Régine André-Obrecht

Initiation à l'Apprentissage Automatique

Chap 1 : Méthodes de décision

Chap 2 : Classification automatique

Chap 3 : Fonctions discriminantes linéaires

Chap 4 : Approche statistique

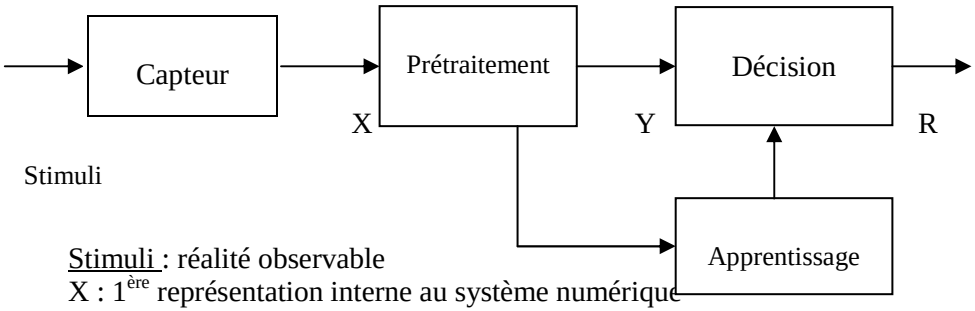
Chap 5 : Méthodes paramétriques et non paramétriques et Décision bayésienne

Chapitre 1 : Méthodes de décision

Reconnaissance des Formes : automatisation d'une tâche de perception réalisée usuellement par le cerveau humain.
Théorie développée depuis les années 1960

Exemples :
- reconnaissances de caractères manuscrits
- reconnaissance d'une personne par sa voix, son visage
- reconnaissance de pathologie musculaire à partir de signaux électromyographiques

I Chaîne de reconnaissance



Stimuli : réalité observable
X : 1ère représentation interne au système numérique
Y : caractéristiques **pertinentes** ou **discriminantes** issues du traitement
Sans perte d'information
R : décision

Capteur : transformation d'un signal physique en un signal numérique

Prétraitement : transformation pour éliminer la redondance, pour avoir une représentation allégée du signal numérique :
- réduction de la dimension de l'espace
- décorrélation des paramètres
- recherche de paramètres discriminants

ce pré traitement dépend de la tâche de reconnaissance

Décision :
- positionnement de la forme captée par rapport aux formes acquises par apprentissage
- assignation d'une classe à un vecteur

Stratégies de décision ou modélisation du processus de classification :
- approche discriminante (utilisation de prototypes, de régions de l'espace…)
- approche probabiliste (distributions probabilistes des observations Y, recherche du minimum d'erreur de classification, prise en compte de contraintes ou de risques)

Apprentissage :
- références type
- distribution probabiliste

L'apprentissage est dit supervisé si un ensemble d'observations, dit ensemble d'apprentissage, est collecté avec connaissance de la forme que chacune des observations représente. Les références ou modèles sont appris à partir de cet ensemble.
L'apprentissage est non supervise si l'ensemble de données acquises pour définir le module de décision est non étiqueté par une forme. La définition du module de décision passe par la recherche aveugle de classes.

II Exemples

Reconnaissance d'un locuteur par sa voix parmi N locuteurs connus
- enregistrement par un microphone d'une phrase
- transformation de la phrase en suite de vecteurs spectraux (FFT)
- apprentissage : création d'un modèle probabiliste pour chaque locuteur pouvant être reconnu
- décision : choix du locuteur le plus vraisemblable.

Remarque : tout locuteur sera reconnu comme un des N locuteurs, problème du rejet.

Vérification d'une personne à l'aide sa voix et d'une vidéo après déclaration de son identité

- capture du visage par webcam et enregistrement de la voix
- transformation de la phrase en suite de vecteurs spectraux (FFT) et du visage par DCT et ACP
- apprentissage : création d'un modèle probabiliste pour chaque personne pouvant être reconnu
- décision : acceptation ou rejet par vraisemblable.

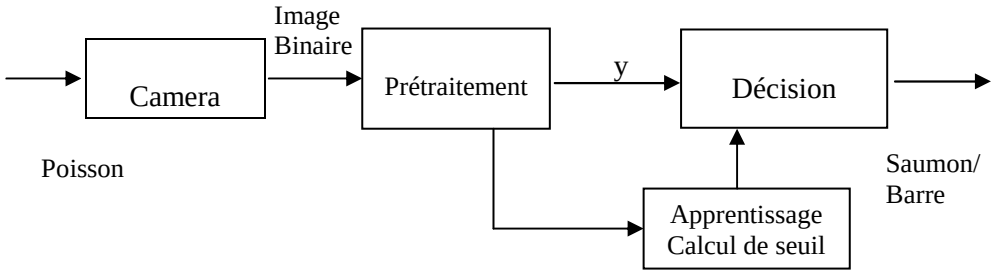
Reconnaissance de caractères manuscrits, d'empreintes digitales

- enregistrement et binarisation de la page manuscrite
- segmentation en caractères (recherche de la ligne de base, de la boîte englobante…)
- apprentissage : recueil d'exemples et /ou création d'un modèle de bruit
- décision : calcul de distances (au sens probabiliste ou non)



Cas d'école : recherche du type de poisson dans une pisciculture

Dans une pisciculture, les poissons sont extraits du bassin automatiquement, il faut ensuite sur une prise d'image du poisson décider sir le poisson est un saumon ou un barre.



Problème à deux classes {saumon, barre}

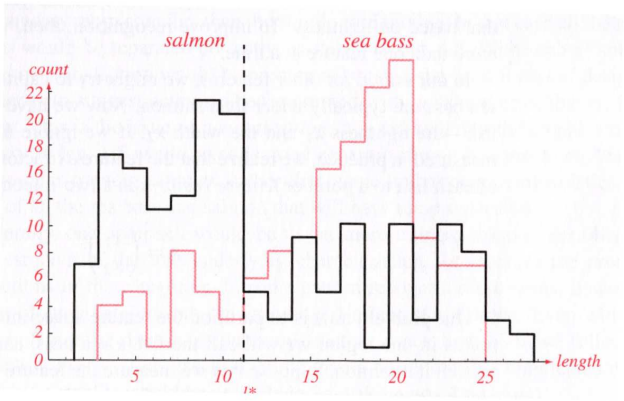
1er cas : approche probabiliste

Le système de décision mis en place est basé sur la luminance de l'image de la planche.

Prétraitement : longueur du poisson sur l'image, $y \in R$.

Apprentissage (ou recherche des connaissances a priori)

- Pourcentage a priori de saumon et de barre => probabilité a priori de chaque classe P(classe)
- Distribution des réalisations de y en fonction de la connaissance du poisson => calcul d'histogrammes => densités de probabilité P(y/classe)



Règle de décision :

Définition d'un seuil λ probabiliste

Minimisation de l'erreur moyenne (critère de base pour l'approche probabiliste bayésienne)

Introduction du coût d'une mauvaise décision

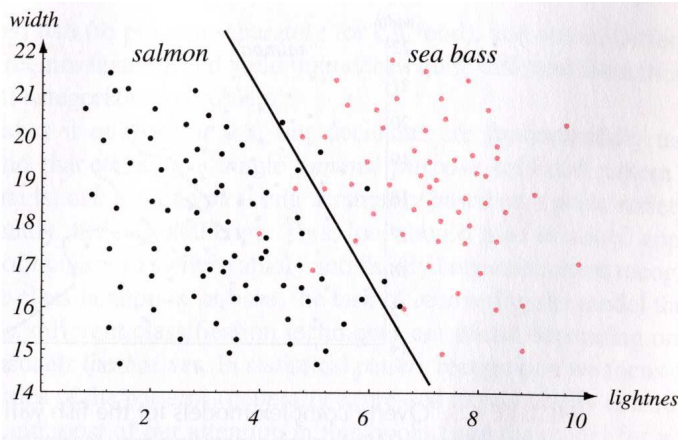
2e cas : approche discriminante

Le système de décision mis en place est basé sur deux paramètres calculés sur l'image (largeur et longueur)

Prétraitement : $y \in R^2$.

Apprentissage :

- Recueil pour chaque classe d'un ensemble d'observations
- Recherche d'une droite séparatrice de deux régions



Règle de décision :

Appartenance à une région

Chapitre 2: Classification automatique

Classification non supervisée d'un ensemble de points,

$$Y = \{y_1, y_2, ..., y_N\}, y_i \in R^d \ :$$

~ trouver une partition du nuage telle qu'au sein de chaque partie les individus se ressemblent et puissent être assimilés à une classe
~ trouver des prototypes ou des références représentatives de l'ensemble

une partie = une classe = une référence = un prototype

I Algorithme de quantification vectorielle ou algorithme des « K means » ou Algorithme des Nuées Dynamiques

Le nombre de classes que l'on recherche est fixé et égal à K.

Par définition, on appelle dictionnaire ou codebook, l'ensemble des références, par analogie avec le codage ; une référence ou un prototype est un codeword.

Notation : d(x,y) désigne une distance entre deux éléments x et y de R^d (distance euclidienne par exemple)

Critère de recherche du dictionnaire D = {d_1, …, d_K} représentant au mieux *Y* :

Trouver D tel qu'il réalise le minimum de la fonction suivante :

$$Crit(D) = \sum_{n=1}^{N} d(y_n, d_{\hat{n}})^2$$

$$\hat{n} = \arg\min_{1 \le k \le K} d(y_n, d_k)$$

$$\hat{D} = \arg\min_{D} Crit(D)$$

Problème : le nombre d'éléments du dictionnaire, K, doit être fixé a priori

Relation entre une partition de Y et un dictionnaire

1- Soit D = {d_1, …, d_K} un dictionnaire de K éléments, appelés , $d_k \in R^d$
On calcule la partition associée :

$$Y = \coprod_{k=1,K} Y_k \qquad \text{avec}$$

$$Y_k = \{y_n \in Y / d(y_n, d_k) \prec d(y_n, d_j), j \neq k\}$$

Cette partition de *Y* est dite associée au dictionnaire D ; chaque Y_k est appelée une classe associée à une référence ou un prototype.

2- Réciproquement soit une partition $Y = \coprod_{k=1,K} Y_k$, on définit le dictionnaire
D = {d_1, …, d_K} avec d_k centre de gravité au sens de la distance employée de la partie Y_k

L'algorithme de quantification vectorielle est basé sur cette alternance de constructions de dictionnaire et de partitionnement de *Y*, ce à partir d'un dictionnaire initial donné que l'on cherche à optimiser au sens du critère *Crit*.

Algorithme de quantification vectorielle :

Soit D_0, un dictionnaire initial de K éléments, t = 0.
Soit D_{t-1}, le dictionnaire obtenu avant la tième itération,

1. effectuer la partition de $Y = \coprod_{k=1,K} Y_k$ associée au dictionnaire D_{t-1},
2. calculer les centres de gravité de chacune des parties de cette partition pour former un nouveau dictionnaire D_t :
   $D_t = \{d^t_1, \ldots, d^t_K\}$
3. calculer $Crit(D_t)$.
4. test d'arrêt : si $(Crit(D_{t-1}) - Crit(D_t)) / Crit(D_{t-1})$ est inférieur à un seuil λ, le dictionnaire optimal est trouvé, $D_{final} = D_t$.
   sinon t = t+1, et retour à l'étape 1.

Mise en œuvre : choix de K et λ.

Variantes :

- algorithme de splitting
- perturbation des classes les plus dispersées
- prise en compte de la distorsion interne des classes

2 Classification hiérarchique - Arbres de décision
Le nombre de classes n'est pas fixé a priori.

2.1 Quelques mesures de dissimilarités entre points et nuages de points

Mesures de similarité entre individus de R^d

$\sigma : R^d \times R^d \rightarrow R^+$
$\sigma(x_1, x_2) \geq 0$
σ est symétrique
$\sigma(x_1, x_2) < \sigma(x_1, x_1)$, $x_1 \neq x_2$

Plus $\sigma(x_1, x_2)$ augmente, plus x_1 et x_2 se ressemblent

Mesure de dissimilarité associée à une mesure de similarité (bornée)

Si σ désigne une mesure de similarité bornée par $\sigma_{max} = \max_{x_1,x_2} \sigma(x_1, x_2)$, on peut définir une mesure de dissimilarité par $dis(x_1,x_2) = \sigma_{max} - \sigma(x_1, x_2)$. Une mesure de dissimilarité proche de la notion de distance, mais elle ne possède pas la propriété de l'inégalité triangulaire.

Distances :

- distance euclidienne (invariante par translation et rotation)
- distance quadratique : $d^2(x_1, x_2) = (x_1 - x_2)^t M (x_1 - x_2)$, avec M matrice dxd définie positive symétrique
- distance de Mahalanobis : distance quadratique avec $M = \Sigma^{-1}$, Σ étant la matrice de covariance d'un nuage de points
- distance de Minkowsky : $d(x_1, x_2) = (\sum_{k=1}^{d} | x_{1,k} - x_{2,k} |^q)^{1/q}$, $q \geq 1$, avec $x_{i,k}$ la $k^{ième}$ coordonnée de x_i.

Dissimilarités entre nuages de points

Soit $(A_j)_{j=1,J}$ un ensemble de J nuages de points de R^d, soit N_i le cardinal de A_i et soit d une distance dur R^d.

- distance min : $d(A_i, A_j) = \min_{x \in A_i, y \in A_j} d(x, y)$
- distance max : $d(A_i, A_j) = \max_{x \in A_i, y \in A_j} d(x, y)$
- distance moyenne : $d(A_i, A_j) = \frac{1}{N_i N_j} \sum_{x \in A_i, y \in A_j} d(x, y)$

- distance entre les moyennes : $d(A_i, A_j) = d(g_i, g_j)$, avec g_i le centre de gravité (ou moyenne) du nuage A_i.

Inertie d'un nuage A de points autour d'un point a (N est le cardinal de A) :

$$I_a(A) = \sum_{x \in A} d^2(x, a)$$

Si a est le centre de gravité g de A, $I_a(A) = I(A)$ est l'inertie du nuage A.

Théorème de Huyghens : Soient $(A_j)_{j=1,J}$ un ensemble de J nuages de points de R^d, N_j le cardinal de A_j, g_j le centre de gravité de A_j, les nuages sont disjoints deux à deux. Alors :

$$I_a(\bigcup_j A_j) = \sum_j I(A_j) + \sum_j N_j d^2(g_j, a)$$

Interprétation : Si a est le centre de gravité de l'union des A_j, cette formule s'interprète comme :
Inertie totale = sommes des inerties intra-classes et de inerties interclasses
On montre que pour J=2 et g centre de gravité de $A_1 \cup A_2$, alors

$$N_1 d^2(g_1, g) + N_2 d^2(g_2, g) = \frac{N_1 N_2}{N_1 + N_2} d^2(g_1, g_2)$$

→ pseudo distance (moment intra nuage) :

$$psd(A_i, A_j) = \frac{N_i N_j}{N_i + N_j} d^2(g_i, g_j)$$

2.2 Construction hiérarchique d'un ensemble de classes sur Y
$Y = \{y_1, y_2, ..., y_N\}, y_i \in R^d$

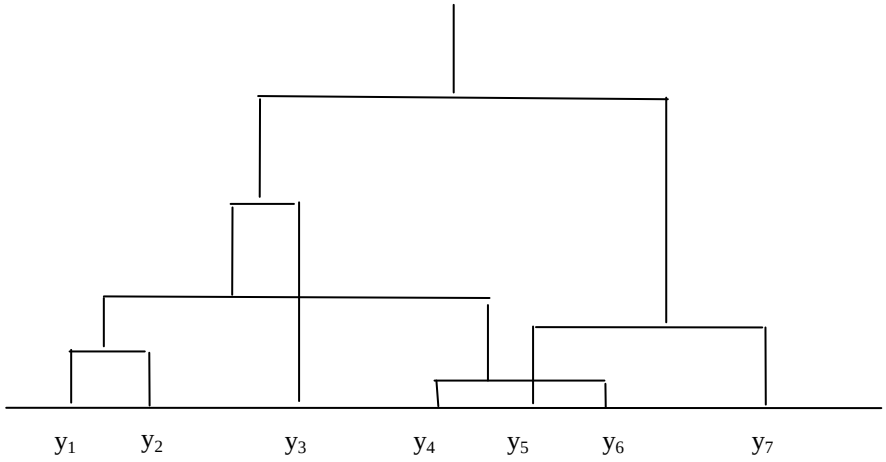
- hiérarchie ascendante :
  - méthode des distances
  - méthode des moments d'ordre 2
- hiérarchie descendante :
  - méthode des moments d'ordre 2

Attention la méthode de splitting (paragraphe 1) est une méthode descendante non hiérarchique
Soit d une mesure de dissimilarité entre nuages de points (groupes de points).

Hiérarchie ascendante :

Processus itératif :
- A la première itération, chaque groupe est constitué d'un seul point de Y, il y a donc N groupes.
- A chaque itération, on regroupe, les deux groupes les plus proches au sens de la dissimilarité d.
- Le processus s'arrête lorsqu'il n'y a plus qu'un groupe : Y est regroupé en une seule classe.

A l'issue du processus, est ainsi construit un arbre (ou dendogramme).
La méthode est dite méthode des distances lorsqu' est utilisée une distance, et elle est dite des moments d'ordre 2, lorsque la dissimilarité est la pseudo distance des moments intra classes.



Hiérarchie descendante :

Processus itératif :

- Avant la première itération, il existe un seul groupe, l'ensemble Y.
- A chaque itération,
  - on choisit un groupe à scinder
  - on scinde le groupe en 2
  - on réaffecte les éléments internes à ces deux nouveaux groupes.
- Le processus s'arrête lorsque les groupes sont tous réduits à un seul élément.
- A l'issue du processus, est aussi construit un arbre (ou dendogramme)

<u>Méthode des moments 2 descendante :</u>

Pour scinder un groupe A_i en deux sous groupes :

prendre chaque paire d'éléments i_1 et i_2 de A_i

- Affecter les éléments de Ai à deux A_{i1} et A_{i2} par la règle des plus proches voisins
- Calculer la pseudo distance des moments pour ces deux groupes $psd(A_{i1}, A_{i2}) = \frac{N_{i1} N_{i2}}{N_{i1} + N_{i2}} d^2(g_{i1}, g_{i2})$

choisir la paire (i_1, i_2) qui maximise cette pseudo distance.

Chapitre 3 : Fonctions discriminantes linéaires

(attention, lorsque l'indice est en exposant, il indique une coordonnée d'un vecteur dans R^d, sinon il s'agit d'un indice d'individu : y_i^j représente la $j^{ème}$ coordonnée de l'individu y_i)

Les observations appartiennent à R^d. Soit $\{k_1, k_2, \ldots, k_K\}$ l'ensemble des classes.
Dans la plupart des cas, un classifieur peut être représenté sous la forme d'un ensemble de fonctions, appelées fonctions discriminantes, $g_i(x)$, i=1,…,K, telles que :

$$y \in k_i \text{ est équivalent à } g_i(y) \geq g_j(y), \forall j .$$

- Les fonctions discriminantes résultent d'une recherche de frontières ou régions, sous forme d'optimisation.
- La discrimination est dite linéaire si les frontières recherchées sont des hyperplans.

Il est classique de rechercher, dès lors que l'on dispose d'un ensemble d'apprentissage étiqueté (apprentissage supervisé) des fonctions discriminantes linéaires, c'est-à-dire des frontières de type hyperplan.

I Cas de deux classes - Algorithme du perceptron

Rappel : <u>équation d'un hyperplan de R^d</u>

Un hyperplan de R^d est déterminé par un point A et un vecteur normal $\vec{\vec{\omega}}$

Tout point M de cet hyperplan vérifie $\overrightarrow{AM}.\vec{\vec{\omega}} = 0$. Il s'en suit que si M a pour coordonnées $(x^1, \ldots, x^d)'$ et $\vec{\vec{\omega}}(\omega^1, \omega^2, ..., \omega^d)'$, l'équation de l'hyperplan est de la forme :

$$\omega^1 x^1 + \omega^2 x^2 + ... + \omega^d x^d + \omega^{d+1} = 0$$

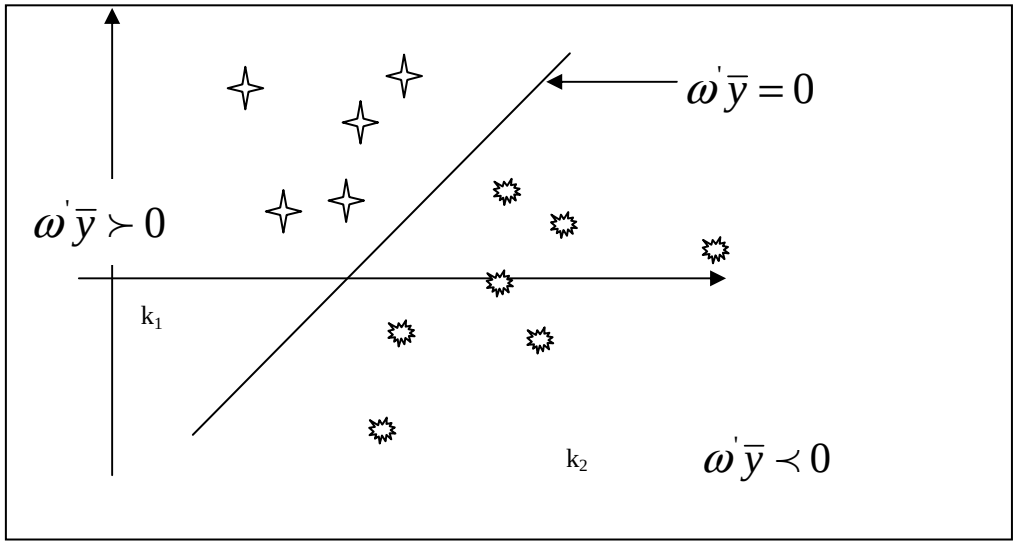
En posant :

$$\bar{x} = \begin{vmatrix} x^1 \\ \ldots \\ x^d \\ 1 \end{vmatrix} \text{ et } \omega = \begin{vmatrix} \omega^1 \\ \ldots \\ \omega^d \\ \omega^{d+1} \end{vmatrix}$$, l'équation devient $\omega' \bar{x} = 0$

<u>Définition</u> : Soit un ensemble de données $\{y_1, y_2, \ldots y_N\}$ de R^d, issues de deux classes $\{k_1, k_2\}$. On dit que les données sont séparables s'il existe

$$\omega = \begin{vmatrix} \omega^1 \\ \ldots \\ \omega^d \\ \omega^{d+1} \end{vmatrix} \text{ tel que}$$

$$\begin{aligned} \omega' \bar{y} &\succ 0 \quad \forall y \in k_1 \\ \omega' \bar{y} &\prec 0 \quad \forall y \in k_2 \end{aligned}$$



Dans le cas de données séparables, une des méthodes pour trouver un hyperplan solution de ce problème, est de rechercher à minimiser la fonction de coût suivante :

$$J(\omega) = \sum_{y \in Y_f} -\delta_y \omega^{'} \bar{y}$$

où Y_f représente l'ensemble des données mal classées avec cet ω et δ_y représente la sortie désirée, c'est-à-dire :

$\delta_y = 1 \; si \; y \in k_1$

$\delta_y = -1 \; si \; y \in k_2$

Algorithme du perceptron (algorithme itératif, de type descente du gradient) pour la mise à jour du vecteur ω :

A la nième itération :

$$\omega^{(n)} = \omega^{(n-1)} + \mu \sum_{y \in Y_f} \delta_y \bar{y}$$

(Attention, il n'y a pas unicité de la solution).

Dans le cas de données non séparables, on recherche un hyperplan séparateur sous optimal qui donnera des résultats satisfaisants en minimisant la fonction de coût suivante :

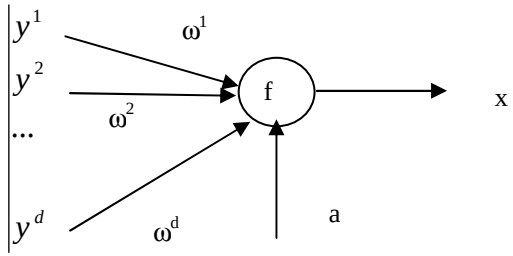
$$J(\omega) = \sum_{y \in Y_f} (\delta_y - \omega^{'} \bar{y})^2$$

la résolution est aussi obtenue par une méthode itérative de type descente du gradient.

II Réseaux de neurones

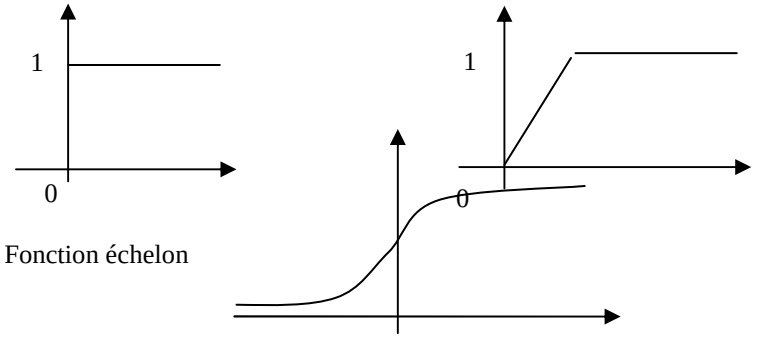
a) la cellule élémentaire d'un réseau de neurones

Elle est constituée de d entrées (*y*), d'une sortie *x*, de d poids synaptiques (ω^i, i=1,…d), et d'un biais *a*



$$x = f(\sum_{i=1}^{d} \omega^i y^i - a)$$

Quelques exemples de fonctions f dites d'activation :

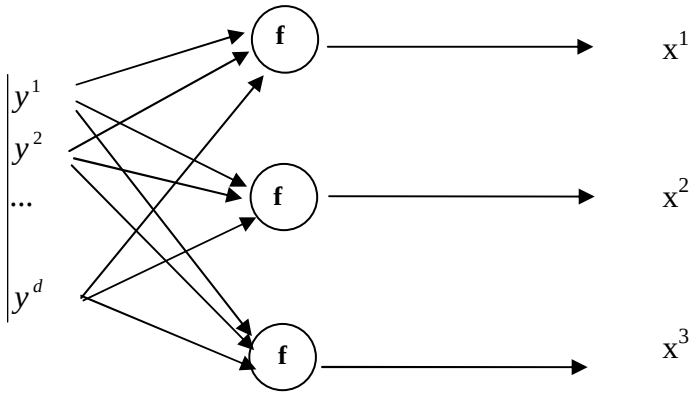


Fonction échelon

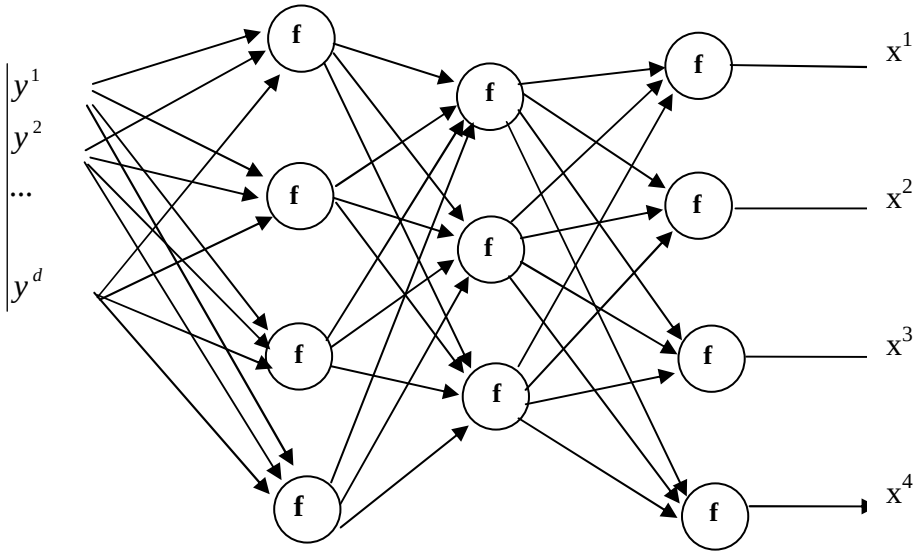
Fonction sigmoïde

Le problème précédent de classification peut se résoudre avec une telle cellule et la fonction échelon.

b) <u>le perceptron à une couche</u> : mise en parallèle de plusieurs cellules élémentaires



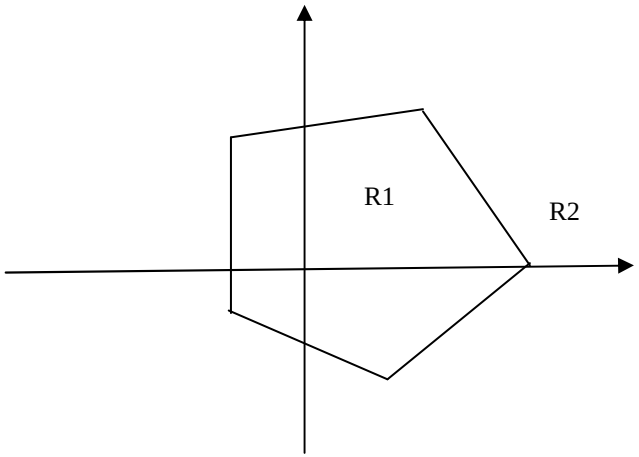
c) le perceptron multicouche (réseau de neurones multicouche)



Chaque sortie peut correspondre à une classe .

<u>Perceptron à deux couches</u> : frontières de décision pour un problème à deux classes de type polygone convexe

- autant de cellules sur la 1ere couche que d'arêtes
- une cellule de sortie qui somme simplement ses entrées pour rendre 1 ou -1 selon que l'individu y appartient à R1 ou R2.



<u>Perceptron à trois couches</u> : union de polygones

Chapitre 4 : Approche statistique

I Exemple à deux classes

Introduction de distributions de probabilités dans l'exemple « cas d'école de la pisciculture » (cf chapitre 1) :

- Deux classes notées k_1 et k_2 (bar et saumon)
- Des observations (la longueur, luminance) appartenant à R, $y \in \Re$

Entrée : facture globale pour l'arrivée des têtards

⇨ définition d'une probabilité *a priori* caractérisant la proportion de saumon et bars a priori en élevage : $P(k_1)$ et $P(k_2)$ (la somme est égale à 1 !)

Histogrammes des observations par type de poisson (cf chapitre 10)

⇨ définition d'un pas de quantification D

⇨ normalisation des histogrammes par le nombre d'individus

⇨ définition d'une densité de probabilité sur *R* par type de poisson : $p(y/k_1)$ et $p(y/k_2)$.

Interprétation : $p(y/k_i)$ est la probabilité d'avoir y comme longueur sachant que l'on observe le poisson k_i

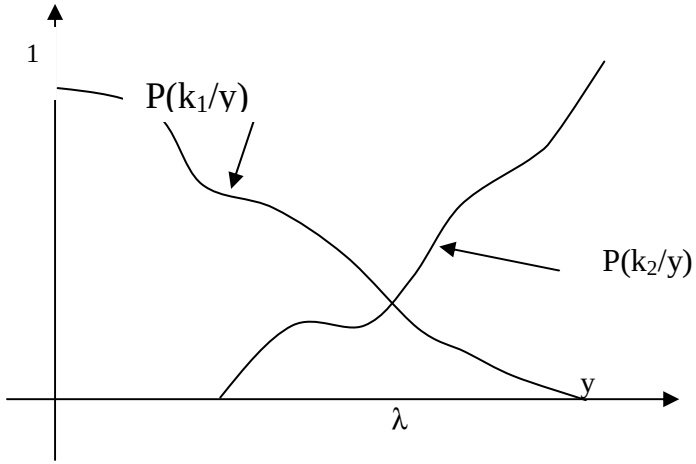
Règle de Bayes : $p(A/B) = p(A \cap B)/P(B)$

Définition de la probabilité *a posteriori* sur l'ensemble des classes :

$$P(k_i/y) = \frac{p(y/k_i)P(k_i)}{p(y)}$$

$$p(y) = \sum_{i=1}^{2} p(y/k_i)P(k_i)$$

Interprétation : $P(k_i/y)$ est, ayant observé la longueur y, la probabilité d'avoir un poisson de type k_i



Règle de décision :

⇨ décision k_1 si et seulement si $P(k_1/y) > P(k_2/y)$ (i différent de j)

⇨ décision k_1 si et seulement si $y < \lambda$

Etude de l'erreur

Si la décision est k_1, la probabilité d'erreur est $P(k_2/y)$, l'erreur est minimisée pour tout y, donc en moyenne également.

Règle bayésienne = minimisation de l'erreur conditionnelle

II Stratégie bayésienne

Objectif : modéliser statistiquement le problème de décision, en prenant en compte toutes les connaissances a priori.

II.1 Ensemble des classes

$K = \{k_1, k_2, ..., k_{N_K}\}$ espace discret probabilisé

$$\{P(k_i), i = 1, N_K\}$$
$$\sum_{i=1}^{N_K} P(k_i) = 1$$

exemple : les caractères manuscrits, les locuteurs, les mots de la langue

II.2 Ensemble des observations

$y \in Y$, Y est un ensemble mesurable discret ou continu de type R^d

Exemples :

- ⇨ représentation spectrale d'un morceau de signal
- ⇨ tableau bidimensionnel de pixels

Définition d'une distribution de probabilité conditionnelle sur Y, pour chaque classe k_i :

$$p(y/ k_i).$$

Définition d'une distribution de probabilité (globale) sur Y :

$$p(y) = \sum_{i=1}^{N_K} p(y / k_i) P(k_i)$$

II.3 Ensemble des décisions et fonction de coût

D : ensemble des décisions

$D = K$

$D = K \cup \{rejet\}$

$D = KxK$

...

Définition d'une fonction de coût appelé fonction de perte:

$\lambda : KxD \dashrightarrow R^+, \lambda(k, j)$ désigne le coût d'une décision j alors qu'il s'agit d'une forme de classe k.

II. 4 Règle de décision (cas déterministe)

Objectif : à partir d'une observation y prendre une décision j (élément de D)

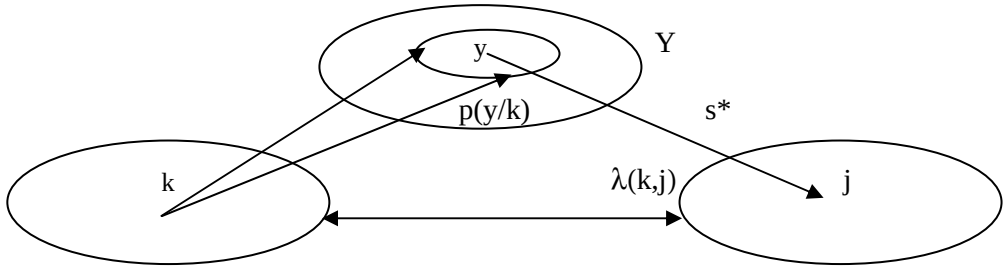
Définition (ou proposition) du risque conditionnel : le risque de prendre une décision j sachant que l'on observe y est la moyenne des coûts entre j et les différents choix possibles d'appartenance à une classe k pondérés par la probabilité a posteriori de cette classe, y étant connu.

$$R(j / y) = \sum_{i=1}^{N_K} \lambda(k_i, j) P(k_i / y)$$

Stratégie bayésienne : la décision prise pour chaque observation y est celle qui minimise le risque conditionnel.

$$s^*(y) = \arg\min_j R(j / y)$$

Schéma récapitualif :



Notations :

- p désigne la probabilité sur *Y* , un ensemble mesurable discret ou continu de type R^d , donc p est une probabilité (cas discret) ou une densité (cas continu)
- P est une probabilité sur l'ensemble K, ensemble discret. P est une probabilité discrète.

III Cas les plus usuels

III. 1 D = K

$$\lambda(k, j) = 1 - \delta(k, j)$$

δ est la fonction de Kronecker, vaut 0 (k=j) ou 1 sinon.

Propositions :

1) s* correspond à la règle du **maximum de vraisemblance**
$$s^*(y) = \arg\max_k P(k / y)$$
2) s* minimise la probabilité d'erreurs en moyenne
3) $s^*(y) = \arg\max_k P(k)\, p(y / k)$

III. 2 Cas de deux classes

Règle de décision dite de rapport de vraisemblance :

La décision est k_1 si et seulement si : $\dfrac{p(y / k_1)}{p(y / k_2)} \geq \gamma$

Exemples : problème de vérification « un contre tous »

III. 3 Cas du rejet

$$D = K \cup \{rejet\}$$

Règle de décision basée sur un seuil de rejet

s*(y) = *rejet* si et seulement si $a \succ P(k / y), \forall k \in K$

sinon

il existe une classe k' telle que $P(k' / y) \geq a$
et $s^*(y) = \arg\max_k P(k / y)$

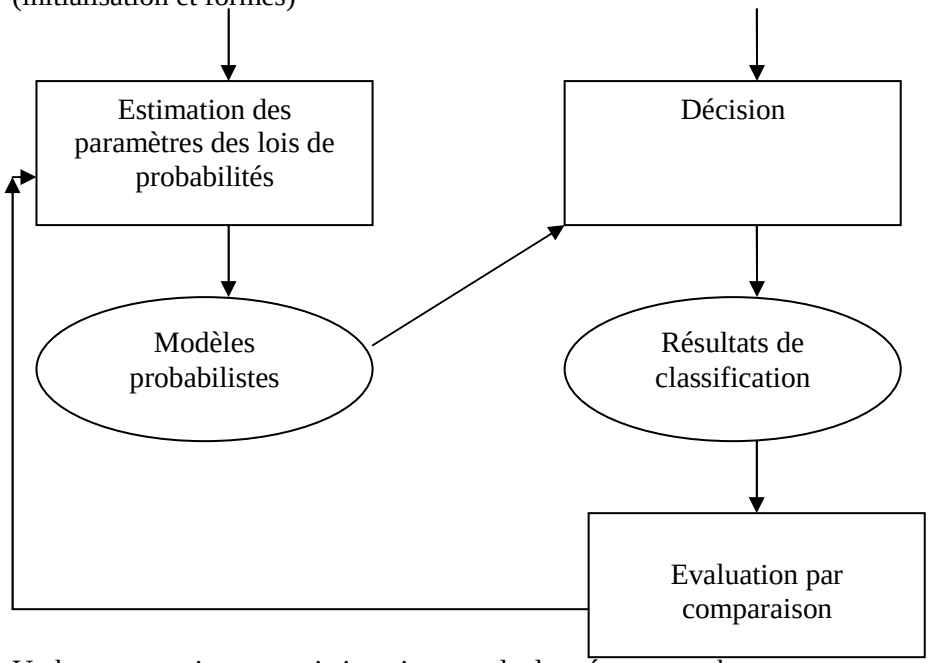
Interprétation : il y a rejet dès lors qu'aucune vraisemblance n'est suffisamment élevée pour avoir confiance en la décision

V Mise en œuvre et Apprentissage

Définition de données d'apprentissage et de données de développement ;
les données sont étiquetées et permettent un apprentissage supervisé.
Apprentissage des distributions de probabilités sur les données d'apprentissage.
Evaluation et réglage des heuristiques sur les données de développement.
Adaptation de distributions sur des données d'adaptation en quantité plus faible que les données d'apprentissage.

Données d'apprentissage
Hypothèses sur les lois de probabilité (initialisation et formes)

Données de développement



Un bon apprentissage statistique impose de données en nombre « suffisant »

Chapitre 5 : Exemples de Méthodes paramétriques et non paramétriques en Décision bayésienne

Les probabilités des observations conditionnellement à chaque classe k, p(y/k), sont données par :

- Des lois paramétriques comme les lois de Poison, les lois gaussiennes, des mélanges de lois : les lois sont décrites par des paramètres
- Des lois non paramétriques définies au travers du seul ensemble d'apprentissage (type histogrammes)

Attention : pour être cohérent avec les chapitres précédents, il sera utilisé les notations suivantes
- un indice figurant en bas désigne un numéro d'individu
- un indice figurant en haut désigne la coordonnée de l'individu si cet individu appartient à R^d.

Exemple : y_n désigne le $n^{ième}$ individu d'un ensemble alors que y^i désigne la $i^{ième}$ coordonnée de l'individu y. Par conséquent, y^i_n désigne la $i^{ième}$ coordonnée du $n^{ième}$ individu d'un ensemble

I Cas discret paramétrique : L'ensemble des observations est N, les lois sont de type Poisson

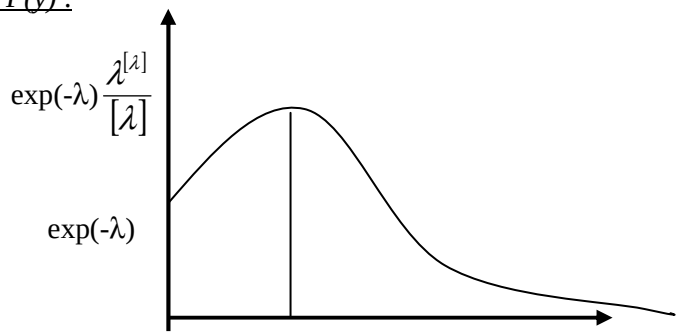
a) Définition d'une loi de Poisson

Une variable aléatoire Y discrète à valeurs dans N, suit une loi de Poisson de paramètre λ si $P(y) = \exp(-\lambda)\frac{\lambda^{y}}{y!}$. λ est positif, supérieur à 1, et différent d'un entier.

Propriété : E(Y)= λ.

Apprentissage d'une loi de Poisson Soit $\mathbf{Y} = \{y_1, y_2 ..., y_N\}, y_n \in N$, un ensemble de N réalisations de la variable Y, suivant une loi de Poisson de paramètre λ. Une estimation de λ est donnée par : $\frac{\sum_{i=1}^{N} y_n}{N}$

Allure de la loi *P(y)* :



[λ] (partie entière de λ)

b) Décision bayésienne

Lorsque les lois conditionnelles par rapport à chaque classe sont définies à partir de lois de Poisson :

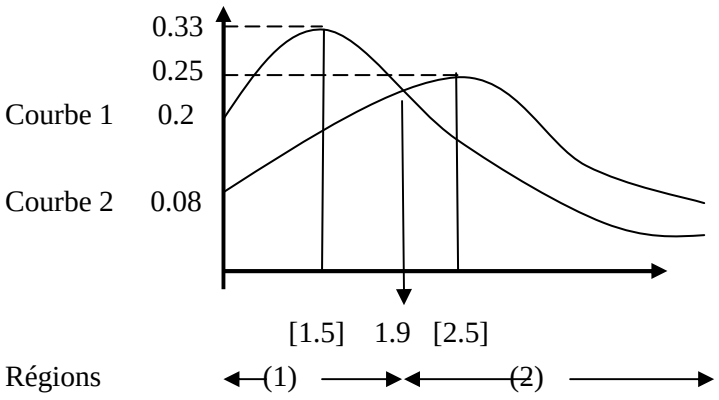
$$P(y/k_i) = \exp(-\lambda_i)\frac{\lambda_i^{\,y}}{y!},$$

λ_i sont des paramètres réels non entiers.

Proposition : dans le cas de la décision par maximum de vraisemblance,

$$s^*(y) = \arg\max_i \exp(-\lambda_i)\frac{\lambda_i^{\,y}}{y!}P(k_i)$$

Région dans le cas de deux classes (classes équiprobables) (λ_1=1.5, λ_2=2.5)



Remarque : La limite entre les deux régions se déplacera à gauche (resp. à droite) si la probabilté a priori de la classe (1) diminue (resp. augmente).

II Cas réel paramétrique : L'ensemble des observations est et les lois sont de type lois gaussiennes

a) Définition d'une loi gaussienne sur R^d

Soit Y le vecteur aléatoire

$$Y = \begin{pmatrix} Y^1 \\ Y^2 \\ \dots \\ Y^d \end{pmatrix}$$, Y^i est une variable aléatoire sur R.

La variable aléatoire Y suit une loi gaussienne si elle admet une densité de probabilité f(y), noté N(y,m,Σ), donnée par :

Cas d=1 :

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{(y-m)^2}{\sigma^2})$$

Cas d quelconque :

$$f(y) = \frac{1}{(\sqrt{2\pi})^d \det(\Sigma)^{1/2}} \exp(-\frac{1}{2}(y-m)'\Sigma^{-1}(y-m))$$

avec E(Y) = m, moyenne de la loi, vecteur de R^d

Σ la matrice de covariance, matrice de dimension dxd,

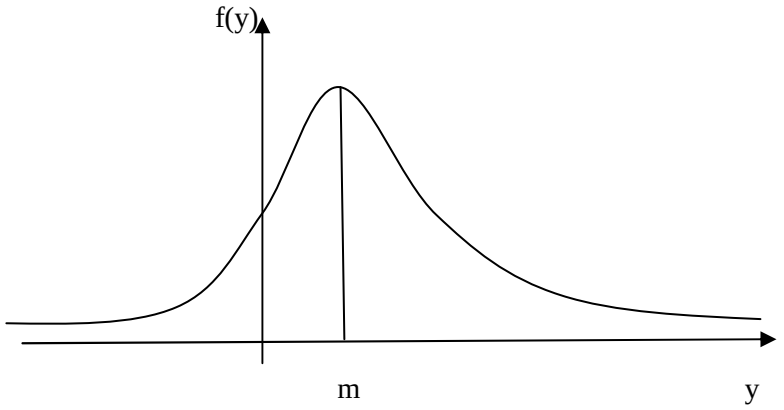
$\Sigma^{r,s} = \text{cov}(Y^r, Y^s) = E((Y^r - E(Y^r)(Y^s - E(Y^s))$

Propriété

Si les coordonnées de la variable Y sont indépendantes, la densité de probabilité dans R^d est le produit de d densité de probabilités gaussiennes définies sur R.

|| *cov(Y^r,Y^s) = 0* si *r* est différent de *s*. ||

Allure de la densité :



L'étalement de la cloche dépend de la valeur de la variance σ^2.

Courbes de niveau : f(y) = constante
Les courbes de niveau d'une densité gaussienne sont des ellipsoïdes dans R^d, définies par la distance de Mahalanobis

$$C_{y_0} = \{y, f(y) = f(y_0)\} = \{y, D_{\Sigma^{-1}}(y,m) = D_{\Sigma^{-1}}(y_0,m)\}$$
$$D_{\Sigma^{-1}}(y,m) = (y-m)'\Sigma^{-1}(y-m)$$

apprentissage d'une loi gaussienne
Soit $\mathbf{Y} = \{y_1, y_2 ..., y_N\}, y_n \in R^d$, un ensemble de N réalisations de la variable Y, suivant une loi gaussienne de paramètre m et Σ. Une estimation des paramètres dite *par maximum de vraisemblance* est donnée par:

$$\hat{m} = \frac{1}{N}\sum_{n=1}^{N} y_n, \quad \hat{\Sigma}^{r,s} = \frac{1}{N}\sum_{i=1}^{N}(y_n^r - \hat{m}^r)(y_n^s - \hat{m}^s)$$

(r = 1,…,d, s= 1,…,d)

b) Décision bayésienne

La loi des observations conditionnellement à une classe *k* est une loi gaussienne sur R^d : f(y/k) = $N(y,m_k,S_k)$

Proposition : dans le cas de la décision par maximum de vraisemblance,

$$s^*(y) = \arg\max_k N(y, m_k, \Sigma_k)P(k)$$

La **discrimination est dite quadratique**, car les équations définissant les frontières de régions sont d'ordre 2 (ellipsoïdes, paraboloïdes, hyperboloïdes)

Cas particulier de discrimination linéaire
Proposition : sous les hypothèses restrictives où $\Sigma_k = \Sigma$, matrice constante éventuellement non diagonale, pour toute classe *k*, la discrimination est linéaire. La frontière entre la classe *i* et la classe *j* est donnée par l'équation *:*

$$(y - \frac{1}{2}(m_i + m_j))'\Sigma^{-1}(m_i - m_j) = \log(p(j)) - \log(p(i))$$

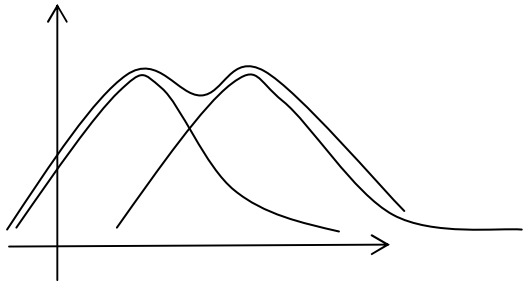
C'est un hyperplan perpendiculaire au vecteur $\Sigma^{-1}(m_i - m_j)$.
Si les classes sont équiprobables, cet hyperplan passe par le point $\frac{1}{2}(m_i + m_j)$.

Cas particulier : Si de plus $\Sigma_k = \sigma^2 I$, pour toute classe *k*, avec *I* la matrice identité de R^d, et si les classes sont équiprobables

$$s^*(y) = \arg\min_k \|y - m_k\|$$

c) Mélanges de lois gaussiennes.

Existence de plusieurs modes ou « moyennes »



$$f(y/k) = \sum_{l=1}^{L_k} \lambda_k^l N(y, m_k^l, \Sigma_k^l)$$

III Cas non paramétrique : Estimation locale d'une densité

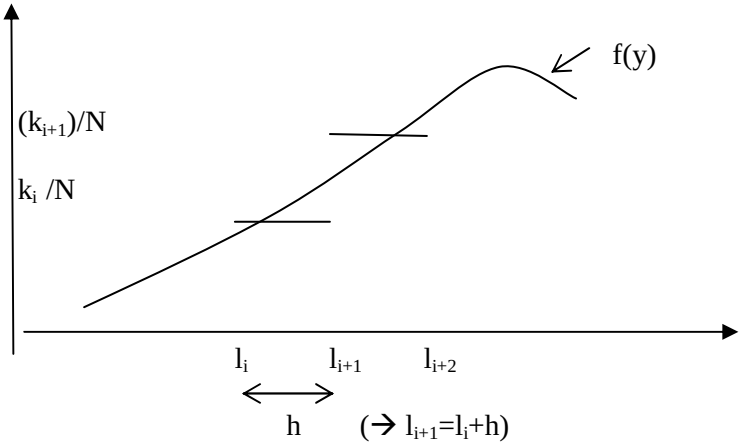
Aucune connaissance sur la loi de la variable aléatoire Y définie sur R^d, si ce n'est qu'elle admet une densité de probabilité en tout point y de R^d. Il s'agit d'estimer localement la valeur de f(y), à partir d'un ensemble d'apprentissage, $\mathbf{Y} = \{y_1, y_2 ..., y_N\}, y_n \in R^d$, donné.

a) <u>estimation statistique de la densité</u>

L'idée est basée sur la notion d'histogramme normalisé.

<u>Examen du cas d=1</u> :

Construction de l'histogramme normalisé, avec un pas de quantification égal à h, à partir de l'ensemble $\mathbf{Y}$:



k_i= (nombre d'observations de Y de valeur comprise entre l_i et l_{i+1})

$$P(l_i<y<l_{i+1}) = \int_{l_i}^{l_{i+1}} f(t)dt$$

$$k_i/N \sim f(\tilde{t}) \text{ x h}$$

<u>Théorème</u> : sous les quatre hypothèses suivantes :

h $\rightarrow$ 0

N $\rightarrow$ + infini

$k_i/N \rightarrow 0$ } $\frac{k_i}{Nh} \rightarrow f(y)$

$k_i \rightarrow$ + infini

<u>Généralisation au cas de R^d</u>

L'histogramme consiste à compter le nombre d'individus dans des hypercubes de côté h, de volume h^d, en un point y.

<u>Définition</u> : Indicatrice de l'hypercube centré en y d'arête h.

$\Phi(\frac{x-y}{h})$= 1 si $|x^i-y^i| < h/2$, pour tout *i*, *i* = 1,…,d

$\Phi(\frac{x-y}{h})$ =0 sinon.

<u>Histogramme</u> : $\sum_{n=1}^{N} \Phi(\frac{y_n - y}{h})$ est le nombre de points de Y qui appartiennent à l'hypercube de centre y, d'arête h.

Théorème (Noyaux de Parzen) : Sous les hypothèses que h tende vers 0, que N tende vers l'infini et que $N \times h^d$ tende vers l'infini, alors

$$\frac{1}{h^d N} \sum_{n=1}^{N} \Phi(\frac{y_n - y}{h}) \text{ tend vers f(y)}$$

<u>Remarque :</u> le problème consiste à fixer h pour trouver le nombre de voisins donné par la fonction Φ. Une alternative consiste à fixer le nombre de voisins de y à une valeur k et de déterminer la valeur de h_k pour qu'il y ait exactement k éléments de Y dans l'hypercube (**estimation par la règle des k plus proches voisins**), l'estimation de f(y) est alors :

$$\frac{k}{h_k^{\,d} N}$$

b) utilisation dans le cadre de la décision bayésienne

$C=\{c_1, c_2, \ldots, c_L\}$ L classses
L'ensemble des observations appartient à R^d.
Soit un ensemble d'apprentissage constitué de données étiquetées :
$\mathbf{Y} = \{(y_1, c_{i1}), (y_2, c_{i2}), \ldots, (y_N, c_{iN})\}$. La réalisation y_n est une réalisation de la classe c_{in}.

Loi a priori sur les classes : si N(c) représente le nombre de réalisations de **Y** qui appartient à la classe c, alors p(c) = N(c)/N

Loi sur **Y** relative à la classe c définie par la densité f(y/c) :

1ère solution – estimation par la règle des k plus proches voisins pour chaque classe **– k est fixé -**

pour chaque classe, on estime $h_{k,c}$ la valeur de l'arête de l'hypercube centré en y contenant k éléments de la classe c.

$$f(y/c) = \frac{k}{h_{k,c}^{\,d} N(c)}$$

La règle du maximum de vraisemblance s'écrit :

$$s^*(y) = \arg\max_c f(y/c)\,p(c)$$

$$= \arg\max_c \frac{k}{h_{k,c}^{\,d} N(c)} \frac{N(c)}{N}$$

$$= \arg\max_c \frac{k}{h_{k,c}^{\,d}}$$

$$y \in \hat{c} \Leftrightarrow h_{k,\hat{c}} \prec h_{k,c} \quad \text{pour toute classe c différente de } \hat{c}.$$

En clair, y appartient à la classe qui correspond à l'hypercube de plus petite arête qui contient k voisins.

2e solution « solution sous optimale par rapport au théorème »

L'arête de l'hypercube est fixée pour toutes les classes et est noté h, cette valeur est fixée de telle sorte que **l'hypercube centré en y contienne k éléments de Y, indépendamment de leur appartenance à une classe.**

Soit $k_{h,c}$ le nombre d'éléments de Y appartenant à cet hypercube et à la classe c ; d'après le théorème,

$$f(y/c) = \frac{k_{h,c}}{h^{\,d} N(c)}$$

La règle du maximum de vraisemblance s'écrit :

$$s^*(y) = \arg\max_c f(y/c)\,p(c)$$

$$= \arg\max_c \frac{k_{h,c}}{h^{\,d} N(c)} \frac{N(c)}{N}$$

$$= \arg\max_c \frac{k_{h,c}}{h^{\,d}}$$

$$= \arg\max_c k_{h,c}$$

$$y \in \hat{c} \Leftrightarrow k_{h,\hat{c}} \succ k_{h,c} \quad \text{pour toute classe c différente de } \hat{c}.$$

Règle de décision par les k plus proche voisins : y appartient à la classe qui est majoritairement représentée parmi les k plus proches voisins de y dans **Y**.