le cnam

Reconnaissance des formes et méthodes neuronales

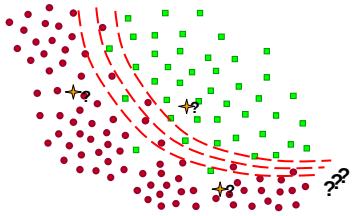
Chapitre 4 : Décision bayesienne, discrimination paramétrique, discrimination non paramétrique

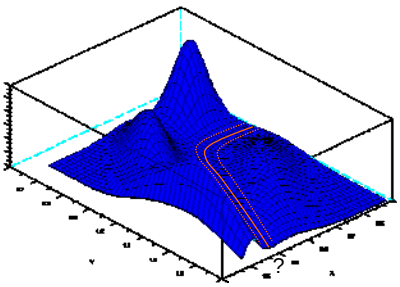
Michel Crucianu
http://cedric.cnam.fr/~crucianm/rfmn.html

le cnam

Décision

- Question (contexte : problème de classement) : sur quelle base décider d'affecter une nouvelle observation à une classe?





et si nous disposions d'un modèle probabiliste de chaque classe, ainsi que du coût des mauvaises décisions ?

le cnam

Théorie bayesienne de la décision

- Cadre probabiliste général : théorie bayesienne de la décision
- Définition du cadre :
  - Observations $\mathbf{x}$ dans un espace de description $\mathcal{X}$
  - Présence de *k* classes c_j, mutuellement exclusives, de probabilités *a priori* $P(c_j)$ supposées connues, avec $\sum_{j=1}^{k} P(c_j) = 1$
  - On considère *k*+1 décisions possibles :
    - a_j : « affectation à la classe *j* » pour $1 \leq j \leq k$
    - a_0 : « refus d'affectation »
  - Les probabilités conditionnées par l'appartenance à une classe, $p(\mathbf{x} | c_j)$ (aussi vraisemblance – *likelihood* – de c_j quand $\mathbf{x}$ est observé), sont supposées connues (notation : $l(c_j | \mathbf{x}) = p(\mathbf{x} | c_j)$)
  - Fonction de coût (ou perte) $\lambda(a_i | c_j) : \{a_0, \ldots, a_k\} \times \{c_1, \ldots, c_k\} \rightarrow \mathbb{R}^+$ quantifie la perte subie en prenant la décision a_i alors que $\mathbf{x} \in c_j$

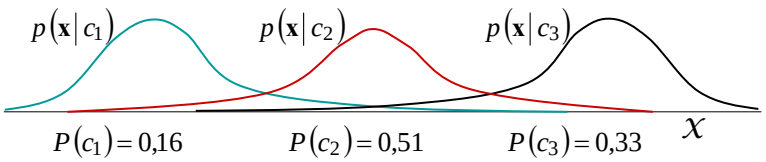
25 septembre 2013 RCP208 3

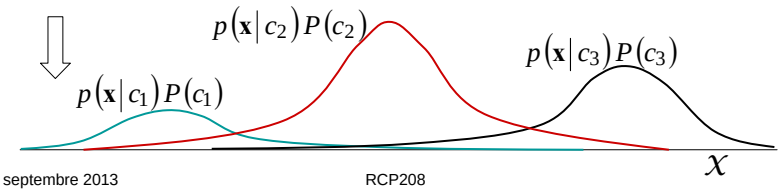
le cnam

Décision bayesienne : cas continu

- Soit $\mathcal{X} \subset \mathbb{R}^d$, et $p(\mathbf{x} | c_j)$ des densités de probabilité
- La relation de Bayes permet d'obtenir les probabilités *a posteriori*

$P(c_j | \mathbf{x}) = \dfrac{p(\mathbf{x} | c_j) P(c_j)}{p(\mathbf{x})}$, où $p(\mathbf{x}) = \sum_{j=1}^{k} p(\mathbf{x} | c_j) P(c_j)$ est l'évidence

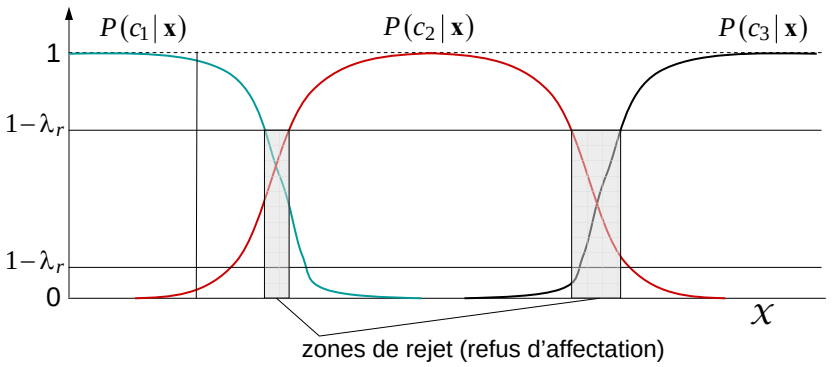




25 septembre 2013 RCP208 4

le cnam

Décision bayesienne : cas continu (2)



le cnam

Décision bayesienne : cas continu (3)

- L'espérance de la perte (ou coût, ou risque conditionnel) subie en prenant la décision a_i alors que $\mathbf{x}$ est observé sera
  $R(a_i \mid \mathbf{x}) = \sum_{j=1}^{k} \lambda(a_i \mid c_j)\, P(c_j \mid \mathbf{x}) \qquad (R(a_i \mid \mathbf{x}) \geq 0)$
- Pour une règle (fonction) de décision $a(\mathbf{x})$, $a : \mathcal{X} \to \{a_0, \ldots, a_k\}$, la perte totale (ou risque total) sera alors
  $R = \int_{\mathcal{X}} R(a(\mathbf{x}) \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$
- Minimiser R : règle de décision bayesienne pour un risque minimal
  $\mathbf{x}$ observé, calculer $R(a_i \mid \mathbf{x})$ pour tout *i* et définir $a^*(\mathbf{x}) = \arg\min_{a_i} \{R(a_i \mid \mathbf{x})\}$
- Le risque total minimal résultant, $R^* = R(a^*)$, est appelé risque bayesien et constitue la meilleure performance atteignable pour une fonction de perte λ donnée (avec les connaissances disponibles $P(c_j)$, $p(\mathbf{x} \mid c_j)$)

le cnam

Coût symétrique et règle du MAP

- Considérons la fonction de coût $\lambda(a_i \mid c_j) = \begin{cases} 0 & \text{si } i = j \\ 1 & \text{si } i \neq j,\ i \neq 0 \\ \lambda_r & \text{si } i = 0,\ 0 < \lambda_r \leq 1 \end{cases}$
  (appelée aussi coût symétrique ou 0-1)
- Alors $R(a_0 \mid \mathbf{x}) = \lambda_r \sum_{j=1}^{k} P(c_j \mid \mathbf{x}) = \lambda_r$

  et, pour $i \neq 0$, $R(a_i \mid \mathbf{x}) = \sum_{j=1}^{k} \lambda(a_i \mid c_j)\, P(c_j \mid \mathbf{x})$
  $= \sum_{j \neq i} \lambda(a_i \mid c_j)\, P(c_j \mid \mathbf{x})$
  $= 1 - P(c_i \mid \mathbf{x})$
- La règle de décision bayesienne devient

  $$a^*(\mathbf{x}) = \begin{cases} a_i & \text{si } P(c_i \mid \mathbf{x}) = \max_j P(c_j \mid \mathbf{x}) > 1 - \lambda_r \\ a_0 & \text{si } \max_j P(c_j \mid \mathbf{x}) \leq 1 - \lambda_r \end{cases}$$

  ce qui explique le nom « maximum *a posteriori* » (MAP) de cette règle

25 septembre 2013 RCP208 7

le cnam

Coût symétrique et règle du MAP (2)

- La décision $a^*(\mathbf{x}) = a_i, i \neq 0$ prise suivant cette règle est erronée si et seulement si $\mathbf{x} \in c_j,\ j \neq i$
- La probabilité d'une telle erreur est

  $$P\left(\bigcup_{j \neq i} c_j \mid \mathbf{x}\right) = \sum_{j \neq i} P(c_j \mid \mathbf{x}) = 1 - P(c_i \mid \mathbf{x})$$
- Maximisant l'*a posteriori*, la décision bayesienne minimise cette probabilité conditionnelle d'erreur pour tout $\mathbf{x}$, donc également la probabilité totale d'erreur

  $$P_E^* = \int_{\mathcal{X}} \left[1 - P\left(a^*(\mathbf{x}) \mid \mathbf{x}\right)\right] p(\mathbf{x})\, d\mathbf{x}$$

  $\rightarrow$ règle de décision à probabilité d'erreur minimum

25 septembre 2013 RCP208 8

le cnam

Critère Minimax

- Comment décider lorsque les *a priori* $P(c_1), P(c_2)$ sont inconnus ?
  - Considérer les probabilités *a priori* égales : arbitraire…
  - → Choisir la règle de décision qui minimise le risque dans le pire des cas (c'est à dire pour le pire choix valide des *a priori*) → règle de décision Minimax
- Considérons un cas avec 2 classes, sans refus d'affectation. Le risque est $R = \int_{\mathcal{X}_1} R(a_1 \mid \mathbf{x}) p(\mathbf{x})\, d\,\mathbf{x} + \int_{\mathcal{X}_2} R(a_2 \mid \mathbf{x}) p(\mathbf{x})\, d\,\mathbf{x}$
  où $\mathcal{X}_1$ est la région de décision a_1 et $\mathcal{X}_2$ la région de décision a_2
- En écrivant les risques conditionnels en fonction des probabilités *a priori* et en utilisant les conditions de normalisation pour $p(\mathbf{x} \mid c_1)$ et $p(\mathbf{x} \mid c_2)$ on obtient une expression de type $R = f_1(\mathcal{X}_1) + P(c_1)\, f_2(\mathcal{X}_1)$
- Quand $f_2(\mathcal{X}_1)$ (coefficient de $P(c_1)$) s'annule, le risque ne dépend plus des probabilités *a priori* ; la règle de décision Minimax est obtenue
- *A priori* partiellement inconnus : on trouve pour chaque *a priori* la règle de décision bayesienne, on garde celle dont le risque bayesien est maximal (⇒ risque minimum pour le pire des cas)

le cnam

Modélisation pour la discrimination

- Problème : les $p(\mathbf{x} \mid c_j)$ sont en général inconnues *a priori*
- Construction de modèles pour la discrimination :
  1. Approche par modélisation des classes (ou générative) : modéliser explicitement $p(\mathbf{x} \mid c_j)$ et se servir des $P(c_j)$ pour obtenir $P(c_j \mid \mathbf{x})$
     - Avantage : fournit les probabilités *a posteriori* des classes
     - Désavantage : très exigeante (densité des données, complexité)
  2. Approche par régression : modéliser $P(c_j \mid \mathbf{x})$ explicitement
  3. Approche discriminante : modéliser directement la frontière de décision (sans faire appel à la théorie bayesienne de la décision)
     - Avantages : moins complexe et s'accommode mieux d'une faible densité des données d'apprentissage
     - Désavantage : fournit le moins d'information (en général uniquement la frontière)

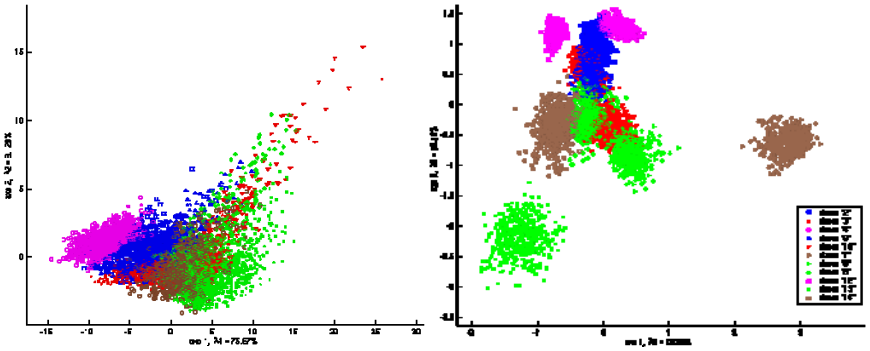
le cnam

Analyse discriminante linéaire

- Méthode d'analyse multivariée descriptive **et** décisionnelle, introduite par Fisher en 1936
- Données : ensemble d'individus décrits par un ensemble de variables quantitatives et une variable nominale « classe »
- Objectifs :
  - Étape descriptive : identifier des « facteurs discriminants » (combinaisons linéaires des variables explicatives) qui permettent de différencier au mieux les classes
  - Étape décisionnelle : sur la base des valeurs prises par les variables explicatives, décider à quelle classe affecter un nouvel individu
- Utilisations :
  - Condenser la représentation des données en conservant au mieux la séparation entre les classes
  - Prendre des décisions de classement pour de nouveaux individus à partir du sous-espace linéaire qui optimise la séparation

le cnam

Analyse discriminante linéaire



Données Textures (40 dimensions) : 2 premières composantes principales

Données Textures (40 dimensions) : 2 premières composantes discriminantes

le cnam

Les données

- Tableau de données $\mathbf{X}$ à *n* lignes (les *individus*, ensemble $\mathcal{E}$) et *p* colonnes (les variables, à valeurs dans $\mathbb{R}$)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Individu *i* :

$$\mathbf{e}_i = {}^T\begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}$$

Variable *j* :

$$\mathbf{x}_j = {}^T\begin{bmatrix} x_{1j} & x_{2j} & \cdots & x_{nj} \end{bmatrix}$$

- Les individus sont répartis en *k* classes $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k$ disjointes telles que $\mathcal{E} = \bigcup_{j=1}^{k} \mathcal{E}_j$

le cnam

Centres de gravité

- Centre de gravité (barycentre) de l'ensemble $\mathcal{E}$ d'individus :

$$g_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \qquad j = 1,2,\ldots,p$$

- Centres de gravité des classes :

$$g_j^{(q)} = \frac{1}{n_q} \sum_{\mathbf{e}_i \in \mathcal{E}_q} x_{ij} \qquad q = 1,2,\ldots,k \qquad \sum_{q=1}^{k} n_q = n$$

- Le centre de gravité de l'ensemble est également centre de gravité des centres des classes

$$g_j = \sum_{q=1}^{k} \frac{n_q}{n} g_j^{(q)}$$

le cnam

Covariances

- Covariance empirique totale entre les variables *j* et *l* :
  $s(j,l) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - g_j)(x_{il} - g_l) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, x_{il} - g_j\, g_l$
  en explicitant les classes, $s(j,l) = \frac{1}{n}\sum_{q=1}^{k}\left[\sum_{\mathbf{e}_i \in \mathcal{E}_q}(x_{ij} - g_j)(x_{il} - g_l)\right]$
- Covariance intra-classes : $d(j,l) = \frac{1}{n}\sum_{q=1}^{k}\left[\sum_{\mathbf{e}_i \in \mathcal{E}_q}\left(x_{ij} - g_j^{(q)}\right)\left(x_{il} - g_l^{(q)}\right)\right]$
- Covariance interclasses : $e(j,l) = \sum_{q=1}^{k}\frac{n_q}{n}\left(g_j^{(q)} - g_j\right)\left(g_l^{(q)} - g_l\right)$
- Relation de Huygens : $s(j,l) = d(j,l) + e(j,l)$
  ou, sous forme matricielle, $\mathbf{S} = \mathbf{D} + \mathbf{E}$

le cnam

Covariances (2)

- Avec les notations $\mathbf{R} = \frac{1}{\sqrt{n}}\begin{bmatrix}(x_{11} - g_1) & \cdots & (x_{1p} - g_p)\\ \vdots & \ddots & \vdots \\ (x_{n1} - g_1) & \cdots & (x_{np} - g_p)\end{bmatrix}$

$$\mathbf{T} = \begin{bmatrix}\sqrt{\frac{n_1}{n}}\left(g_1^{(1)} - g_1\right) & \cdots & \sqrt{\frac{n_1}{n}}\left(g_p^{(1)} - g_p\right)\\ \vdots & \ddots & \vdots \\ \sqrt{\frac{n_k}{n}}\left(g_1^{(k)} - g_1\right) & \cdots & \sqrt{\frac{n_k}{n}}\left(g_p^{(k)} - g_p\right)\end{bmatrix}$$

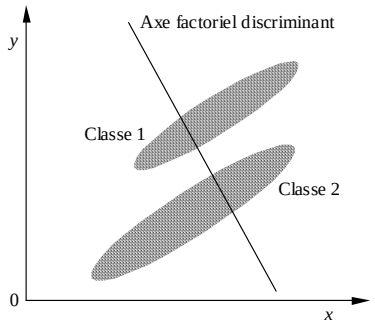
on peut écrire $\mathbf{S} = {}^T\mathbf{R}\,\mathbf{R}$ et $\mathbf{E} = {}^T\mathbf{T}\,\mathbf{T}$

le cnam

Étape descriptive

- Chercher les sous-espaces linéaires de $\mathbb{R}^p$ qui maximisent la séparation entre les projections des individus appartenant à des classes différentes

Axe discriminant pour un exemple avec 2 variables explicatives et 2 classes



le cnam

Étape descriptive : 1er axe discriminant

- Quelle droite arrive au meilleur compromis entre la maximisation de l'écart entre les projections d'individus appartenant à des classes différentes et la minimisation de l'écart entre les projections d'individus appartenant à une même classe ?
- Projection de l'individu *i* sur un vecteur $\mathbf{u}$ ($\hat{\mathbf{e}}_i$ est l'individu centré)
  $u(i) = {}^T\hat{\mathbf{e}}_i\,\mathbf{u} = \sum_{j=1}^{p}\left(x_{ij} - g_j\right)u_j$
- Variance des projections des individus sur la droite de direction $\mathbf{u}$:
  $v(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}\left[u(i) - \bar{u}\right]^2$
  ou encore (sachant que les individus sont centrés)
  $v(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}u^2(i) = \frac{1}{n}\,{}^T\mathbf{u}\left(\sum_{i=1}^{n}\hat{\mathbf{e}}_i\,{}^T\hat{\mathbf{e}}_i\right)\mathbf{u}$
- Comme $\sum_{i=1}^{n}\hat{\mathbf{e}}_i\,{}^T\hat{\mathbf{e}}_i = {}^T\mathbf{R}\,\mathbf{R} = \mathbf{S}$, on obtient $v(\mathbf{u}) = {}^T\mathbf{u}\,\mathbf{S}\,\mathbf{u}$

le cnam

Étape descriptive : 1er axe discriminant

- La relation de Huygens implique $^T\mathbf{u}\,\mathbf{S}\,\mathbf{u} = {}^T\mathbf{u}\,\mathbf{D}\,\mathbf{u} + {}^T\mathbf{u}\,\mathbf{E}\,\mathbf{u}$
- Condition que doit respecter la droite recherchée :
$$\arg\max_{\mathbf{u}} \frac{{}^T\mathbf{u}\,\mathbf{E}\,\mathbf{u}}{{}^T\mathbf{u}\,\mathbf{D}\,\mathbf{u}} = \arg\min_{\mathbf{u}} \frac{{}^T\mathbf{u}\,\mathbf{D}\,\mathbf{u}}{{}^T\mathbf{u}\,\mathbf{E}\,\mathbf{u}} = \arg\min_{\mathbf{u}} \frac{{}^T\mathbf{u}\,\mathbf{S}\,\mathbf{u}}{{}^T\mathbf{u}\,\mathbf{E}\,\mathbf{u}} = \arg\max_{\mathbf{u}} \frac{{}^T\mathbf{u}\,\mathbf{E}\,\mathbf{u}}{{}^T\mathbf{u}\,\mathbf{S}\,\mathbf{u}}$$
- De façon équivalente : $\arg\max_{\mathbf{u}} {}^T\mathbf{u}\,\mathbf{E}\,\mathbf{u}$, sous la contrainte $^T\mathbf{u}\,\mathbf{S}\,\mathbf{u} = \text{constante}$
- En utilisant les multiplicateurs de Lagrange, on obtient l'équation que $\mathbf{u}$ doit satisfaire : $\mathbf{E}\,\mathbf{u} = \lambda\,\mathbf{S}\,\mathbf{u}$ (équation de valeurs propres généralisée)
- Si la matrice de covariance totale est inversible, on arrive à $\mathbf{S}^{-1}\,\mathbf{E}\,\mathbf{u} = \lambda\,\mathbf{u}$
- On a également $\lambda = \frac{{}^T\mathbf{u}\,\mathbf{E}\,\mathbf{u}}{{}^T\mathbf{u}\,\mathbf{S}\,\mathbf{u}} = \frac{\left({}^T\mathbf{u}\,\mathbf{S}\,\mathbf{u} - {}^T\mathbf{u}\,\mathbf{D}\,\mathbf{u}\right)}{{}^T\mathbf{u}\,\mathbf{S}\,\mathbf{u}} \leq 1$

  donc la solution recherchée est le vecteur propre correspondant à la plus grande valeur propre de $\mathbf{S}^{-1}\,\mathbf{E}$

25 septembre 2013 RCP208 19

le cnam

Étape descriptive : axes suivants

- Suivant le même procédé, on obtient des axes discriminants successifs correspondant aux vecteurs propres successifs de l'équation $\mathbf{E}\,\mathbf{u} = \lambda\,\mathbf{S}\,\mathbf{u}$ (ou de la matrice $\mathbf{S}^{-1}\,\mathbf{E}$, si $\mathbf{S}$ est inversible)
- Remarques :
  - La matrice $\mathbf{S}^{-1}\,\mathbf{E}$ n'étant pas symétrique en général, les vecteurs propres correspondant à des valeurs propres différentes ne seront pas orthogonaux ! Ils doivent seulement satisfaire la contrainte $^T\mathbf{u}_i\,\mathbf{S}\,\mathbf{u}_j = {}^T\mathbf{u}_i\,\mathbf{D}\,\mathbf{u}_j = {}^T\mathbf{u}_i\,\mathbf{E}\,\mathbf{u}_j = 0$ pour $i \neq j$
  - Il y a au plus $k - 1$ valeurs propres non nulles car le rang de $\mathbf{E}$ est au plus égal à $k - 1$! Pour un problème à 2 classes il y aura donc au plus une valeur propre non nulle, donc un seul axe discriminant !

25 septembre 2013 RCP208 20

le cnam

Traitement algorithmique

- Difficultés :
  A. Résolution d'un problème de valeurs propres généralisé
  B. Singularité (déterminant = 0) ou mauvais conditionnement (λ_{max} >> >> $\lambda_{min} (\neq 0)$) de la matrice des covariances empiriques totales
- Solution commune en deux étapes :
  1. Réduction de la dimension par ACP afin d' obtenir (dans le sous-espace résultant) une matrice $\tilde{\mathbf{S}}$ non singulière et bien conditionnée
  2. Multiplication de la matrice $\mathbf{R}$ par $\sqrt{n}\,\mathbf{V}\,\mathbf{M}^{-1/2}$ (où $\mathbf{S} = \mathbf{V}\,\mathbf{M}\,^{T}\mathbf{V}$), pour arriver à un problème de valeurs propres ordinaire
     $\tilde{\mathbf{S}} = {}^{T}\tilde{\mathbf{R}}\,\tilde{\mathbf{R}} = {}^{T}\left(\sqrt{n}\,\mathbf{R}\,\mathbf{V}\,\mathbf{M}^{-1/2}\right)\left(\sqrt{n}\,\mathbf{R}\,\mathbf{V}\,\mathbf{M}^{-1/2}\right) = n\,\mathbf{M}^{-1/2}\,{}^{T}\mathbf{V}\,\mathbf{S}\,\mathbf{V}\,\mathbf{M}^{-1/2} = n\,\mathbf{I}$
     - La nouvelle équation à résoudre sera $\tilde{\mathbf{E}}\,\mathbf{u} = \lambda\,\tilde{\mathbf{S}}\,\mathbf{u}$ (donc $\tilde{\mathbf{E}}\,\mathbf{u} = \lambda\,\mathbf{u}$)
- Si la réduction élimine trop d'information discriminante, une régularisation sera préférée (par ex., ajout de $\alpha\,\mathbf{I}_p$ à $\mathbf{S}$, $\mathbf{I}_p$ étant la matrice identité d'ordre *p* et α le coefficient de régularisation)

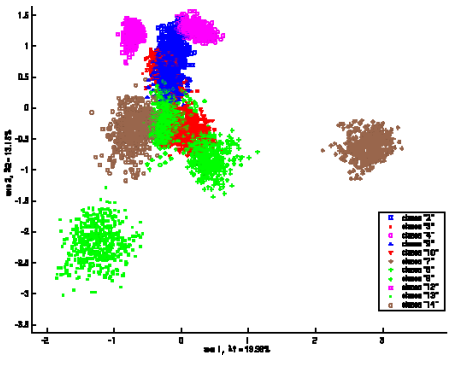
le cnam

Un exemple

- Textures (exemple déjà utilisé pour l'ACP)
- Individus (anonymes, 5500 au total) : 500 pixels pour chacune des **11** micro-textures différentes
- 40 variables : moments statistiques modifiés d'ordre 4, déterminés pour 4 orientations différentes (0°, 45°, 90°, 135°), tenant compte des relations avec les voisins d'ordres 1 et 2

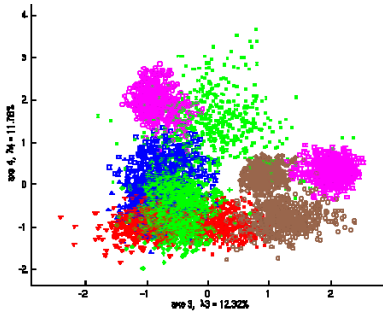
(Données obtenues et mises à disposition par le Laboratoire de Traitement d'Image et de Reconnaissance de Formes (LTIRF) de l'Institut National Polytechnique de Grenoble (INPG) dans le cadre du projet ESPRIT III ELENA (No. 6891) et du groupe de travail ESPRIT ATHOS (No. 6620))

Projections sur les 2 premiers axes discriminants

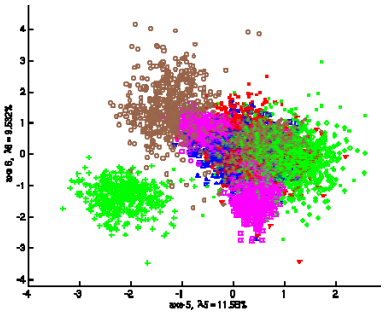


le cnam

Exemple : axes discriminants suivants



Projections sur les axes 3 et 4



Projections sur les axes 5 et 6

25 septembre 2013 RCP208 23

le cnam

Étape décisionnelle

- Classer un nouvel individu dans une des classes sur la base des valeurs prises par les variables explicatives
- L'étape descriptive a permis de conserver le sous-espace linéaire le plus discriminant
- Approche géométrique simple : le nouvel individu **a** est affecté à la classe dont il est le plus « proche »
- Évaluation de la proximité entre un individu et une classe :
  - Distance entre l'individu et le centre de gravité de la classe :
    - Métrique de Mahalanobis unique ⇒ frontières linéaires
    - Métrique de Mahalanobis adaptée à la classe ⇒ frontières quadratiques
  - Distances entre l'individu et un ensemble de points de la classe ⇒ frontières en général non linéaires

25 septembre 2013 RCP208 24

le cnam

Discrimination linéaire

- La proximité entre un nouvel individu $\mathbf{a}$ et une classe (quelle qu'elle soit) est donnée par une même distance de Mahalanobis au centre de gravité $\mathbf{g}^{(q)}$ de la classe :
  $d^2\left(\mathbf{a},\mathbf{g}^{(q)}\right) = {}^T\left(\mathbf{a}-\mathbf{g}^{(q)}\right)\mathbf{S}^{-1}\left(\mathbf{a}-\mathbf{g}^{(q)}\right)$
- On a $d^2\left(\mathbf{a},\mathbf{g}^{(q)}\right) = {}^T\mathbf{a}\,\mathbf{S}^{-1}\,\mathbf{a} - {}^T\mathbf{g}^{(q)}\,\mathbf{S}^{-1}\,\mathbf{a} - {}^T\mathbf{a}\,\mathbf{S}^{-1}\,\mathbf{g}^{(q)} + {}^T\mathbf{g}^{(q)}\,\mathbf{S}^{-1}\,\mathbf{g}^{(q)}$ et, ${}^T\mathbf{a}\,\mathbf{S}^{-1}\,\mathbf{a}$ ne dépendant pas de *q*, lors de la comparaison des distances $d^2\left(\mathbf{a},\mathbf{g}^{(q)}\right)$ on tiendra compte uniquement des fonctions linéaires discriminantes (linéaires en $\mathbf{a}$)
  $v_q(\mathbf{a}) = {}^T\mathbf{g}^{(q)}\,\mathbf{S}^{-1}\left(\mathbf{g}^{(q)} - 2\,\mathbf{a}\right)$ (ou $v_q(\mathbf{a}) = d^2\left(\mathbf{a},\mathbf{g}^{(q)}\right) - {}^T\mathbf{a}\,\mathbf{S}^{-1}\,\mathbf{a}$)
- L'individu $\mathbf{a}$ sera alors affecté à la classe pour laquelle $v_q(\mathbf{a})$ est minimale

le cnam

Discrimination linéaire entre 2 classes

- Règle de décision : affecter $\mathbf{a}$ à la classe 1 si $v_2(\mathbf{a}) - v_1(\mathbf{a}) > 0$, sinon à la classe 2
- En développant on obtient
  $v_2(\mathbf{a}) - v_1(\mathbf{a}) = 2\,{}^T\left(\mathbf{g}^{(1)} - \mathbf{g}^{(2)}\right)\mathbf{S}^{-1}\,\mathbf{a} - {}^T\left(\mathbf{g}^{(1)} - \mathbf{g}^{(2)}\right)\mathbf{S}^{-1}\left(\mathbf{g}^{(1)} + \mathbf{g}^{(2)}\right)$
- L'expression ${}^T\left(\mathbf{g}^{(1)} - \mathbf{g}^{(2)}\right)\mathbf{S}^{-1}\,\mathbf{a}$ est la fonction linéaire discriminante de Fisher
- La frontière de discrimination, donnée par l'équation $v_2(\mathbf{a}) - v_1(\mathbf{a}) = 0$ sera linéaire en $\mathbf{a}$ (suivant la dimension : droite, plan, hyperplan)
- La règle de décision devient : affecter $\mathbf{a}$ à la classe 1 si
  ${}^T\left(\mathbf{g}^{(1)} - \mathbf{g}^{(2)}\right)\mathbf{S}^{-1}\left(\mathbf{a} - \frac{\mathbf{g}^{(1)} + \mathbf{g}^{(2)}}{2}\right) > 0$
  sinon à la classe 2

le cnam

Choix du nombre d'axes significatifs

- Pouvoir discriminant de l'axe factoriel *i* : λ_i (car $\sum_{j=1}^{k-1} \lambda_j = 1$)
- Tests statistiques (hypothèses : lois normales, matrices de variances-covariances identiques pour les classes) :
  - Égalité à 0 de la *q*-ème valeur propre (test de Rao) : $(n-k)\lambda_q$ suit une loi $\chi^2(p+k-2q)$
  - Signification de l'apport des axes au-delà de *q* (test du Lambda de Wilks) : $-[n-(p+k)/2-1]\ln(L^*)$ (avec $L^* = \prod_{i=q+1}^{k-1} \lambda_i$, produit de valeurs propres de $\mathbf{S}^{-1}\mathbf{E}$) suit une loi $\chi^2((p-q)(k-q-1))$
  - Signification de l'apport de l'axe *q*+1 : $V_{q+1} - V_q$ suit une loi $\chi^2(k-1)$ (où $V = \sum_{i=1}^{k} n_i^T (\mathbf{g}^{(i)} - \mathbf{g}) \mathbf{E}^{-1} (\mathbf{g}^{(i)} - \mathbf{g})$)
- Sans ces hypothèses, faire appel à l'étape décisionnelle via la méthode de l'échantillon test : on met des données de côté (l'échantillon test), on calcule les fonctions discriminantes sur les données restantes pour différents nombres d'axes, on évalue le classement sur l'échantillon test

25 septembre 2013 RCP208 27

le cnam

Discrimination quadratique

- Les matrices de variances-covariances des différentes classes sont rarement identiques entre elles et à la matrice de variances-covariances totale, donc l'emploi d'une métrique différente pour chaque classe est préférable
- Sebestyen (1962) : utiliser comme métrique par rapport à la classe *q* la métrique qui minimise l'inertie intra-classe de la classe *q* (sous une contrainte de normalisation afin d'exclure la solution triviale)
- En utilisant la métrique de matrice $\mathbf{Q}_q$, le carré de la distance moyenne entre les individus de la classe *q* est
$$D_q^2 = \frac{1}{n_q(n_q-1)} \sum_{\mathbf{e}_i \in \mathcal{E}_q} \sum_{\mathbf{e}_j \in \mathcal{E}_q}^{T} (\mathbf{e}_i - \mathbf{e}_j) \mathbf{Q}_q (\mathbf{e}_i - \mathbf{e}_j)$$
- La métrique recherchée doit donc minimiser D_q^2 sous la contrainte de normalisation $\det(\mathbf{Q}_q) = 1$

25 septembre 2013 RCP208 28

le cnam

Discrimination quadratique (2)

- On montre que la métrique recherchée est définie par la matrice
  $\mathbf{Q}_q = \left[\det \mathbf{S}^{(q)}\right]^{\frac{1}{p}} \left[\mathbf{S}^{(q)}\right]^{-1}$
  où $\mathbf{S}^{(q)}$ est la matrice des covariances empiriques de la classe *q*
- Dans le cas de deux classes, la frontière de discrimination sera définie par $d^2_{(1)}\left(\mathbf{x}, \mathbf{g}^{(1)}\right) = d^2_{(2)}\left(\mathbf{x}, \mathbf{g}^{(2)}\right)$ ou, en développant,

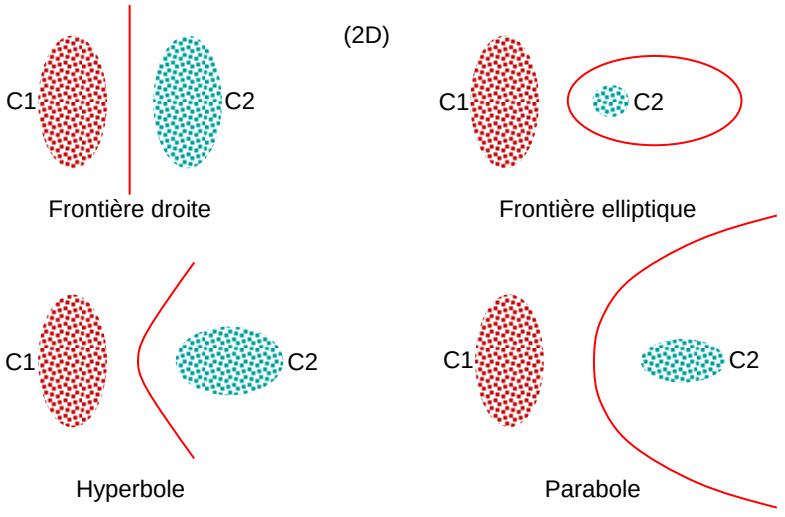
$$\left[\det \mathbf{S}^{(1)}\right]^{\frac{1}{p}} {}^T\left(\mathbf{x}-\mathbf{g}^{(1)}\right)\left[\mathbf{S}^{(1)}\right]^{-1}\left(\mathbf{x}-\mathbf{g}^{(1)}\right) = \left[\det \mathbf{S}^{(2)}\right]^{\frac{1}{p}} {}^T\left(\mathbf{x}-\mathbf{g}^{(2)}\right)\left[\mathbf{S}^{(2)}\right]^{-1}\left(\mathbf{x}-\mathbf{g}^{(2)}\right)$$

- Quand $\mathbf{S}^{(1)} \neq \mathbf{S}^{(2)}$ on obtient une équation de second degré pour les composantes du vecteur $\mathbf{x}$, raison pour laquelle on parle de discrimination quadratique

le cnam

Discrimination quadratique : exemples

(2D)



le cnam

Relation avec la décision bayesienne

- Considérons le cas de 2 classes de probabilités *a priori* $P(c_1)$, $P(c_2)$ et avec

$$p(\mathbf{x}|c_i)=\frac{1}{(2\pi)^{p/2}\left|\mathbf{\Sigma}^{(i)}\right|^{1/2}}\exp\left[-\frac{1}{2}{}^{T}\left(\mathbf{x}-\boldsymbol{\mu}^{(i)}\right)\left[\mathbf{\Sigma}^{(i)}\right]^{-1}\left(\mathbf{x}-\boldsymbol{\mu}^{(i)}\right)\right] \qquad i=1,2$$

où $\boldsymbol{\mu}^{(i)}$ est l'espérance et $\mathbf{\Sigma}^{(i)}$ la matrice de variances-covariances

- Avec un coût symétrique et sans refus d'affectation, la frontière issue de la décision bayesienne est donnée par $P(c_1|\mathbf{x})=P(c_2|\mathbf{x})$, ce qui revient à

$${}^{T}\left(\mathbf{x}-\boldsymbol{\mu}^{(1)}\right)\left[\mathbf{\Sigma}^{(1)}\right]^{-1}\left(\mathbf{x}-\boldsymbol{\mu}^{(1)}\right)+\ln\frac{\left|\mathbf{\Sigma}^{(1)}\right|}{\left|\mathbf{\Sigma}^{(2)}\right|}-\ln\frac{P(c_1)}{P(c_2)}={}^{T}\left(\mathbf{x}-\boldsymbol{\mu}^{(2)}\right)\left[\mathbf{\Sigma}^{(2)}\right]^{-1}\left(\mathbf{x}-\boldsymbol{\mu}^{(2)}\right)$$

ne dépend pas de $\mathbf{x}$
donc frontière quadratique

le cnam

Analyse discriminante : mise en œuvre

Logiciel qui implémente l'AD : *Utilisateur de l'AD* :

Introduction des données étiquetées

ACP centrée ou régularisation de $\mathbf{S}$
→ réduction à un problème de valeurs et vecteurs propres ordinaire

Valeurs et vecteurs propres de $\tilde{\mathbf{S}}^{-1}\mathbf{E}$

Choix nombre d'axes discriminants (utilisant éventuellement des données étiquetées non encore exploitées)

} Étape descriptive

Introduction données non étiquetées

Affectation de chaque donnée non étiquetée à une classe (par calcul des distances selon le cas) :
- discrimination linéaire ou
- discrimination quadratique

} Étape décisionnelle

le cnam

Discrimination non paramétrique

- Méthodes paramétriques : hypothèses sur les classes ou sur la frontière de discrimination (appartenance à une famille paramétrée)
- Méthodes non paramétriques : absence d'hypothèses sur les classes ou sur la frontière de discrimination
- Méthodes populaires de discrimination non paramétrique :
  - Estimation de densité par des noyaux (Parzen) : $p(\mathbf{x} | c_j)$ est modélisée pour chaque classe par un mélange de noyaux centrés sur les exemples d'apprentissage, ensuite on rentre dans le cadre de la décision bayesienne (décision sur la base des $P(c_j | \mathbf{x})$)
  - *k* plus proches voisins (*k*ppv) : aucun modèle n'est construit pour les classes ou pour la frontière de décision, la décision de classement est prise pour chaque nouvel individu en examinant ses *k* plus proches voisins

25 septembre 2013 RCP208 33

le cnam

Méthode des *k* plus proches voisins

(*k nearest neighbours*, *knn*)

- *n* observations $D_n = \{\mathbf{x}_i\}_{1 \le i \le n}$ dans un espace métrique $\mathcal{M}$
- Chaque observation appartient à une des *c* classes mutuellement exclusives
- Affectation d'une nouvelle observation $\mathbf{a}$ à une des classes :
  1. Pour *k* fixé, recherche des *k* plus proches voisins (suivant la métrique de $\mathcal{M}$), avec choix aléatoire en cas d'égalité entre distances
  2. Affectation de l'observation à la classe la plus représentée parmi ces *k* plus proches voisins, avec choix aléatoire si égalité entre classes
  - Possibilités de rejet (refus d'affectation) :
    1. Rejet de non représentativité : toutes les distances entre $\mathbf{a}$ et ses *k* plus proches voisins sont supérieures à un seuil (par exemple $\bar{d}_{kppv} + 2\sigma$)
    2. Rejet d'ambiguïté : aucune classe ne dépasse un seuil de représentation (par exemple 66%) parmi les *k* (> 1) plus proches voisins

25 septembre 2013 RCP208 34

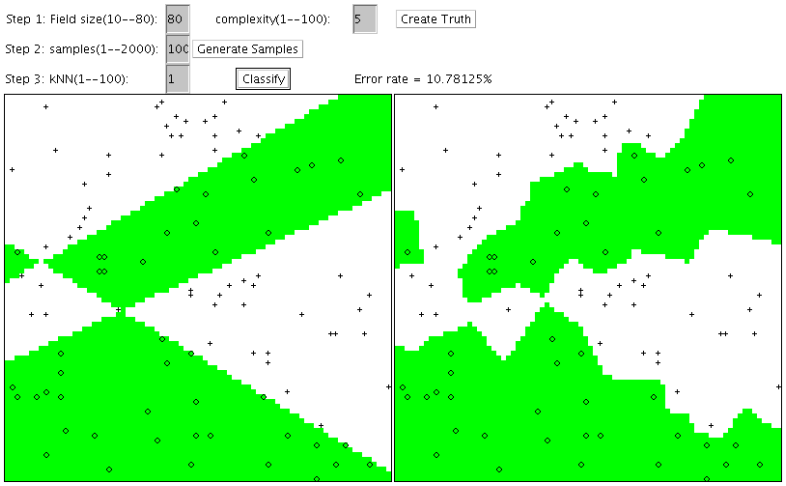
le cnam

K plus proches voisins

- Aucun modèle des classes ou de la frontière de discrimination n'est déterminé avant de prendre une décision !
- Si l'espace métrique de description $\mathcal{M}$ est $\mathbb{R}^p$ muni d'une métrique, on peut montrer que si la taille de l'échantillon $n \to \infty$ alors $P_E^*(\mathbf{x}) \leq \ldots \leq P_{E\,kppv}(\mathbf{x}) \leq P_{E\,(k-1)ppv}(\mathbf{x}) \leq \ldots \leq P_{E\,1ppv}(\mathbf{x}) \leq 2\,P_E^*(\mathbf{x})$ où $P_E^*(\mathbf{x})$ est la probabilité d'erreur (pour $\mathbf{x}$) obtenue par la règle de décision de Bayes et $P_{E\,kppv}(\mathbf{x})$ la probabilité d'erreur obtenue par la règle des *k* plus proches voisins (cas sans rejet)
- Le comportement avec un échantillon D_n limité peut être éloigné de ce comportement asymptotique !

le cnam

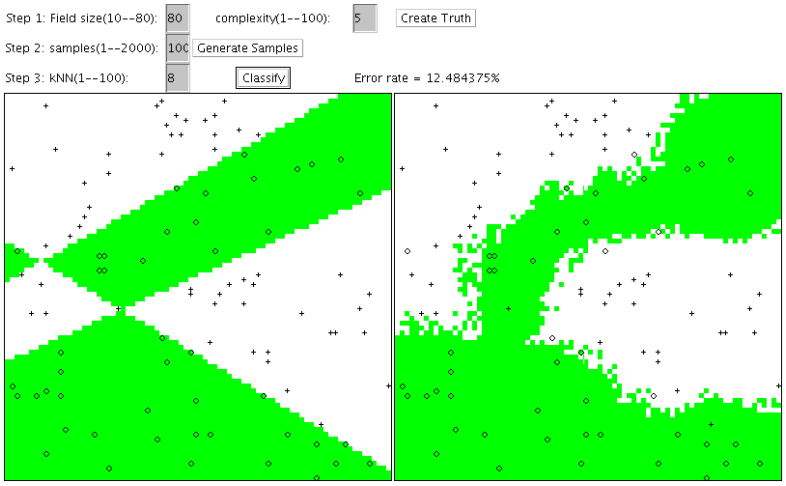
K plus proches voisins : exemple



(issu de http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html)

le cnam

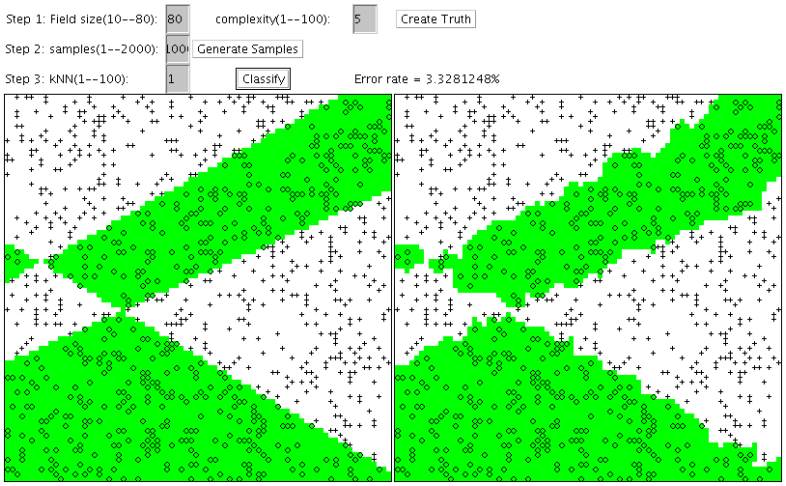
K plus proches voisins : exemple (2)



(issu de http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html)

le cnam

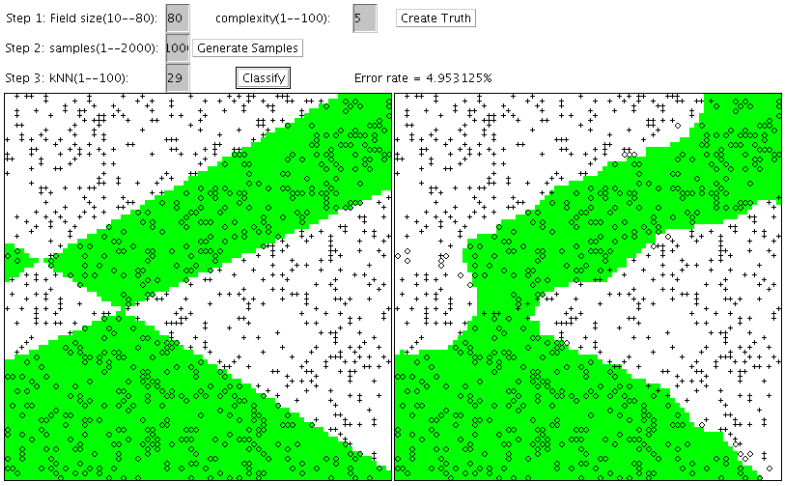
K plus proches voisins : exemple (3)



(issu de http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html)

le cnam

K plus proches voisins : exemple (4)



(issu de http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html)

le cnam

Recherche des *k* plus proches voisins

- Pour déterminer les *k*ppv d'une nouvelle observation, une implémentation basique calculera *n* distances (entre la nouvelle observation et toutes les observations pour lesquelles la classe est connue) $\Rightarrow$ prise très lente de décision pour *n* élevé !
- Quelques solutions :
  - Approximative : nettoyage et condensation : le nettoyage élimine les observations isolées, la condensation retire les observations situées à l'intérieur des classes et loin de la frontière
  - Exactes, basées sur une *indexation* préalable des observations connues et permettant de réduire la complexité à $O(\log n)$:
    - $\mathcal{M}$ est un espace vectoriel : SR-*tree*, etc.
    - La structure métrique est la seule structure connue de $\mathcal{M}$: M-*tree*, etc.