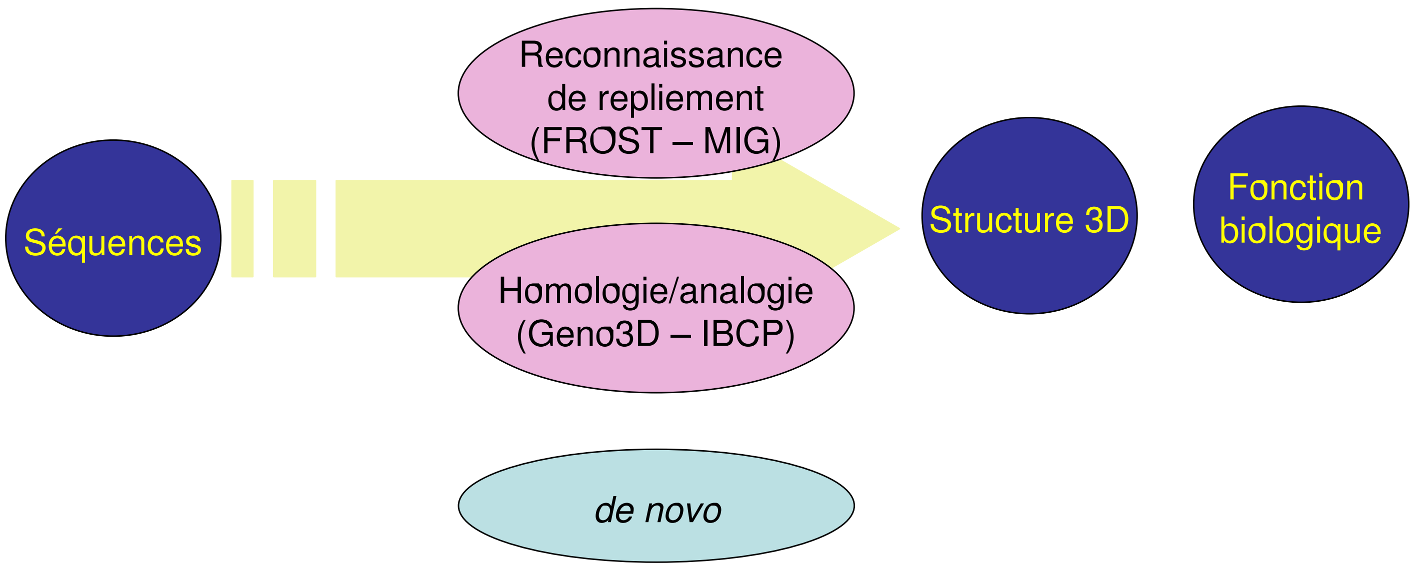
FIG. 7 – Elements structuraux périodiques : hélice α (haut) et brins β (bas)

Différents niveaux d'organisation structurale des protéines

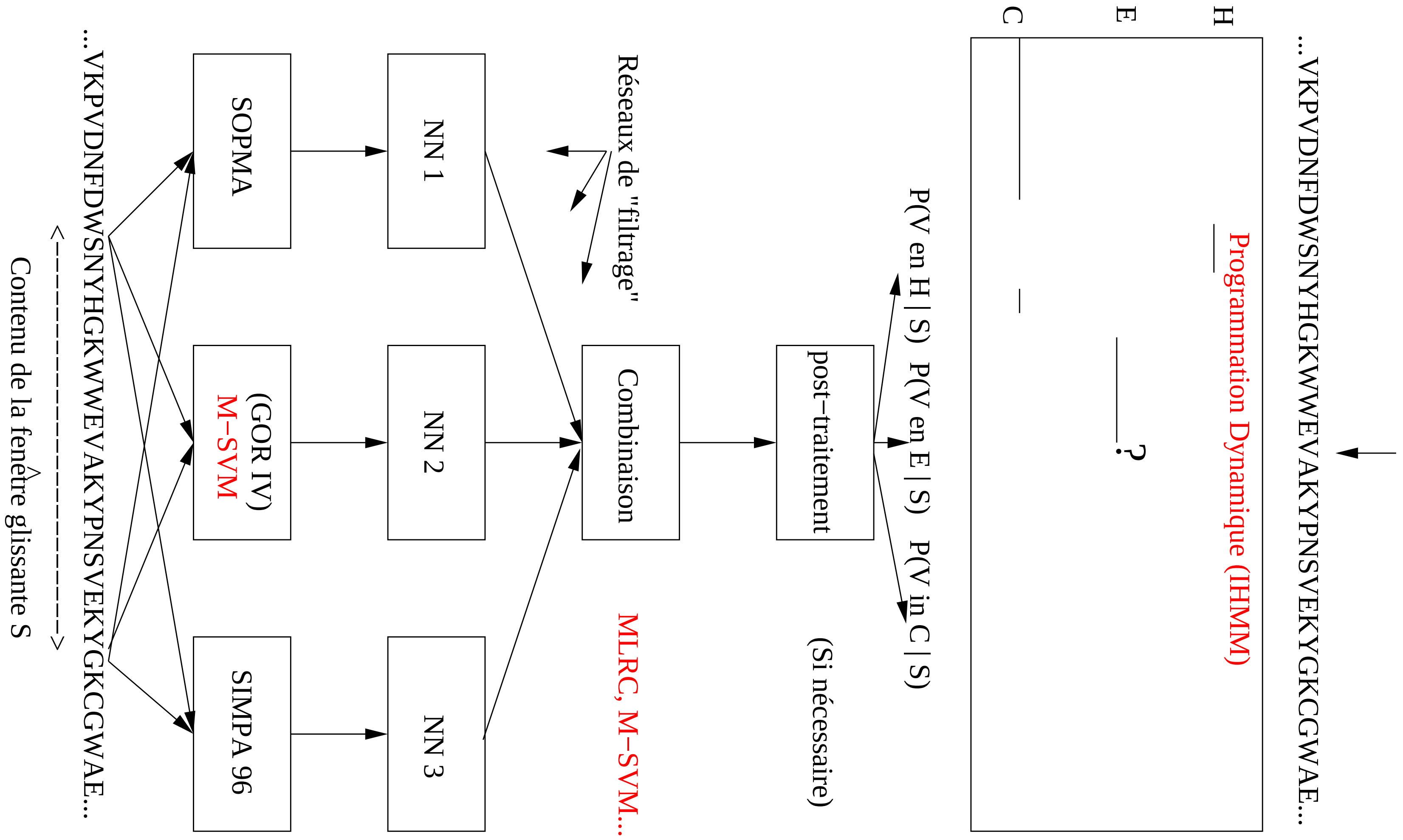
- **Structure tertiaire (27 112 structures 3D connues)**



Stratégie “diviser pour régner” pour prédire la structure 3D



Approche hybride pour la prédiction de la structure secondaire



Classifieurs mettant en œuvre une approche locale

Système connexionniste hybride

- Réseau “séquence-vers-structure” et réseau “structure-vers-structure”
- Codage adaptatif “compact” du contenu de la fenêtre d’analyse
- Extension de la taille du contexte efficace par introduction de connexions récurrentes
- Incorporation de données physico-chimiques

M-SVM

- Noyau dédié au traitement des séquences protéiques

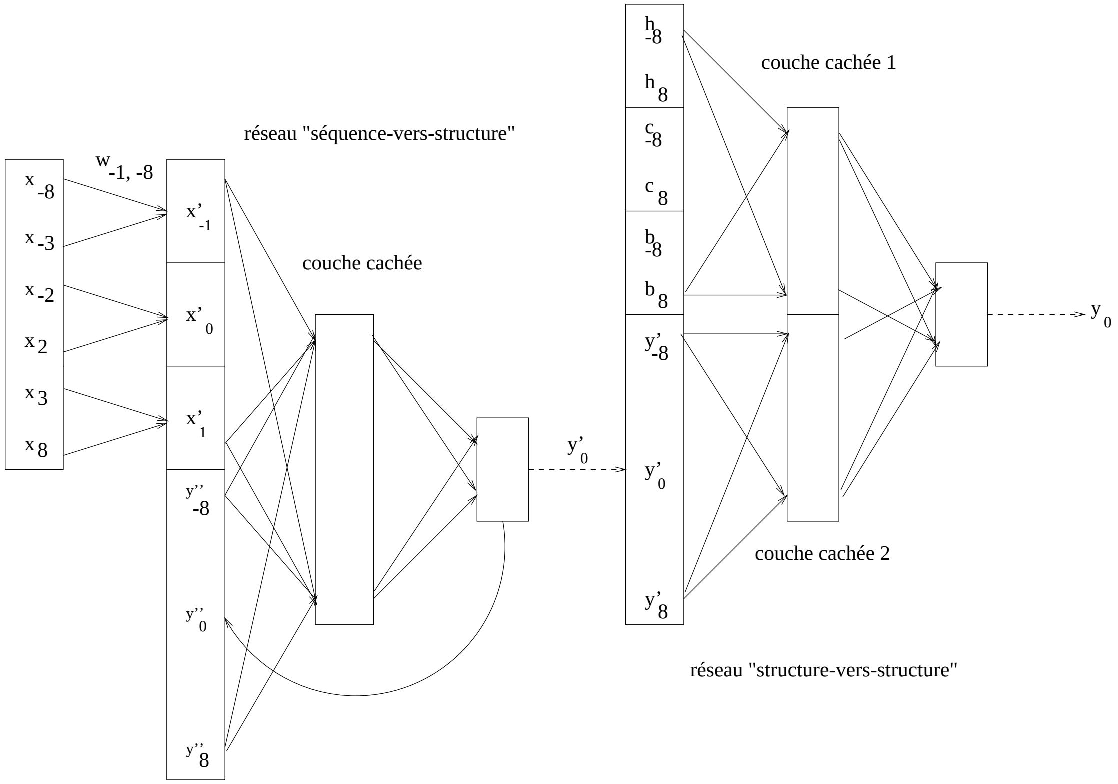
Module statistique

- Prise en compte explicites de dépendances d’ordres supérieurs

Algorithme des plus proches voisins

- Pondération des composantes de la (pseudo) métrique afin d’optimiser le pouvoir discriminant du modèle

Modèle neuronal hiérarchique



$$x'_i = \sum_{j=1+\delta_i}^{|W_i|+\delta_i} w_{ij} x_j \quad w_{ij} = sigm(w'_{ij}) \quad \Delta w'_{ij} = -\epsilon \frac{\partial J}{\partial w'_{ij}}$$

- En test, le premier réseau évolue librement vers un état stable (prédiction équilibrée)

Nouveau noyau pour le traitement des séquences protéiques

- **Expression analytique (structure primaire seule)**

(x, x') : vecteurs codant les contenus des deux fenêtres d'analyse à comparer

$$\kappa_{\mu,\theta,A}(x, x') = \exp\left(-\mu \sum_{i=-n}^{n} \theta_i^2 (\|x_i\|^2 + \|x'_i\|^2 - 2\langle x_i, x'_i\rangle)\right)$$

- **Extension au traitement des alignements multiples**

Immédiate : x est remplacé par $\tilde{x}$ tel que $\tilde{x}_i = \sum_{j=1}^{22} \theta_{ij} a_j$

$$\langle \tilde{x}_i, \tilde{x}'_i \rangle = \langle \sum_{j=1}^{22} \theta_{ij} a_j, \sum_{k=1}^{22} \theta'_{ik} a_k \rangle = \sum_{j=1}^{22} \sum_{k=1}^{22} \theta_{ij} \theta'_{ik} \langle a_j, a_k \rangle$$

Prise en compte de la nature des substitutions (matrice A)

Matrices de similarité standard

	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W
G	2																			
P	1	3																		
D	0	0	2																	
E	0	-1	1	2																
A	0	-1	0	1	2															
N	0	0	1	0	0	3														
Q	0	0	0	1	0	1	2													
S	0	0	0	0	1	0	0	2												
T	0	0	0	0	0	0	0	0	2											
K	0	0	0	0	0	1	0	0	0	2										
R	0	0	0	0	0	0	0	0	0	1	2									
H	0	0	0	0	0	0	0	0	0	0	0	2								
V	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	2							
I	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	2						
M	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	0	2					
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2				
L	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	0	2	0	2			
F	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	1	0	-1	0	2		
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	-1	0	1	2	
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	0	0	0	-1	0	0	0	2

FIG. 8 – Matrice de similarité en prédiction de la structure secondaire **(Levin *et al.*, 1986)**

Matrice M caractérisant les similarités entre acides aminés : fournie par le biologiste.

Approximation de M par une matrice de Gram

- $A = [a_1, \dots, a_i, \dots, a_{22}]^T \in \mathcal{M}_{22}(\mathbb{R})$: representations (implicites) des acides aminés
- AA^T : matrice des produits scalaires = approximation symétrique semi-définie positive de M

La diagonalisation de M étant donnée par :

$$M = PDP^{-1} = PDP^T$$

(P est orthogonale puisque M est symétrique), alors

$$AA^T = M_+ = PD_+P^T$$

où D_+ est déduite de D en annulant les valeurs propres négatives

On obtient ainsi :

$$A = P\sqrt{D_+}$$

Estimation du vecteur θ

- **Alignement noyau-cible : cas multi-classe (Vert, 2002)**

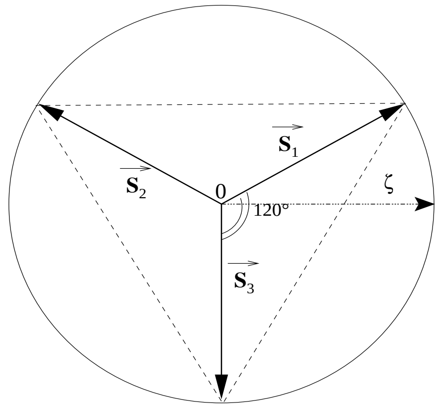


FIG. 9 – Choix optimal des représentants des catégories (cas de 3 classes)

$$\kappa_t(x, x') = \begin{cases} \zeta^2 & \text{si } C(x) = C(x') \\ -\zeta^2/(Q-1) & \text{sinon} \end{cases}$$

- **Algorithme d'apprentissage**

Descente de gradient stochastique

Module statistique

Développement de Bahadur-Lazarsfeld (1961)

• Représentation de la probabilité jointe p des réponses $x = [x_1, \ldots, x_i, \ldots, x_n]^T$ à n questions dichotomiques ($X = (X_1, \ldots, X_i, \ldots, X_n)$ prend ses valeurs dans $\{0, 1\}^n$) :

$$p(X = x) = p_1(X = x) f(x)$$

$$p_1(X = x) = \prod_i \alpha_i^{x_i} (1 - \alpha_i)^{(1-x_i)}$$

$$f(x) = 1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<k} r_{ijk} z_i z_j z_k + \ldots + r_{1\ldots i\ldots n} z_1 \ldots z_i \ldots z_n$$

$\alpha_i = \mathbb{E}_p(X_i),\ Z_i : X_i$ normalisée

$r_{ij} = \mathbb{E}_p(Z_i Z_j), \ldots, r_{1\ldots i\ldots n} = \mathbb{E}_p(Z_1 \ldots Z_i \ldots Z_n)$

Module statistique

Application à la prédiction de la structure secondaire des protéines

x_{ij} : 1 si le i-ième résidu de la fenêtre d'analyse est l'acide aminé d'indice j, 0 sinon Fenêtre glissante : matrice $x = [x_{ij}]$, $(1 \leq i \leq |W|)$, $(1 \leq j \leq 22)$

$$\hat{p}(x|C_k) = p_2(x|C_k) = p_1(x|C_k)\left(1 + \sum_{22.i+j<22.l+m} r_{ijlmk} z_{ijk} z_{lmk}\right)$$

• Estimations des probabilités a posteriori des classes : obtenues par application de la formule de Bayes

$$\hat{p}(C_1|x) \propto \hat{p}(x|C_k)\hat{p}(C_k)$$

• Différentes solutions pour assurer la positivité des estimations

Algorithme des plus proches voisins

Pseudo-distance entre deux segments

- Pseudo-distance entre les contenus de deux fenêtres

$$d(x,x') = \sum_{i=1}^{|W|} d_i(x,x')$$

- Pseudo-distance entre les acides aminés de même rang dans les deux fenêtres

$$d_{1i}(x,x') = \sum_{k=1}^{3} |p(C_k|x_i) - p(C_k|x'_i)| + \lambda_1 \sum_{k=1}^{3} \sum_{j=i+1}^{|W|} |p(C_k|x_i,x_j) - p(C_k|x'_i,x'_j)|$$

$$\begin{aligned} d_{2i}(x,x') \quad &= \quad \lambda_2 \textstyle\sum_{k=1}^{3} \left(1 - \frac{min(p(C_k|x_i),p(C_k|x'_i))}{max(p(C_k|x_i),p(C_k|x'_i))}\right) \\ &+ \quad \lambda_3 \textstyle\sum_{k=1}^{3} \sum_{j=i+1}^{|W|} \left(1 - \frac{min(p(C_k|x_i,x_j),p(C_k|x'_i,x'_j))}{max(p(C_k|x_i,x_j),p(C_k|x'_i,x'_j))}\right) \end{aligned}$$

$$\forall i \in \{1,\ldots,|W|\}\,,\quad d_i(x,x') = d_{1i}(x,x') + d_{2i}(x,x')$$

Algorithme des plus proches voisins

Apprentissage discriminant : détermination du vecteur λ

$\lambda = [\lambda_1, \lambda_2, \lambda_3]^T \in \mathbb{R}_+^3$

λ : choisi de manière à optimiser un critère *inertiel* (principe emprunté à l'algorithme des centres mobiles)

$\Rightarrow$ λ est la solution d'un problème de programmation mathématique

Estimation des probabilités a posteriori des classes

Algorithme en quatre étapes

1. Déterminer l'ensemble E des 30 plus proches voisins du segment à assigner
2. Répartir les élements de E en fonction de l'état conformationnel de leur résidu central
3. Pour chaque classe, calculer la somme des inverses des distances correspondantes
4. Normaliser

Sommet de la hiérarchie, IHMM : traitements de haut niveau

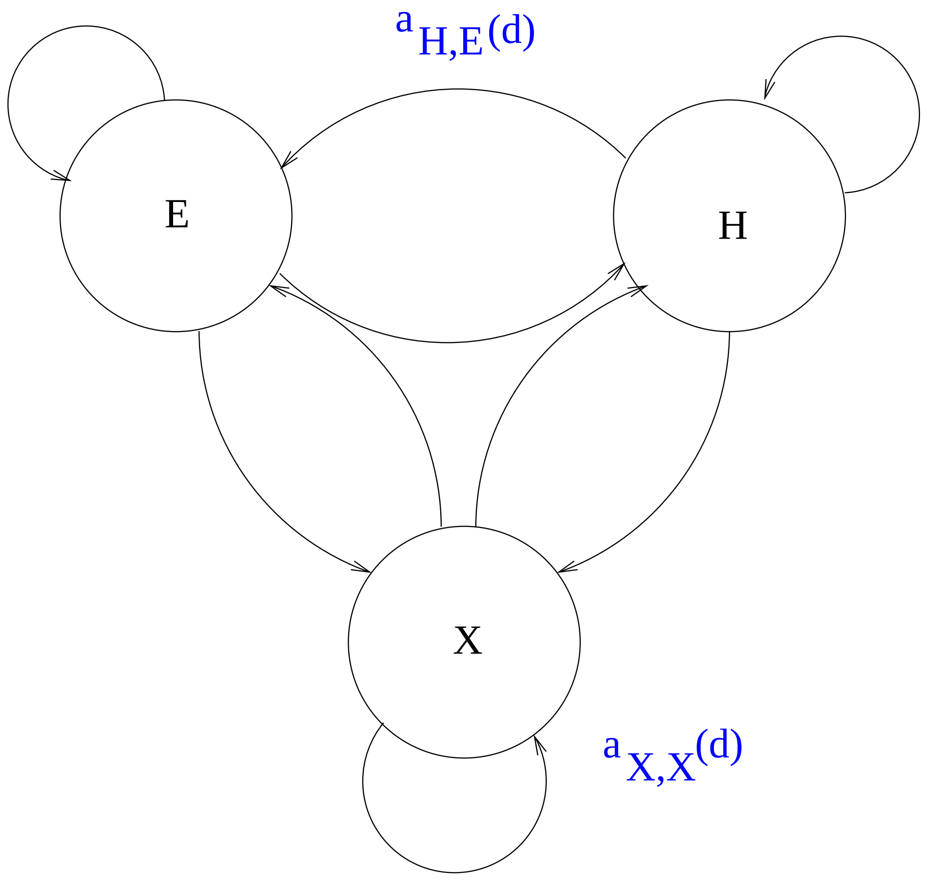


FIG. 10 – Topologie du IHMM utilisé pour le post-traitement des sorties des experts

$\longrightarrow$ distributions des longueurs des éléments structuraux mieux respectées

Bibliographie

M. Anthony et P.L. Bartlett (1999). *Neural Network Learning : Theoretical Foundations*. Cambridge University Press.

M. Anthony (1997). Probabilistic Analysis of Learning in Artificial Neural Networks : The PAC Model and its Variants. *Neural Computing Surveys*, **1**, 1-47.

N. Aronszajn (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **68**, 337-404.

P. Baldi et S. Brunak (1998). *Bioinformatics, the machine learning approach*. The MIT Press, Cambridge, Ma.

P. Baldi, S. Brunak, P. Frasconi, G. Soda et G. Pollastri (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937-946.

P. Barde et M. Ledoux (1998). *Probabilité*. Belin.

B. Boser, I. Guyon et V.N. Vapnik (1992). A training algorithm for optimal margin classifiers. Actes du cinquième workshop annuel *COLT*, 144-152.

L. Bottou (1997). La mise en œuvre des idées de Vladimir N. Vapnik. Dans *Statistique et méthodes neuronales*, édité par S. Thiria, Y. Lechevalier, O. Gascuel et S. Canu, DUNOD, Chap. 16, 262-274.

G. Brightwell, C. Kenyon et H. Paugam-Moisy (1997). Multilayer neural networks : one or two hidden layers ? *NIPS'9*, 148-154.

C. Combet, M. Jambon, G. Deléage et C. Geourjon (2002). Geno3D : automatic comparative molecular modelling of protein. *Bioinformatics*, **18**, 213-214.

C. Cortes et V.N. Vapnik (1995). Support-vector networks. *Machine Learning*, **20**, 273-297.

K. Crammer and Y. Singer (2001). On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research*, **2**, 265-292.

N. Cristianini et J. Shawe-Taylor (2000). *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

F. Cucker et S. Smale (2001). On the Mathematical Foundations of Learning. *Bulletin of the American Mathematical Society*. **39**, 1-49.

G. Cybenko (1989). Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals and Systems*, **2**, 303-314.

R.J. Elliott, L. Aggoun et J.B. Moore (1995). *Hidden Markov Models, estimation and control.* Springer-Verlag, N.Y.

R. Fletcher (1987). *Practical methods of optimization*. John Wiley & Sons.

F. Fogelman-Soulié et *al.* (1997). *Statistiques et méthodes neuronales*. S. Thiria, Y. Lechevallier, O. Gascuel et S. Canu éds., DUNOD.

Y. Guermeur (2002). Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, **5**, 168-179.

Y. Guermeur, C. Geourjon, P. Gallinari et G. Deléage (1999). Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, **15**, 413-421.

Y. Guermeur et O. Teytaud (2006). Estimation et contrôle des performances en généralisation des réseaux de neurones. Dans *Apprentissage connexionniste*. Y. Bennani éditeur, Hermès, Chap. 10, 279-342.

D. Haussler (1992). Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Information and Computation*, **100**, 78-150.

J. Hertz, A. Krogh et R.G. Palmer (1991). *Introduction to the theory of neural computation*. Addison-Wesley, Reading.

W. Hoeffding (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13-30.

A.N. Kolmogorov (1963). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *American Mathematical Society*, **28**, 55-59.

A. Krogh (1998). An introduction to Hidden Markov Models for biological sequences. *Computational Methods in Molecular Biology*, chap. 4. S.L. Salzberg, D.B. Searls et S. Kasif éds., Elsevier.

Y. Lee, Y. Lin et G. Wahba (2001). Multicategory Support Vector Machines. Rapport technique TR-1043, University of Wisconsin, Madison, Department of Statistics.

J.M. Levin, B. Robson et J. Garnier (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS*, **205**, 303-308.

G. Lugosi (2003). Concentration-of-measure inequalities. *Lecture notes, Summer School on Machine Learning at the Australian National University, Canberra.*

G. Lugosi et K. Zeger (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. on Information Theory*, **41**, 677-687.

G. Lugosi et K. Zeger (1996). Concept learning using complexity regularization. *IEEE Trans. on Information Theory*, **42**, 48-54.

A. Marin, J. Potier, K. Zimmermann et J.-F. Gibrat (2002). FROST : A Filter-Based Fold Recognition Method. *PROTEINS : Structure, Function, and Genetics*, **49**, 493-509.

A.B.J. Novikoff (1962). On the convergence proofs on perceptrons. *Proceedings of the Symposium on the Mathematical Theory of Automata*, **XII**, 615-622.

D. Pollard (1984). *Convergence of Stochastic Processes.* Springer Series in Statistics, Springer-Verlag, N.Y.

B. Rost et C. Sander (1993). Prediction of protein secondary structure at better than 70 % accuracy. *J. Mol. Biol.*, **232**, 584-599.

D.E. Rumelhart, G.E. Hinton et R.J. Williams (1986). Learning internal representations by error propagation. *Parallel Distributed Processing - Explorations of the Microstructure of Cognition*, **1**, the MIT Press, Cambridge Ma.

N. Sauer (1972). On the density of families of sets. *Journal of Combinatorial Theory (A)*, **13**, 145-147.

M.J. Schervish (1995). *Theory of Statistics*. Springer Series in Statistics, Springer-Verlag, N.Y.

B. Schölkopf, K. Tsuda et J.-P. Vert (2004). *Kernel Methods in Computational Biology*. The MIT Press, Cambridge.

J. C. Setubal et J. Meidanis (1997). *Introduction to Computational Molecular Biology*. PWS Publishing Company.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson et M. Anthony (1996). Structural risk minimization over data-dependent hierarchies. Rapport technique NC-TR-96-053, NeuroCOLT.

M. Talagrand (1994). Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, **22**, 28-76.

V.N. Vapnik (1982). *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, N.Y.

V.N. Vapnik (1998). *Statistical Learning Theory*. John Wiley & Sons, Inc., N.Y.

V.N. Vapnik et A.Y. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, **16**, 264-280.

V.N. Vapnik et A.Y. Chervonenkis (1974). Ordered risk minimization (I et II). *Automated and Remote Control*, **34**, 1226-1235 et 1403-1412.

V.N. Vapnik et A.Y. Chervonenkis (1991). The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recognition and Image Analysis*, **1**, (3), 284-305.

J. Weston and C. Watkins (1998). Multi-class Support Vector Machines. Rapport technique

CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science.