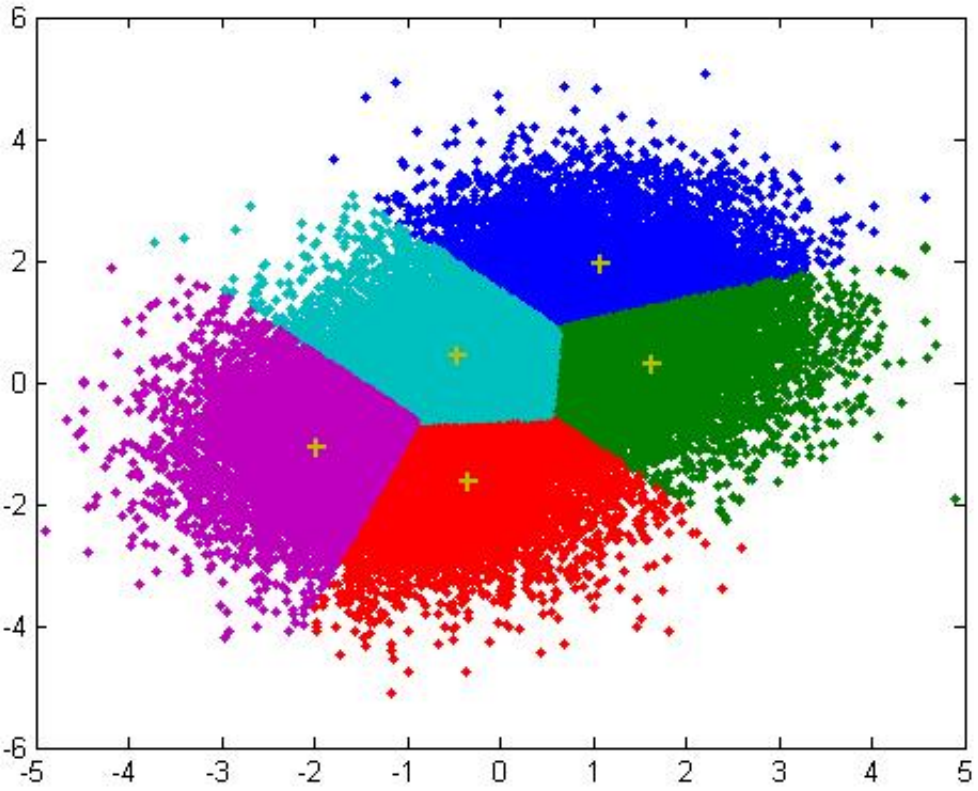
UNIVERSITE DE LORRAINE
Ecole Nationale Supérieure d'Electricité et de Mécanique

Eléments de classification de données

Didier MAQUIN

6 octobre 2020



AVERTISSEMENT

Ces notes de cours ne sont pas originales. Elles ne sont que la compilation de différentes sources citées en référence et légèrement amendées.

Table des matières

1 Analyse discriminante

1.1 Les iris de Fisher

Les données dites "iris de Fisher" sont issues d'une étude du botaniste Anderson [1] et ont été utilisées en 1937 par le célèbre statisticien Sir Ronald Fisher [2] pour démontrer la pertinence de ses méthodes, dites d'analyse discriminante. Elles sont constituées de 150 mesures faites sur 3 variétés de fleurs d'iris : *setosa*, *versicolor*, *virginica*. De manière précise, 4 mesures sont effectuées sur chaque fleur : largeur et longueur du sépale, largeur et longueur du pétale. Fisher a cherché à identifier les caractères ou les combinaisons de caractères qui permettent de distinguer au mieux les espèces d'iris. Il s'agit donc des quantités qui discriminent le plus les espèces, d'où le terme générique utilisé d'analyse discriminante.

Iris *Setosa*

Iris *Versicolor*

Iris *Virginica*

De façon générale, deux types de questions se posent naturellement :

— **Analyse discriminante descriptive**. Il s'agit ici, comme lorsqu'on réalise une ACP, de représenter les données dans un espace ad hoc, qui permette de bien mettre en évidence les variables liées à l'espèce de la fleur. En d'autres termes, les techniques qui seront mises sur pied vont chercher à répondre à la question : quelles variables, quels groupes de variables, quels sous-espaces discriminent-ils au mieux les 3 espèces d'iris ?
On retrouve souvent ce type de problème dans des applications médicales, financières ou socio-économiques.

— **Analyse discriminante décisionnelle**. On cherche ici à affecter une nouvelle fleur à une espèce en connaissant les valeurs des 4 variables quantitatives qui la décrivent. On est passé d'un objectif de description à un objectif de prévision ; c'est pourquoi des concepts probabilistes peuvent être utilisés pour traiter ces questions. Les domaines d'application de l'analyse discriminante décisionnelle sont tellement vastes qu'il serait illusoire d'en faire un recensement. Notons en premier lieu la théorie de la reconnaissance des formes qui a pour objectif de reconnaître des objets.

Examinons ces données en les représentant graphiquement (figure 1). On observe que les variables relatives aux pétales semblent davantage discriminer les espèces que les données de sépales et que l'espèce *setosa* semble facile à distinguer des deux autres. Par contre, *versicolor* et *virginica* paraissent assez mêlées.

1.2 Formalisation de l'analyse

On considère une population de n individus indexés par $i, 1 \leq i \leq n$, chaque individu numéro i étant de poids p_i. Ces individus sont caractérisés par deux types de variables :

1. E. Anderson. The irises of the Gaspé peninsula, *Bulletin of the American Iris Society*, 59, p. 2-5, 1935.
2. R.A. Fisher. The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7, p. 179-188, 1936.

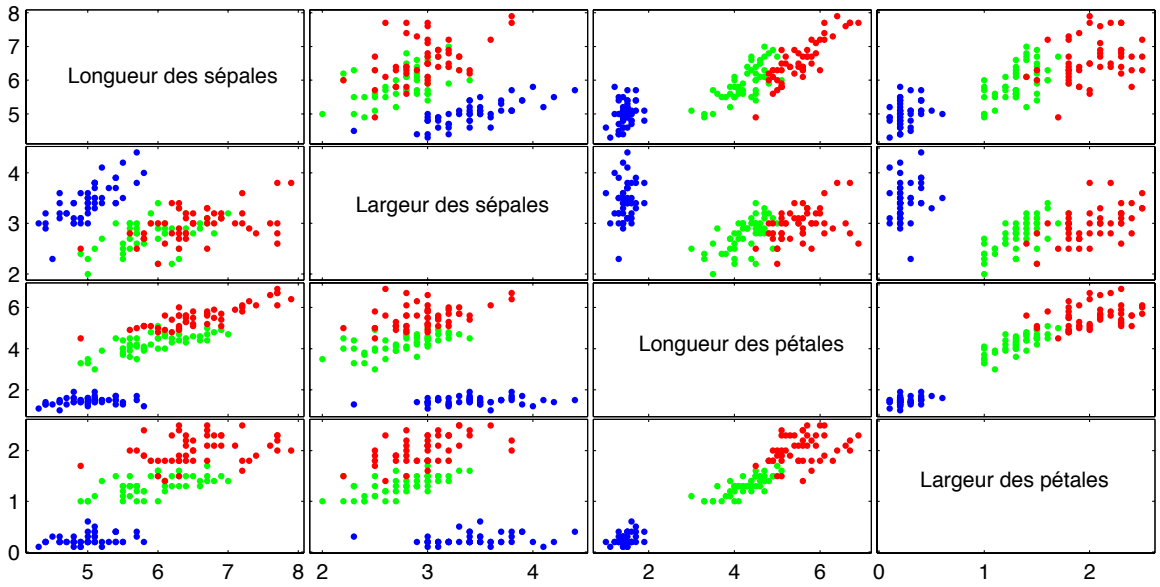


FIGURE 1 – *setosa* : bleu, *versicolor* : vert, *virginica* : rouge

— p variables X_j, le plus souvent quantitatives ;
— l'appartenance à un groupe qui se traduit par une variable qualitative Y possédant m modalités $y_h, 1 \leq h \leq m$

On note G_h le groupe $\{i, Y(i) = y_h\}$.

Analyse discriminante descriptive

Ici, il s'agit avant tout d'identifier les variables, les groupes ou combinaisons de variables, qui "expliquent" au mieux l'appartenance au groupe d'un individu. Calculons quelques indicateurs sur les variables de l'exemple considéré :

	Longueur sépale	Largeur sépale	Longueur pétale	Largeur pétale
Minimum	4.30	2.00	1.00	0.10
1er quartile	5.10	2.80	1.60	0.30
Médiane	5.80	3.00	4.35	1.30
Moyenne	5.84	3.06	3.76	1.20
3ème quartile	6.40	3.30	5.10	1.80
Maximum	7.90	4.40	6.90	2.50

TABLE 1 – Indicateurs statistiques

Pour trouver les variables les plus discriminantes, le tracé de boîtes à moustaches, ou box-plots, est très informative (figure 2).

Cette représentation donne les mêmes informations que la figure 1, de manière beaucoup plus lisible. On voit ainsi très simplement que les mesures concernant les pétales différencient mieux les espèces que celles qui concernent les sépales, et que toute mesure du pétale sépare très bien l'espèce *setosa* des deux autres. Il est par contre plus difficile de distinguer *versicolor* et *virginica* à partir de ces variables.

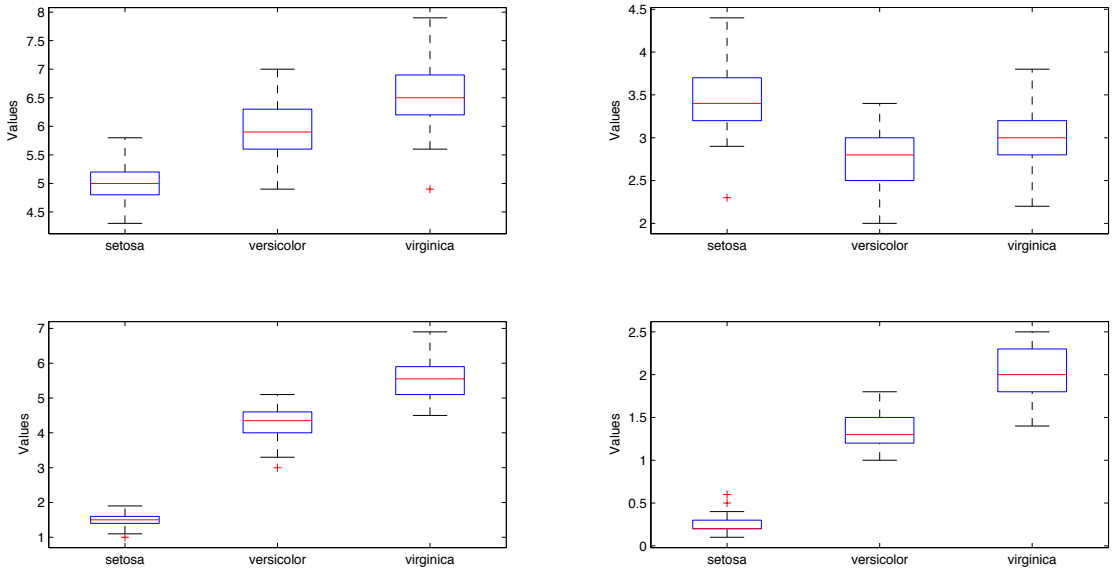


FIGURE 2 – Boîtes à moustaches des variables mesurées selon les espèces d'iris

Pouvoir discriminant et variance

Si l'on mesure la dispersion avec une variance, ou plutôt un écart-type noté σ, on peut construire le tableau 2, où par exemple σ_{set} désigne l'écart-type de la variable observée pour l'espèce *setosa*. La dernière colonne, intitulée *mean*, représente la moyenne des 3 rapports σ_{set}/σ, σ_{ver}/σ et σ_{vir}/σ.

Variable	σ	σ_{set}	σ_{ver}	σ_{vir}	σ_{set}/σ	σ_{ver}/σ	σ_{vir}/σ	*mean*
Longueur sépale	0.83	0.35	0.52	0.64	0.43	0.62	0.77	0.61
Largeur sépale	0.44	0.38	0.31	0.32	0.87	0.72	0.74	0.78
Longueur pétale	1.77	0.17	0.47	0.55	0.10	0.27	0.31	0.23
Largeur pétale	0.76	0.11	0.20	0.27	0.14	0.26	0.36	0.25

TABLE 2 – Premiers indices évaluant le pouvoir discriminant de chaque variable

On observe clairement ici que le pouvoir discriminant d'une variable est lié à la façon dont la loi de la variable considérée est concentrée du fait que l'observation de la variable est faite sur les seuls individus d'une espèce. Ainsi, plus la quantité *mean* est faible, plus la variable est utile pour discriminer les espèces. Et ceci est évidemment cohérent avec les observations précédentes faites sur les distributions des diverses variables.

Introduisons alors un modèle probabiliste ; on considère que les individus indexés par i, $(1 \leq i \leq n)$ sont affectés d'un poids p_i, ce qui définit une probabilité P sur l'ensemble des individus (ici, P est uniforme sur l'ensemble des 150 iris). Sur cet espace probabilisé sont définies les variables aléatoires quantitatives X_j $(1 \leq j \leq p)$ et la variable aléatoire qualitative Y de modalités y_h $(1 \leq h \leq m)$. Rappelons également la formule de la variance totale, appliquée à toute v.a. X_j et à la v.a. Y (théorème de la variance conditionnelle) :

$$\mathrm{Var}(X_j) = \mathrm{Esp}(\mathrm{Var}(X_j \,|\, Y)) + \mathrm{Var}(\mathrm{Esp}(X_j \,|\, Y)) \tag{1}$$

Comme la variable Y est qualitative, les espérance et variance conditionnelles s'explicitent simplement, car :

$$P(Y = y_h)\text{Esp}(X_j \,|\, Y = y_h) = \sum_{Y(i)=y_h} p_i X_j(i)$$

$$P(Y = y_h)\text{Var}(X_j \,|\, Y = y_h) = \sum_{Y(i)=y_h} p_i \left(X_j(i) - \text{Esp}(X_j \,|\, Y = y_h)\right)^2$$

Considérons la première ligne du tableau qui précède, relatif à la variable $X =$ longueur sépale. Les termes précédents s'interprètent facilement.

- σ^2 = Var(longueur sépale)
- σ^2_{set} = Var(longueur sépale $\,|\, Y = setosa$)
- σ^2_{ver} = Var(longueur sépale $\,|\, Y = versicolor$)
- σ^2_{vir} = Var(longueur sépale $\,|\, Y = virginica$)

Par suite, le premier terme du second membre de l'équation (1) n'est autre que :

$$\sigma^2_{set}P(Y = setosa) + \sigma^2_{ver}P(Y = versicolor) + \sigma^2_{vir}P(Y = virginica)$$

Soit encore, du fait que les 3 espèces comportent chacune 50 individus :

$$\text{Esp}(\text{Var}(\text{longueur sépale} \,|\, Y)) = (\sigma^2_{set} + \sigma^2_{ver} + \sigma^2_{vir})/3$$

d'où, en divisant l'identité par Var(longueur sépale) :

$$\text{Esp}(\text{Var}(\text{longueur sépale} \,|\, Y))/\text{Var}(\text{longueur sépale}) = (\sigma^2_{set}/\sigma^2 + \sigma^2_{ver}/\sigma^2 + \sigma^2_{vir}/\sigma^2)/3$$

L'indicateur *mean* introduit dans le tableau qui précède n'est rien d'autre, aux carrés près, que le quotient $\text{Esp}(\text{Var}(X \,|\, Y)/\text{Var}(X)$. Ce rapport est compris entre 0 et 1 et la variable X sera d'autant plus discriminante pour Y qu'il est petit.

Cette remarque conduit à chercher à interpréter tous les termes de l'équation (1). Le premier, $\text{Var}(X)$, est très simple ; il s'agit de la variance – dite variance totale – de la v.a. X et mesure sa dispersion.

Le second est la moyenne des variances dans chaque groupe, on l'appellera plus loin **variance intra-classes**.

Le dernier représente la variance des espérances dans chaque groupe, et nous le nommerons **variance inter-classes**. Ce terme est d'autant plus fort que les espérances sont distinctes et le rapport correspondant $\text{Var}(\text{Esp}(X \,|\, Y))/\text{Var}(X)$ est appelé indice de Sobol [3]. Il est, comme l'indice précédent, compris entre 0 et 1, mais plus il est proche de 1, plus la variable X est discriminante. Les indices de Sobol correspondants aux données des iris sont reproduites dans le tableau ci-après, et confirment les observations faites précédemment, à savoir que les deux variables concernant la taille des pétales discriminent beaucoup mieux les 3 espèces que ne le font les mesures faites sur les sépales.

3. I.M. Sobol. Sensitivity analysis for non-linear mathematical models, *Mathematical Modelling and Computer Experiments*, 1, p. 407-414, 1993.

Variable	Indice de Sobol
Longueur sépale	0.61
Largeur sépale	0.39
Longueur pétale	0.94
Largeur pétale	0.93

TABLE 3 – Indices de Sobol pour les données d'iris

Poursuivons l'étude à partir des indices numériques que nous avons commencé à mettre en évidence. Etendons tout d'abord la formule de la variance totale au cas vectoriel. Si en effet X représente le vecteur aléatoire formé des variables $X_j, 1 \leq j \leq p$ on obtient :

$$\mathrm{Var}(X) = \mathrm{Esp}(\mathrm{Var}(X \,|\, Y)]) + \mathrm{Var}(\mathrm{Esp}(X \,|\, Y)) \tag{2}$$

formule en apparence identique à la formule (1), à la différence près que dans le cas présent, les espérances s'appliquent à des vecteurs ou des matrices et que $\mathrm{Var}(X)$ est la matrice de covariance du vecteur aléatoire X. Explicitons les coefficients de chacun des termes de cette formule, en utilisant la notation $g = \mathrm{Esp}(X)$ et $g^h = \mathrm{Esp}(X \,|\, Y = y_h)$. Comme leur nom l'indique, g représente le centre de gravité du nuage complet et g_h celui des données du groupe G_h. Comme précédemment, la variable Y étant qualitative, les espérance et variance conditionnelles s'explicitent facilement. Ainsi,

$$\mathrm{Var}(X)_{j,k} = \sum_{i=1}^{n} p_i (X_j(i) - g_j)(X_k(i) - g_k)$$

$\mathrm{Var}(X)$ est la **matrice de covariance empirique totale** du nuage de points en dimension p. Elle est notée habituellement **T** (pour Totale).

$$P(Y = y_h)\mathrm{Var}(X \,|\, Y = y_h)_{j,k} = \sum_{Y(i)=y_h} p_i (X_j(i) - g_j^h)(X_k(i) - g_k^h)$$

et donc :

$$\mathrm{Esp}(\mathrm{Var}(X \,|\, Y))_{j,k} = \sum_{h=1}^{m} \sum_{Y(i)=y_h} p_i (X_j(i) - g_j^h)(X_k(i) - g_k^h)$$

soit encore en notant, pour $i \in G_h$, $p(i \,|\, Y = y_h)$ la probabilité conditionnelle $p_i \,|\, P(Y = y_h)$:

$$\mathrm{Esp}(\mathrm{Var}(X \,|\, Y))_{j,k} = \sum_{h=1}^{m} P(Y = y_h) \sum_{Y(i)=y_h} p(i \,|\, Y = y_h)(X_j(i) - g_j^h)(X_k(i) - g_k^h)$$

La quantité $\mathrm{Esp}(\mathrm{Var}(X \,|\, Y))$ s'interprète comme une moyenne, pondérée par leur probabilité, des matrices de covariance dans chaque groupe G_h. On la note **W** (pour *Within*) et on la nomme **matrice de covariance intra-classes**. Pour le dernier terme, on rappelle que le vecteur $\mathrm{Esp}(X \,|\, Y)$ prend m valeurs $g^1, \ldots, g^m$ avec comme probabilités respectives $P(Y = y_1), \ldots, P(Y = y_m)$. Son espérance vaut $g = \mathrm{Esp}(X)$ et sa matrice de covariance se réduit donc à :

$$\mathrm{Var}(\mathrm{Esp}(X \,|\, Y))_{j,k} = \sum_{h=1}^{m} P(Y = y_h)(g_j^h - g_j)(g_k^h - g_k)$$

Ce terme représente la matrice de covariance du nuage des centres de gravité $g^1, \ldots, g^m$ affectés de leur probabilité. Elle est dite **matrice de covariance inter-classes** et est notée **B** pour *Between*. En d'autres termes, la relation (2) s'interprète selon :

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \tag{3}$$

Total = Within + Between

covariance totale = covariance intra-classes + covariance inter-classes

1.3 L'analyse discriminante linéaire (LDA : *Linear Discriminant Analysis)*

La technique d'analyse discriminante linéaire peut maintenant être introduite naturellement. Nous avons vu en effet que l'indice de Sobol permettait de classer les variables quantitatives selon leur pouvoir discriminant. On peut de façon plus générale identifier parmi toutes les combinaisons linéaires de variables celle qui a l'indice de Sobol le plus important. De façon précise, soit $\beta = (\beta_1, \ldots, \beta_p)^T$ le vecteur des coefficients de la combinaison linéaire cherchée de sorte que celle-ci s'exprime selon :

$$\beta^T X = \beta_1 X_1 + \ldots + \beta_p X_p$$

On cherche β tel que l'indice de Sobol $S(\beta) = \mathrm{Var}(\mathrm{Esp}(\beta^T X \,|\, Y)) \,/\, \mathrm{Var}(\beta^T X)$ est maximal. Or,

$$\mathrm{Var}(\mathrm{Esp}(\beta^T X \,|\, Y)) = \mathrm{Var}(\beta^T \mathrm{Esp}(X \,|\, Y)) = \beta^T \mathrm{Var}(\mathrm{Esp}(X \,|\, Y))\beta = \beta^T \mathbf{B} \beta$$
$$\mathrm{Var}(\beta^T X) = \beta^T \mathrm{Var}(X)\beta = \beta^T \mathbf{T} \beta$$

On en déduit que l'indice de Sobol vaut :

$$S(\beta) = \frac{\beta^T \mathbf{B} \beta}{\beta^T \mathbf{T} \beta} \tag{4}$$

On retrouve ici une expression bien connue en analyse en composantes principales. On cherche le vecteur β maximisant $S(\beta)$. La solution est donnée par :

$$\frac{dS(\beta)}{d\beta} = \frac{2(\beta^T \mathbf{T}\beta)\mathbf{B}\beta - \mathbf{2}(\beta^{\mathbf{T}} \mathbf{B}\beta)\mathbf{T}\beta}{(\beta^T \mathbf{T}\beta)^{\mathbf{2}}} = 0 \tag{5}$$

soit :

$$\mathbf{B}\beta = \frac{(\beta^{\mathbf{T}} \mathbf{B}\beta)}{(\beta^{\mathbf{T}} \mathbf{T}\beta)} \mathbf{T}\beta \tag{6}$$

ou encore :

$$\mathbf{T^{-1}B}\beta = \frac{(\beta^{\mathbf{T}} \mathbf{B}\beta)}{(\beta^{\mathbf{T}} \mathbf{T}\beta)} \beta = \mathbf{S}(\beta)\beta \tag{7}$$

La recherche du vecteur β maximisant l'indice de Sobol se ramène donc à un calcul de vecteur propre associé à la plus grande valeur propre de la matrice $\mathbf{T^{-1}B}$.

La propriété suivante va permettre de conduire les calculs.

> La matrice $\mathbf{B}$, symétrique positive de taille $n \times n$, est de rang r inférieur à $m-1$. Il existe une base $(\beta_1, \ldots, \beta_n)$ orthonormale pour la forme quadratique associée à la matrice $\mathbf{T}$ et orthogonale pour la forme quadratique associée à la matrice $\mathbf{B}$. En d'autres termes, si $\mathbf{M}$ désigne la matrice de changement de base, on a :
>
> $$M^T \mathbf{B} M = \Lambda \quad \text{et} \quad M^T \mathbf{T} M = I \tag{8}$$
>
> où Λ est une matrice diagonale, dont les valeurs propres sont ordonnées de manière décroissante : $1 \geq \lambda_1 \geq \lambda_2 \ldots \geq \lambda_r > 0$ et $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_n = 0$

La nullité des valeurs propres à partir du rang $r+1$ n'est qu'une définition du rang de la matrice $\mathbf{B}$; le rang r est inférieur à $m-1$ du fait que la matrice $\mathbf{B}$ est la matrice de covariance des m centres de gravité des groupes. Ce nuage contenant au plus m individus est en effet inclus dans un espace affine de dimension $m-1$ (pour l'exemple des iris, les trois centres de gravité sont situés dans un plan). Enfin, la majoration par 1 est obtenue immédiatement si on rappelle que l'indice de Sobol $S(\beta)$ est inférieur à 1, conséquence immédiate de l'identité (3).

Rappelons que les termes de valeurs et vecteurs propres sont pleinement justifiés du fait que la deuxième égalité de l'identité (8) donne $M^T = (\mathbf{T}M)^{-1} = M^{-1}\mathbf{T}^{-1}$. Par suite, on a :

$$M^{-1}(T^{-1}\mathbf{B})M = \Lambda \tag{9}$$

On en déduit que les valeurs diagonales de la matrice Λ sont les valeurs propres de la matrice $T^{-1}\mathbf{B}$ et que les vecteurs propres sont ceux qui forment la matrice M.

L'ensemble des vecteurs propres $\beta_1, \ldots, \beta_n$, rangés par ordre de valeurs propres décroissantes, définissent les variables discriminantes ou variables canoniques. En termes d'interprétation probabiliste des quantités introduites, la variable aléatoire $\beta_1 X$, combinaison linéaire des variables aléatoires $X_1, \ldots, X_p$, est la variable dont la variabilité est réduite au minimum par la variable qualitative Y qui désigne le groupe. Elle résout donc le problème de maximisation de l'indice de Sobol donné par (4). Cette analyse peut être étendue aux espaces engendrés par les k ($1 \leq k \leq q$) premiers vecteurs propres. La démarche correspondante est intitulée analyse discriminante linéaire ou Linear Discriminant Analysis (LDA).

On notera la similitude entre ce qui précède et l'ACP. Celle-ci dépasse la simple analogie puisque l'on peut énoncer le résultat qui suit.

L'algorithme de la LDA est le même que celui de l'ACP du nuage des m centres de gravité des groupes, affectés de leur poids respectif, où l'espace des individus est muni de la métrique dite de Mahalanobis [4] $\mathbf{T}^{-1}$. Ce résultat est une simple conséquence de l'identité (9).

Signalons que l'analyse discriminante linéaire est quelquefois développée en remplaçant la matrice $\mathbf{T}$ par la matrice $\mathbf{W}$. C'est le cas notamment lorsque l'on effectue une analyse décisionnelle car les rapports du type de ceux donnés dans la formule (4) s'interprètent alors en termes de statistique de Fisher. Donnons très rapidement le lien, très simple, entre ces deux analyses.

On réécrit la formule (9) sous la forme :

$$\mathbf{B}M = \mathbf{T}M\Lambda$$

En remplaçant ensuite la variance totale par son expression (1), on a :

$$\mathbf{B}M(I-\Lambda) = \mathbf{W}M\Lambda$$

$$M^{-1}\mathbf{W}^{-1}\mathbf{B}M(I-\Lambda) = \Lambda$$

$$M^{-1}\mathbf{W}^{-1}\mathbf{B}M = \Lambda(I-\Lambda)^{-1}$$

Mais la matrice $\Lambda(I-\Lambda)^{-1}$ est diagonale, de terme général $\lambda_i/(1-\lambda_i)$. On déduit donc que les vecteurs propres $E_1, \ldots, E_n$ de la matrice $\mathbf{T}^{-1}\mathbf{B}$ restent vecteurs propres de la matrice $\mathbf{W}^{-1}\mathbf{B}$ de valeurs propres associées données par $\mu_i = \lambda_i/(1-\lambda_i)$

4. La métrique de Mahalanobis est cependant parfois également définie comme celle associé à la matrice $\mathbf{W}^{-1}$

1.4 Retour sur l'exemple

Appliquons les résultats précédents à l'exemple du fichier des iris. On cherche donc la combinaison linéaire $\beta^T X = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ permettant la meilleure séparation des données. Le calcul des vecteurs propres et valeurs propres de la matrice $\mathbf{T^{-1}B}$ donne :

$$\beta_1 = \begin{pmatrix} 0.2087 & 0.3862 & -0.5540 & -0.7074 \end{pmatrix}^T \text{ avec } S(\beta_1) = 0.9695 \text{ et}$$
$$\beta_2 = \begin{pmatrix} 0.0065 & 0.5866 & -0.2526 & -0.7695 \end{pmatrix}^T \text{ avec } S(\beta_2) = 0.2114.$$

On peut donc engendrer les deux variables discriminantes suivantes : $Z_1 = \beta_1^T X$ et $Z_2 = \beta_2^T X$ pour lesquelles on peut reconduire l'analyse de dispersion :

Variable	σ	σ_{set}	σ_{ver}	σ_{vir}	σ_{set}/σ	σ_{ver}/σ	σ_{vir}/σ	$mean$
Z_1	1.44	0.21	0.26	0.28	0.15	0.18	0.19	0.17
Z_2	0.31	0.25	0.24	0.32	0.81	0.78	1.05	0.88

TABLE 4 – Pouvoir discriminant des variables discriminantes

La projection des observations sur le sous-espace des variables discriminantes est représenté à la figure 3.

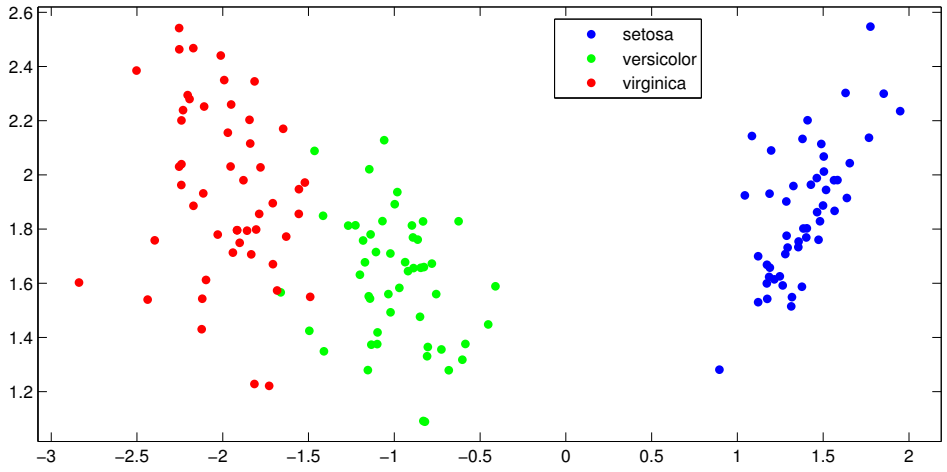


FIGURE 3 – Projection sur l'espace des variables discriminantes

On observe que le premier axe (la première variable discriminante) possède un bon pouvoir discriminant (indice de Sobol égal à 0.97) alors que le second discrimine seulement légèrement les espèces *versicolor* et *virginica*.

Comme en régression, on peut imaginer de transformer les variables initiales ou de les combiner et recommencer la procédure, i.e. identifier parmi les combinaisons linéaires de ces nouvelles variables celle dont l'indice de Sobol est le plus important. L'approche la plus commune consiste à ajouter tous les produits des variables deux à deux (distinctes ou pas).

Notons X_{aug} le vecteur :

$$X_{aug} = \begin{pmatrix} X_1 & X_2 & \ldots & X_p & X_1X_2 & X_1X_3 & \ldots & X_{p-1}X_p & X_1^2 \ldots & X_p^2 \end{pmatrix}^T$$

Ce vecteur appartient à ce que nous appellerons espace quadratique associé aux variables initiales. Appliquons de nouveau la méthodologie LDA. Les deux valeurs propres les plus grandes

(correspondant aux indices de Sobol) sont $\beta_1 = 0.9864$ et $\beta_2 = 0.7466$. L'analyse de dispersion est résumée au tableau 5.

Variable	σ	σ_{set}	σ_{ver}	σ_{vir}	σ_{set}/σ	σ_{ver}/σ	σ_{vir}/σ	$mean$
Z_1	0.435	0.064	0.054	0.0253	0.148	0.125	0.057	0.110
Z_2	0.107	0.031	0.042	0.077	0.290	0.395	0.721	0.469

TABLE 5 – Pouvoir discriminant des variables discriminantes

On remarque que l'indice de Sobol correspondant à la première variable discriminante augmente un peu ; cette augmentation est un phénomène général. Par contre, l'indice de Sobol pour la deuxième variable discriminante est beaucoup plus élevé qu'initialement, passant de 0.21 à 0.75. La projection des observations sur le sous-espace des variables discriminantes est représenté à la figure 4.

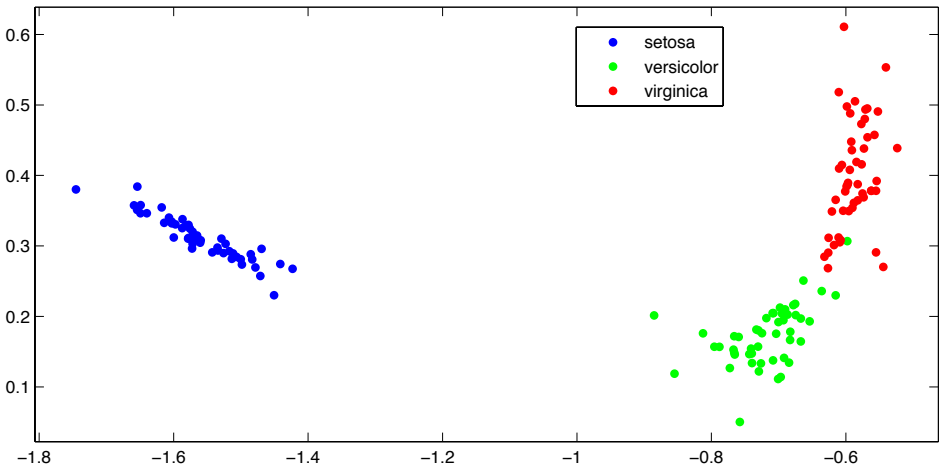


FIGURE 4 – Projection sur l'espace des variables discriminantes

1.5 Analyse discriminante décisionnelle

L'analyse discriminante décisionnelle pose le problème suivant : étant donné un nouvel individu sur lequel on a observé les p variables X_j mais pas la variable qualitative Y (l'appartenance à un groupe), comment décider de la modalité y_h de Y, c'est-à-dire du groupe auquel appartient cet individu ?

Pour cela nous allons définir des règles de décision et nous donner les moyens de les évaluer sur un seul individu. On affectera un individu x à la modalité y_h en minimisant sa distance (dans la métrique de Mahalanobis $\mathbf{W}^{-1}$) aux centres de gravité de chaque classe g^h, i.e.

$$\|x - g^h\|^2_{\mathbf{W}^{-1}} = (x - g^h)^T \mathbf{W}^{-1} (x - g^h) \tag{10}$$

ce qui revient à chercher la modalité y_h qui maximise la quantité :

$$l_h(x) = (g^h)^T \mathbf{W}^{-1} x - \frac{1}{2} (g^h)^T \mathbf{W}^{-1} g^h \tag{11}$$

Chacune de ces expressions, $1 \leq h \leq m$, est linéaire en x ce qui signifie que les séparations entre les classes sont des hyperplans définis par $l_h = l_k, h \neq k$.

1.6 Approche probabiliste

Introduction

Les résultats peuvent également être établis par l'intermédiaire d'une approche probabiliste. Considérons de nouveau l'exemple des iris de Fisher. Peut déterminer une espèce d'iris parmi les 3 en analysant seulement la longueur et la largeur des pétales ainsi que la longueur et la largeur des sépales ?

Pour cela, intéressons nous uniquement à une variable, la longueur des pétales, qui sera notée X. Si nous traçons par espèce (notée Y), les estimations de la densité de la longueur des pétales, nous obtenons 3 estimateurs (figure 5).

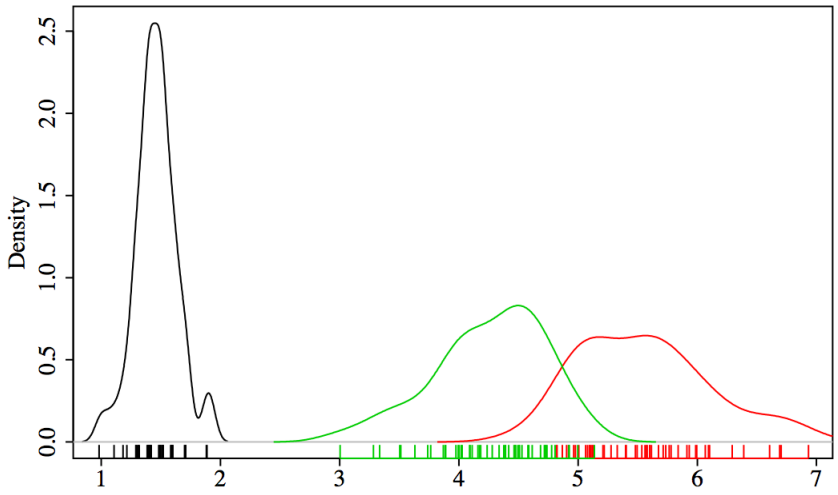


FIGURE 5 – Densités estimées des longueurs des pétales selon les espèces (noir=setosa, vert=versicolor, rouge=virginica)

Supposons que ces densités ont une forme connue, par exemple la loi normale $\mathcal{N}(\mu, \sigma)$, où bien entendu les paramètres de moyennes et de variances seraient différents d'une espèce à l'autre. Pour l'espèce *setosa* la densité s'écrit :

$$f(X, Y = setosa) = \frac{1}{\sqrt{2\pi}\sigma_{set}} \exp\left(-\frac{1}{2\sigma_{set}^2}(X - \mu_{set})^2\right)$$

de même pour la seconde et la troisième espèce :

$$f(X, Y = versicolor) = \frac{1}{\sqrt{2\pi}\sigma_{ver}} \exp\left(-\frac{1}{2\sigma_{ver}^2}(X - \mu_{ver})^2\right)$$

$$f(X, Y = virginica) = \frac{1}{\sqrt{2\pi}\sigma_{vir}} \exp\left(-\frac{1}{2\sigma_{vir}^2}(X - \mu_{vir})^2\right)$$

Ce raisonnement est intéressant pour décrire la variabilité d'une variable, la longueur des pétales, par groupe ou par espèce. Si nous n'avions que 2 variables explicatives, par exemple la longueur et la largeur des pétales, la représentation graphique d'un estimateur des densités (conjointes) par espèce est encore possible.

Cette représentation peu visuelle peut être remplacée par un contour des lignes de niveau de la densité (figure 7).

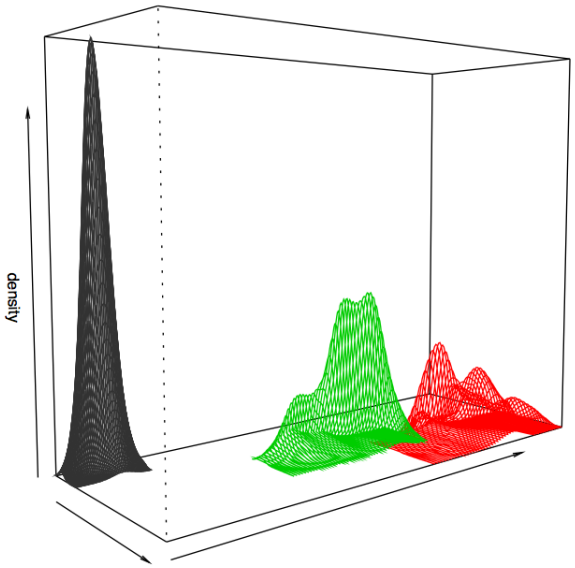


FIGURE 6 – Densités estimées des longueurs et largeurs des pétales selon les espèces (noir=setosa, vert=versicolor, rouge=virginica)

Or rappelons que nous sommes en présence de 4 variables et non pas 1 seule ou 2. L'extension naturelle pour prendre en compte ces 4 variables est simplement une loi de $\mathbb{R}^4$ par exemple une loi multi-normale de dimension 4. Nous aurons donc pour le groupe $G_h, h \in \{1, 2, 3\}$:

$$f(X, Y = y_h) = \frac{1}{(2\pi|\Sigma_h|)^{4/2}} \exp\left(-\frac{1}{2}(X - \mu_h)^T \Sigma_h^{-1}(X - \mu_h)\right)$$

où Σ_h est une matrice de variance, symétrique, carrée d'ordre 4 du groupe j et $\mu_h \in \mathbb{R}^4$ est le vecteur moyenne du groupe G_h.

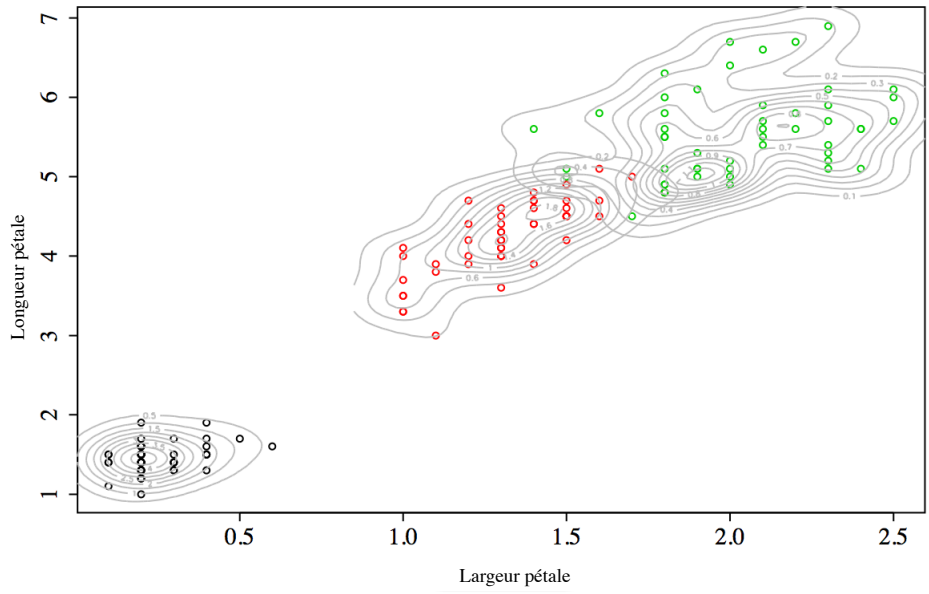


FIGURE 7 – Observation des longueurs et largeurs des pétales selon les espèces (noir=setosa, rouge=versicolor, vert=virginica) et ligne de niveau des densités estimées

Cependant avec ce modèle, dont les paramètres sont inconnus, il est impossible de prévoir l'espèce h au vu d'une observation nouvelle $X = x_0$. En effet ce modèle nous donne la variabilité des X sachant le groupe h, groupe inconnu que l'on souhaite justement connaître. Nous allons donc essayer de prévoir l'appartenance à une espèce d'un iris avec uniquement ses longueurs des pétales et sépales ainsi que ses largeurs des pétales et sépales. Ces mesures sont notées x_0. Lorsque l'on détermine l'appartenance à une espèce, sans connaître cette espèce, avec uniquement les longueurs et largeurs, il est inéluctable de faire des erreurs. Il existe donc une incertitude dans le processus de détermination, incertitude que nous pouvons modéliser par des probabilités d'appartenance à une espèce. Plus la probabilité d'un groupe h est grande, plus on est sûr de son classement parmi ce groupe, au vu des mesures x_0 des longueurs et largeurs.

Nous cherchons donc trois probabilités $P(Y = \mathit{setosa} \,|\, X = x_0)$, $(Y = \mathit{versicolor} \,|\, X = x_0)$ et $P(Y = \mathit{virginica} \,|\, X = x_0)$. Pour déterminer ces trois probabilités de classement nous utilisons le théorème de Bayes qui sous sa forme "probabilités discrètes" s'énonce :

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \,|\, A)P(A)}{P(B)}$$

Cela donne, en remplaçant les probabilités par des densités lorsque la variable est continue, les trois probabilités cherchées :

$$\begin{aligned}
P(Y = \mathit{setosa} \,|\, X = x_0) &= \frac{f(x_0 \,|\, Y = \mathit{setosa})P(Y = \mathit{setosa})}{f(x_0)} \\
P(Y = \mathit{versicolor} \,|\, X = x_0) &= \frac{f(x_0 \,|\, Y = \mathit{versicolor})P(Y = \mathit{versicolor})}{f(x_0)} \\
P(Y = \mathit{virginica} \,|\, X = x_0) &= \frac{f(x_0 \,|\, Y = \mathit{virginica})P(Y = \mathit{virginica})}{f(x_0)}
\end{aligned}$$

Remarquons que $f(x_0)$ au dénominateur est toujours présent dans les trois probabilités. Ce n'est donc pas ce facteur qui détermine l'appartenance à un groupe. De plus, puisque nous savons que ce sont des probabilités, la somme des trois vaut 1 et le dénominateur des fractions précédentes pourra être remplacé par :

$$\begin{aligned}
f(x_0) =& f(x_0 \,|\, Y = \mathit{setosa})P(Y = \mathit{setosa}) + f(x_0 \,|\, Y = \mathit{versicolor})P(Y = \mathit{versicolor}) \\
&+ f(x_0 \,|\, Y = \mathit{virginica})P(Y = \mathit{virginica})
\end{aligned}$$

Les probabilités $P(Y = y_h)$ représentent les probabilités *a priori* d'une espèce, c'est-à-dire la probabilité d'occurrence d'une espèce sans avoir aucune donnée. En général nous n'avons aucun *a priori* et donc ces probabilités sont choisies égales c'est-à-dire ici 1/3. Il est possible aussi de choisir le pourcentage d'observations dans un groupe ce qui ici est toujours 1/3. Enfin, si des études préalables ont donné des indications sur ces probabilités, il sera bon de les utiliser.

Nous pouvons maintenant énoncer de manière générale toutes les considérations vues dans cet exemple.

Analyse discriminante linéaire et quadratique

Nous sommes en présence de n observations d'un couple (Y, X). Pour l'observation i, notée (Y_i, X_i), Y_i est un label qui dénote l'appartenance à un groupe $G_h, h \in 1, \ldots, m$ et $X_i \in \mathbb{R}^p$ est un ensemble de variables explicatives de l'appartenance à un groupe (variable notée Y).

Une nouvelle observation arrive, nous mesurons les variables explicatives, cette mesure est notée $x_0 \in \mathbb{R}^p$ et nous souhaitons connaître son groupe y_0 inconnu. La probabilité conditionnelle de l'appartenance au groupe G_j sachant x_0 s'écrit :

$$P(Y = y_j \,|\, X = x_0) = \frac{f(x_0 \,|\, Y = y_j)P(Y = y_j)}{\sum_{h=1}^{m} f(x_0 \,|\, y = y_h)P(Y = y_h)}, \quad \forall j \in \{1, \ldots, m\} \tag{12}$$

Les probabilités a priori des groupes, notée $P(Y = y_j)$, sont connues. Quand l'utilisateur n'a pas d'*a priori*, il peut, soit choisir des groupes équiprobables $P(Y = y_j) = 1/m$, soit estimer la probabilité à partir des fréquences de chaque groupe dans les observations. Afin de spécifier le modèle de discrimination normal (ou quadratique), nous allons supposer l'hypothèse de normalité ci-dessous.

Discriminante quadratique. La densité des variables explicatives dans chaque groupe j suit une loi multi-normale : $f(X \,|\, Y = y_j) \approx \mathcal{N}(\mu_j, \Sigma_j)$ (hétéroscédasticité des variables).

Nous pouvons ensuite ajouter une hypothèse supplémentaire pour obtenir le modèle de discrimination linéaire.

Discriminante linéaire. La densité des variables explicatives dans chaque groupe j suit une loi multi-normale de même matrice de variance Σ dans chacun des groupes : $f(X \,|\, Y = y_j) \approx \mathcal{N}(\mu_j, \Sigma)$ (hypothèse d'homoscédasticité).

Une fois estimés tous les paramètres des lois normales, il suffit alors d'utiliser l'équation (12) pour connaître les probabilités d'affectation de la nouvelle observation aux différents groupes. Evidemment la prévision par la méthode sera donnée par le groupe le plus probable, c'est-à-dire :

$$j_0 = \underset{j \in \{1,\ldots,m\}}{\operatorname{argmax}} P(Y = y_j \,|\, X = x_0) = \underset{j \in \{1,\ldots,m\}}{\operatorname{argmax}} f(x_0 \,|\, Y = y_j)P(Y = y_j)$$

Estimation des paramètres

Pour chacun des m groupes, nous devons estimer $(\mu_j, \Sigma_j)_{j=1}^m$ où $\mu_j \in \mathbb{R}^p$ et $\Sigma_j \in \mathbb{R}^{p \times p}$. Il y a donc m moyennes à estimer et 1 ou m matrices de variance à estimer. Il existe de nombreuses procédures d'estimation plus ou moins classiques. En ce qui concerne les moyennes par groupes, on calcule le centre de gravité des groupes :

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i \in J} X_i$$

où J est l'ensemble des numéros d'observations qui sont dans le groupe j et n_j le nombre d'observations dans le groupe j (cardinal de J). Pour la méthode discriminante quadratique, les variances sont estimées par :

$$\hat{\Sigma}_j = \frac{1}{n_j - 1} \sum_{i \in J} (X_i - \hat{\mu}_j)(X_i - \hat{\mu}_j)^T$$

alors que pour la méthode discriminante linéaire, la variance est estimée par :

$$\hat{\Sigma} = \frac{1}{n - m} \sum_{j=1}^{m} \sum_{i \in J} (X_i - \hat{\mu}_j)(X_i - \hat{\mu}_j)^T$$

Interprétation géométrique

L'analyse discriminante possède une interprétation géométrique. Cette interprétation n'est pas utile pour faire des calculs ni pour appliquer la méthode mais elle permet d'associer une interprétation visuelle aux calculs précédents.

Afin de pouvoir faire des représentations graphiques, nous supposons que le nombre de variables explicatives p est égal à 2. Par ailleurs supposons que nous n'avons pas d'*a priori*, ce qui permet de ne pas s'occuper des $P(Y = y_j)$ qui sont tous égaux à $1/m$. Discriminer revient à chercher $\text{argmax}_{j\in\{1,\dots,m\}} f(x_0 \,|\, Y = y_j)P(Y = y_j) = \text{argmax}_{j\in\{1,\dots,m\}} f(x_0 \,|\, Y = y_j)$.

Comme $x_0 \in \mathbb{R}^2$ est un point du plan, nous cherchons à savoir, en fonction de la valeur de x_0, la classe que l'on va choisir. Il va y avoir des régions du plan où tous les points seront classés dans le groupe 1, d'autres où le classement sera 2, etc. Nous sommes donc intéressés par les frontières, c'est-à-dire l'ensemble des points x_0 que l'on peut classer soit dans une classe j, soit dans une autre j'. Cette frontière est simplement les points x_0 qui sont tels que :

$$f(x_0 \,|\, Y = y_j) = f(x_0 \,|\, Y = y'_j)$$

$$\frac{1}{2\pi|\Sigma_j|}\exp\left(-\frac{1}{2}(x_0-\mu_j)^T\Sigma_j^{-1}(x_0-\mu_j)\right) = \frac{1}{2\pi|\Sigma'_j|}\exp\left(-\frac{1}{2}(x_0-\mu_{j'})^T\Sigma_{j'}^{-1}(x_0-\mu_{j'})\right)$$

c'est-à-dire :

$$\ln\left(\frac{|\Sigma_j|}{|\Sigma'_j|}\right) - \frac{1}{2}x_0^T(\Sigma_j^{-1}-\Sigma_{j'}^{-1})x_0 + x_0^T(\Sigma_j^{-1}\mu_j - \Sigma_{j'}^{-1}\mu_{j'}) - \frac{1}{2}(\mu_j^T\Sigma_j^{-1}\mu_j - \mu_{j'}^T\Sigma_{j'}^{-1}\mu_{j'}) = 0$$

Si l'on développe cette équation en remplaçant le vecteur x_0 par ses coordonnées (x, y), nous obtenons une équation quadratique en x et y qui permet de dire qu'une frontière sera de la forme d'une conique. Cette constatation donne son nom à la méthode dite de discrimination quadratique.

En revanche, lorsque $\Sigma = \Sigma_j = \Sigma_{j'}$, nous avons alors :

$$x_0^T\Sigma_j^{-1}(\mu_j - \mu_{j'}) - \frac{1}{2}(\mu_j + \mu_{j'})^T\Sigma^{-1}(\mu_j - \mu_{j'}) = 0$$

Si l'on développe cette équation en remplaçant le vecteur x_0 par ses coordonnées (x, y) nous obtenons une équation d'une droite.

Exemple 1 (LDA dans $\mathbb{R}^2$ pour 3 groupes, variables X non corrélées)

Afin de visualiser cela sur un exemple, supposons que $m = 3$ et que $\Sigma = \Sigma_1 = \Sigma_2 = I_2$. Les observations suivent toutes des lois normales $\mathcal{N}(\mu_j, I_2)$. Les moyennes sont choisies égales à $\mu_1 = (2, 2)^T$, $\mu_2 = (-2, 2)^T$ et $\mu_3 = (0, -2)^T$ respectivement. La frontière entre le groupe 1 et le groupe 2 est donc :

$$x_0^T(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T(\mu_1 - \mu_2) = 0$$

$$(x_0 - \frac{1}{2}(\mu_1 + \mu_2))^T(\mu_1 - \mu_2) = 0$$

Soit M le point de coordonnée x_0, G_1 le centre de gravité du groupe 1, de coordonnées μ_1 et G_2 celui du groupe 2 de coordonnées μ_2. Soit G_{12} le milieu des deux points G_1, G_2. Il est de coordonnées $\frac{1}{2}(\mu_1 + \mu_2)$. Cette dernière équation se lit alors :

$$< \overrightarrow{G_{12}M}, \overrightarrow{G_2G_1} >= 0$$

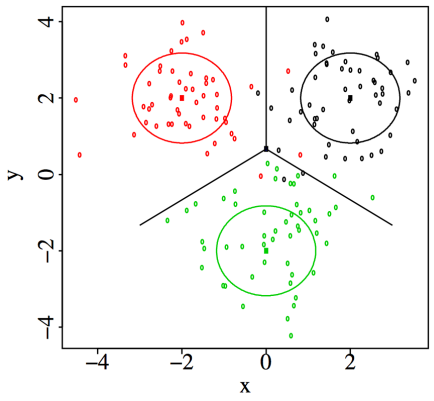


FIGURE 8 – Frontières théoriques. Les cercles correspondent à la région de probabilité 1/2

c'est-à-dire que les points M cherchés sont sur une droite passant par G_{12} et orthogonale à la droite portée par $\overrightarrow{G_2G_1}$ c'est-à-dire la droite (G_1G_2).

Les deux autres frontières théoriques peuvent être obtenues de la même façon. Comme, on ne connaît pas les valeurs de μ_j et Σ_j, on les remplace par leur estimateurs, ce qui fournit des frontières empiriques légèrement différentes.

Si nous changions d'exemple, avec des données $\mathcal{N}(\mu_j, \Sigma)$, la frontière entre les groupes 1 et 2 passerait toujours par G_{12}, mais comme le produit scalaire serait calculé par rapport à Σ^{-1}, l'angle serait différent. De plus les régions de probabilité seraient alors des ellipses.

Exemple 2 (LDA dans $\mathbb{R}^2$ pour 3 groupes (avec covariance))

Reprenons le même exemple avec $g = 3$ groupes, mais cette fois il existe une corrélation entre les 2 variables explicatives, $\Sigma = \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. Les observations suivent toutes des lois normales $\mathcal{N}(\mu_j, \Sigma)$, où μ_j est la moyenne du groupe. Les moyennes sont toujours choisies égales à $\mu_1 = (2, 2)^T$, $\mu_2 = (-2, 2)^T$ et $\mu_3 = (0, -2)^T$ respectivement. Le même calcul que

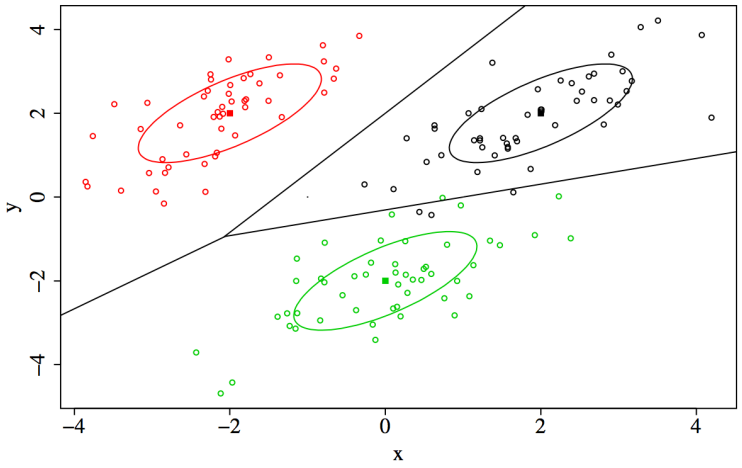


FIGURE 9 – Frontières théoriques. Les cercles correspondent à la région de probabilité 1/2

précédemment sur la frontière entre les groupes 1 et 2 aboutit à

$$< \overrightarrow{G_{12}M}, \overrightarrow{G_2G_1} >_{\Sigma^{-1}} = 0$$

Les deux autres frontières théoriques de la méthode LDA peuvent être obtenues de façon identique ainsi que les frontières empiriques.

2 Classification automatique

2.1 Introduction

Les techniques de classification automatique sont destinées à produire des groupements d'objets ou d'individus décrits par un certain nombre de variables ou de caractères. Le recours aux techniques de classification automatique est sous-tendu par quelques idées générales. On suppose que certains regroupements doivent exister, ou on exige que certains regroupements soient effectués. On manifeste donc un intérêt pour la mise en évidence de classes d'individus et de caractères. La nature des résultats attendus est diverse. Il s'agira soit de partitions des ensembles étudiés (lignes ou colonnes du tableau analysé) soit de hiérarchie de partitions. Quelquefois, il s'agira d'arbres, au sens de la théorie des graphes, où les sommets seront les objets à classer. Enfin, on pourra rechercher des classes empiétantes ou simplement mettre en évidence des zones à forte densité.

Pour l'essentiel, les techniques de classification font appel à une démarche algorithmique : une série d'opérations est définie de façon récursive et répétitive. Il existe plusieurs familles d'algorithmes de classification : les algorithmes ascendants (ou encore agglomératifs) qui procèdent à la construction des classes par agglomération successive des objets et qui fournissent une hiérarchie de partitions d'objets ; les algorithmes descendants (ou encore divisifs) qui procèdent par dichotomies successives de l'ensemble des objets, et qui peuvent également fournir une hiérarchie de partitions ; enfin les algorithmes conduisant directement à des partitions comme les méthodes d'agrégation autour de centres mobiles.

2.2 Classification ascendante hiérarchique : CAH

Présentation

Les principes généraux communs aux diverses techniques de classification ascendante sont extrêmement simples.

— on suppose au départ que l'ensemble des objets à classer est muni d'une distance [5]. Ceci ne suppose pas que les distances soient toutes calculées au départ ; il faut pouvoir alors les calculer ou les recalculer à partir des coordonnées des points-objets ;
— on suppose ensuite qu'il existe des règles de calcul des distances entre groupements disjoints d'objets. Cette distance entre groupements pourra en général se calculer directement à partir de distances des différents éléments impliqués dans le groupement.

Par exemple, si x, y et z sont trois objets et si x et y sont regroupés en un seul élément noté h, on peut définir la distance de ce groupement à z par la plus petite distance des divers éléments de h à z :

$$d(h, z) = \min\{d(x, z) \;\; d(y, z)\}$$

5. Il pourra s'agir d'une simple mesure de dissimilarité pour laquelle l'inégalité triangulaire $d(x, y) \leq d(x, Z) + d(y, z)$ n'est pas exigée

Cette distance s'appelle le *saut minimal* (*single linkage*). On peut également choisir la distance maximale (*complete linkage*) définie par :

$$d(h,z) = \max\{d(x,z)\ \ d(y,z)\}$$

Une autre règle simple et fréquemment employée est celle de la *distance moyenne* (*average linkage*) :

$$d(h,z) = (d(x,z) + d(y,z))/2$$

Si, plus généralement, x et y désignent des sous-ensembles disjoints de l'ensemble des objets ayant respectivement n_x et n_y éléments, h sera alors un sous-ensemble formé de $n_x + n_y$ éléments et on définira la distance moyenne généralisée suivante :

$$d(h,z) = (n_x d(x,z) + n_y d(y,z))/(n_x + n_y)$$

L'algorithme de classification ascendante hiérarchique se déroule de la façon suivante. On désignera par élément soit les objets à classer eux-mêmes, soit les regroupements d'objets générés par l'algorithme.

1. à l'étape 1, il y a n éléments à classer (qui sont les n objets) ;
2. on cherche les deux éléments les plus proches que l'on agrège en un nouvel élément ;
3. on calcule les distances entre le nouvel élément et les éléments restants. On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement $(n-1)$ éléments à classer ;
4. on réitère le processus à l'étape 2 jusqu'à ce qu'il n'y ait plus qu'un seul élément.

Illustrons cette procédure en prenant comme objet à classer cinq points du plan et comme *distance* entre ces objets le carré de leur distance euclidienne.

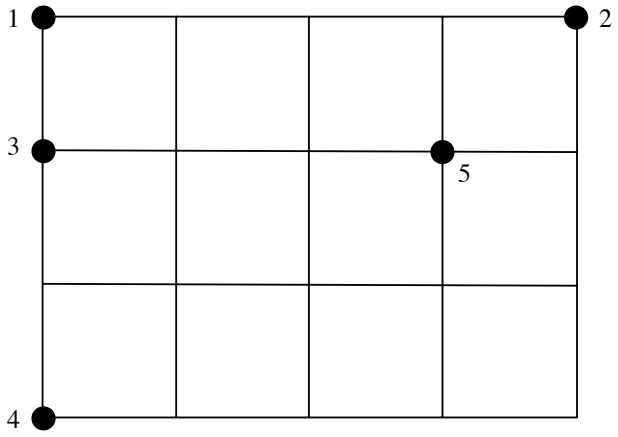


FIGURE 10 – Objets à classer

La matrice des distances ainsi définie et donnée au tableau suivant :

	(1)	(2)	(3)	(4)	(5)
(1)	0	16	1	9	10
(2)	16	0	17	25	2
(3)	1	17	0	4	9
(4)	9	25	4	0	13
(5)	10	2	9	13	0

1. les objets agrégés sont 1 et 3. Il est commode d'appeler 6 le nouvel élément obtenu. La nouvelle matrice des distances est donnée au tableau suivant (a). On a par exemple :

$$d(6,4) = \min\{d(1,4), d(3,4)\} = \min\{9,4\} = 4$$

2. les deux éléments 2 et 5 sont agrégés en l'élément 7. La nouvelle matrice est donnée au tableau (b) ;
3. on agrège en 8 les éléments 6 et 4. La matrice est donnée au tableau (c) ;
4. on agrège les deux éléments restant 8 et 7.

	(2)	(4)	(5)	(6)
(2)	0	25	2	16
(4)	25	0	13	4
(5)	2	13	0	9
(6)	16	4	9	0

a)

	(4)	(6)	(7)
(4)	0	4	13
(6)	4	0	9
(7)	13	9	0

b)

	(7)	(8)
(7)	0	9
(8)	9	0

c)

Finalement, les regroupements successifs ainsi que leur représentation par un arbre ou dendrogramme, où la valeur des distances correspondant aux différents niveaux d'agrégation a été portée en ordonnée, sont présentés à la figure 11.

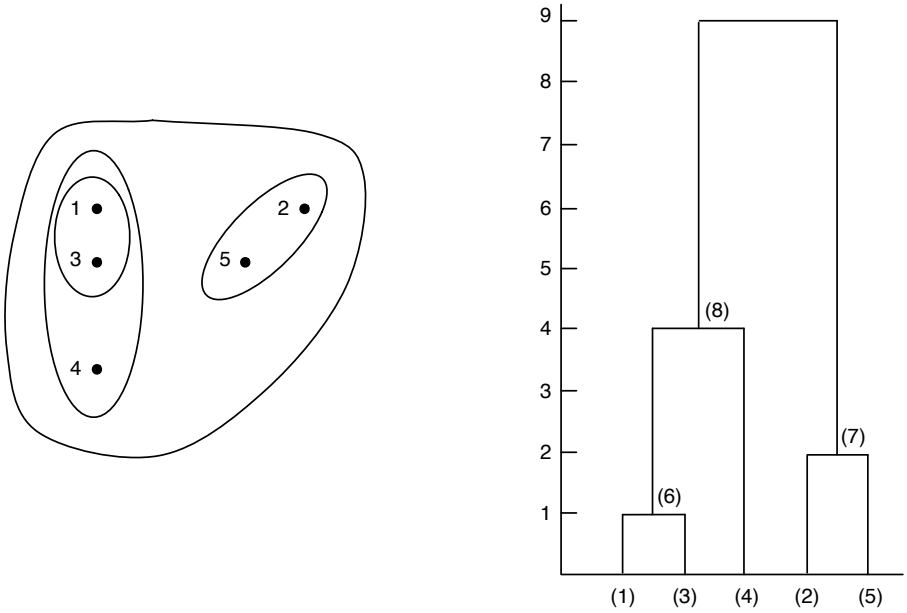


FIGURE 11 – Groupes successifs et dendrogramme

La famille des parties de l'ensemble des objets construite à partir d'algorithmes ascendants forme ce que l'on appelle une **hiérarchie**. On obtient une **hiérarchie indicée** si, à toute partie de h de la hiérarchie, est associée une valeur numérique $v(h) \geq 0$ compatible avec la relation d'inclusion : si $h \subset h'$ alors $v(h) < v(h')$. Pour la figure 11, la hiérarchie est indicée de façon naturelle par la valeur des distances correspondant à chaque étape d'agrégation (ordonnées). En coupant l'arbre par une droite horizontale, on obtient une **partition** d'autant plus fine que la section est proche des éléments terminaux.

Notion d'ultramétrique

La notion de hiérarchie est étroitement liée à une classe de distance entre objets appelée **ultramétrique**. Rappelons qu'un ensemble E est muni d'une métrique ou distance d, si d est une application de $E \times E$ dans $\mathbb{R}^+$ obéissant aux conditions suivantes :

1. $d(x, y) = 0$ si et seulement si $x = y$
2. $d(x, y) = d(y, x)$ (symétrie)
3. $d(x, y) \leq d(x, z) + d(x, z)$ (inégalité triangulaire)

Cette distance sera dite ultramétrique si elle vérifie la condition suivante, plus forte que l'inégalité triangulaire :

4. $d(x, y) \leq \max\{d(x, z), d(y, z)\}$

Il est équivalent de munir l'ensemble E d'une ultramétrique ou de définir une hiérarchie indicée de parties de cet ensemble. Montrons tout d'abord que toute hiérarchie indicée permet de définir une distance. En analysant le dendrogramme de la figure 11, on peut établir la matrice des distances suivantes :

	(1)	(2)	(3)	(4)	(5)
(1)	0	9	1	4	9
(2)	9	0	9	9	2
(3)	1	9	0	4	9
(4)	4	9	4	0	9
(5)	9	2	9	9	0

Montrons que l'on a toujours

$$d(x, y) \leq \max\{d(x, z), d(y, z)\}$$

Rappelons que deux parties de la hiérarchie H sont soit disjointes, soit liées par une relation d'inclusion. Appelons $h(x, z)$ la plus petite partie de H contenant x et z (dont l'indice est $d(x, z)$). Puisque $h(x, z)$ et $h(y, z)$ ne sont pas disjointes, on a, par exemple, $h(x, z) \subset h(y, z)$ (resp. $h(y, z) \subset h(x, z)$). Comme x, y et z sont tous trois contenus dans $h(y, z)$ (resp. $h(x, z)$), on a obligatoirement :

$$\begin{array}{rlll} & h(x, y) \subset h(y, z) & \text{et} & d(x, y) \leq d(y, z) \\ \text{resp.} & h(x, y) \subset h(x, z) & \text{et} & d(x, y) \leq d(x, z) \end{array}$$

ce qui établit l'inégalité.

Réciproquement, à toute ultramétrique d on peut faire correspondre une hiérarchie indicée dont d est l'indice associé. Il suffit d'appliquer l'algorithme du saut minimal au tableau des distances correspondant. On peut alors observer qu'il est inutile de procéder au calcul des distances à chaque étape : il suffit de rayer l'un des deux éléments agrégés.

En effet, si x et y sont agrégés en t, il faut en principe calculer les distances au nouvel élément t (figure 12). Or on a obligatoirement $d(z, x) \geq d(x, y)$ et de même $d(z, y) \geq d(x, y)$ sinon (z, x) ou (z, y) auraient été agrégés à la place de (x, y). Pour une ultramétrique, cela

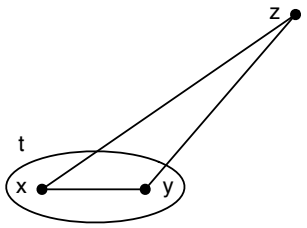


FIGURE 12 – Distance d'un groupe à un objet

implique que $d(z,x) = d(z,y)$, ce que l'on peut exprimer de façon imagée en disant que pour une ultramétrique, tous les triangles sont isocèles, avec le plus petit côté pour base. En effet, on a :

$$d(z,x) \leq \max\{d(x,y), d(z,y)\} \quad \text{donc} \quad d(z,x) \leq d(z,y)$$

De la même façon :

$$d(z,y) \leq \max\{d(z,x), d(x,y)\} \quad \text{donc} \quad d(z,y) \leq d(z,x)$$

Il s'ensuit que $d(z,y) = d(z,x)$. Le calcul des distances de z à t est finalement inutile puisque les deux distances mises en cause sont égales. Ceci montre comment l'algorithme du saut minimal opère sur la matrice de distances : il transforme la métrique initiale en ultramétrique, en diminuant certaines distances à chaque étape.

Agrégation selon un critère portant sur la variance

La technique de classification précédente a l'avantage de conduire à des calculs simples tout en possédant des propriétés mathématiques intéressantes. Cependant pour certaines applications les résultats obtenus sont critiquables. En particulier, le saut minimal a le défaut de produire des "effet de chaîne".

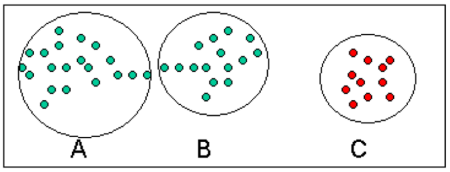


FIGURE 13 – Illustration de l'effet de chaîne

Pour les points de la figure 13, les groupes **A** et **B** ne sont pas discernables. Les techniques d'agrégation selon la variance cherchent à optimiser, à chaque étape, selon des critères liés à l'inertie, la partition obtenue par agrégation de deux éléments.

Nous considérons ici les n objets à classer comme des points d'un espace euclidien à p dimension. Chaque point x_i (vecteur à p composantes) sera muni d'une masse m_i. On note $M = \sum_i m_i$ la masse totale du nuage. Le carré de la distance entre les points x_i et x_j est noté :

$$\|x_i - x_j\|^2 = d^2(x_i, x_j)$$

L'inertie totale du nuage est la quantité :

$$I = \sum_i m_i \|x_i - g\|$$

où g est le centre de gravité global du nuage :

$$g = \frac{1}{M} \sum_i m_i x_i$$

S'il existe une partition de l'ensemble des objets en m classes, la $h^{ième}$ classe ayant pour centre de gravité g^h et pour masse m_h, la relation de Huygens, comme nous l'avons vu précédemment, fournit une décomposition de la quantité I en inerties intra-classes et interclasses selon :

$$I = \sum_h m_h \|g^h - g\|^2 + \sum_h \sum_{i \in G_h} m_i \|x_i - g^h\|^2$$

Soit x_i et x_j deux éléments de masses m_i et m_j que l'on agrège en un seul élément x de masse $m = m_i + m_j$, avec :

$$x = \frac{m_i x_i + m_j x_j}{m_i + m_j}$$

On peut décomposer l'inertie I_{ij} du groupement x par rapport à g selon :

$$I_{ij} = m_i \|x_i - x\|^2 + m_j \|x_j - x\|^2 + m \|x - g\|^2$$

Seul le dernier terme subsiste si x_i et x_j sont remplacés par leur centre de gravité. La perte d'inertie ΔI_{ij} vaut donc :

$$\Delta I_{ij} = m_i \|x_i - x\|^2 + m_j \|x_j - x\|^2$$

En remplaçant x par son expression en fonction de x_i et x_j, on obtient :

$$\Delta I_{ij} = \frac{m_i m_j}{m_i + m_j} \|x_i - x_j\|^2 = \frac{m_i m_j}{m_i + m_j} d^2(x_i, x_j)$$

La stratégie d'agrégation fondée sur le critère de la perte d'inertie minimale est donc la suivante : au lieu de chercher les deux éléments les plus proches, on cherche les éléments x_i et x_j correspondant à ΔI_{ij} minimal, ce qui revient à considérer les ΔI_{ij} comme de nouveaux indices de dissimilarité.

Si l'on travaille sur les coordonnées des points, on effectue les calculs des centres de gravité (x pour x_i et x_j). Par contre, si l'on travaille sur les distances, il est commode de calculer les nouvelles distances à partir des anciennes. Le carré des distances entre un point quelconque z et le centre de classe x s'écrit, en fonction des distances à x_i et x_j :

$$d^2(x, z) = \frac{1}{m_i + m_j} (m_i d^2(x_i, z) + m_j d^2(x_j, z) - \frac{m_i m_j}{m_i + m_j} d^2(x_i, x_j)$$

La méthode utilisant cet indice de dissimilarité est connue sous le nom de méthode de Ward.

2.3 Agrégation autour de centres mobiles

Soit à partitionner un ensemble I de n individu caractérisés par p paramètres ou variables. On suppose que l'espace $\mathbb{R}^p$ supportant les n points-individus est muni d'une distance appropriée notée d (fréquemment la distance euclidienne). On désire constituer au maximum q classes. L'algorithme d'agrégation autour des centres mobiles est le suivant :

Etape 0 : $k = 0$; déterminer q centres provisoires de classes. Le choix de ces centres est important pour la rapidité de la convergence, et les connaissances *a priori* doivent ici être mises à profit, s'il y en a. Dans le cas contraire, le plus courant, il suffit de tirer aléatoirement ces centres par un tirage sans remise. Ces centres $\{C_1^0, C_2^0, \ldots, C_q^0\}$ induisent une partition P^0 de I en q classes $\{I_1^0, I_2^0, \ldots, I_q^0\}$. Ainsi l'individu i appartient à I_k^0 si le point i est plus proche de C_k^0 que de tous les autres centres. Les classes sont délimitées dans l'espace par des polytopes convexes formées par les plans médiateurs des segments joignant tous les couples de centres (diagramme de Voronoï).

Etape k : $k = k+1$; déterminer q nouveaux centres de classes $\{C_1^k, C_2^k, \ldots, C_q^k\}$ en prenant les centres de gravité des classes $\{I_1^{k-1}, I_2^{k-1}, \ldots, I_q^{k-1}\}$. Ces nouveaux centres induisent une nouvelle partition P^k construite selon la même règle et formée des classes $\{I_1^k, I_2^k, \ldots, I_q^k\}$. Retour à l'étape k.

L'algorithme s'arrête soit lorsque deux itérations successives conduisent à la même partition, soit lorsqu'un critère convenablement choisi (par exemple la variance intra-classe) cesse de décroître de façon sensible, soit encore parce qu'un nombre maximal d'itérations défini *a priori* est atteint.

Cet algorithme est illustré sur la figure 14 dans le cas où $q = 2$. Le tirage aléatoire des centres provisoires C_1^0 et C_1^0 et la construction de la première partition $P^0 = \{I_1^0, I_2^0\}$ en affectant chaque individu au sous-nuage dont le centre est le plus proche. L'étape 1 présente les nouveaux centres et les sous-nuages dont ils sont les centres de gravité. De nouveau, l'étape 2 fournit les centres de gravité des nouveaux sous-nuages I_1^2, I_2^2.

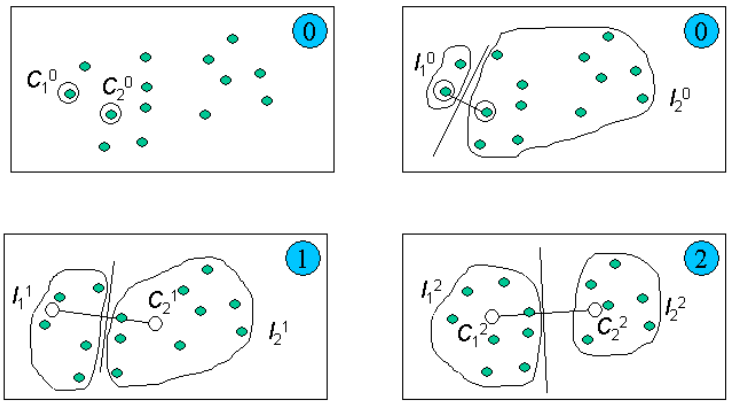


FIGURE 14 – Algorithme des centres mobiles

De nombreux algorithmes sont fondés sur un principe similaire. Les deux principaux sont les nuées dynamiques et les *k-means* ou **k-moyennes**. La différence pour la méthode des nuées dynamiques se situe au niveau de la réaffectation des individus à chaque classe. Après avoir déterminé les centres de gravité, un noyau est déterminé pour chaque classe comme étant l'individu le plus proche du centre de gravité de chaque classe. La réaffectation se fait alors en fonction de la distance des autres individus aux noyaux de chaque classe. Ce formalisme a permis plusieurs généralisations de la méthode.

La méthode des *k-means* après avoir choisi une première fois les centres mobiles, recalcule le centre de chaque classe dès lors qu'un individu y est affecté. La position du centre est donc modifiée à chaque affectation, ce qui permet d'avoir une bonne partition en peu d'itérations.

3 Les séparateurs à vaste marge (SVM : Support Vector Machine)

3.1 Introduction

Les *Support Vector Machines* ou, en français, Séparateurs à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination, c'est-à-dire la prévision d'une variable qualitative binaire $(-1, +1)$.

Ils sont basés sur la recherche de l'**hyperplan de marge optimale** qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de généralisation (qualité de prévision) est la plus grande possible.

Cette approche découle directement des travaux de Vapnik en théorie de l'apprentissage à partir de 1995. Elle s'est focalisée sur les propriétés de généralisation (ou prévision) d'un modèle en contrôlant sa complexité. L'autre idée directrice de Vapnik est d'éviter de substituer à l'objectif initial : la discrimination, un ou des problèmes qui s'avèrent finalement plus complexes à résoudre comme par exemple l'estimation non paramétrique de la densité d'une loi multidimensionnelle en analyse discriminante.

Un problème de discrimination est dit **linéairement séparable** lorsqu'il existe une fonction de décision linéaire (appelé aussi séparateur linéaire) de la forme $D(x) = \text{signe}(f(x))$ avec $f(x) = w^T x + b$, $w \in \mathbb{R}^p$, $b \in \mathbb{R}$, classant correctement toutes les observations de l'ensemble d'apprentissage.

La fonction f est appelée fonction caractéristique. C'est un problème particulier qui semble très spécifique, mais qui permet d'introduire de manière pédagogique les principaux principes des SVM : marge, programmation quadratique, vecteur support, formulation duale et matrice de Gram.

On décide donc qu'une observation x est de classe 1 si $f(x) \geq 0$ et de classe -1 sinon. La frontière de décision $f(x) = 0$ est un hyperplan, appelé hyperplan séparateur, ou séparatrice. Le but d'un algorithme d'apprentissage supervisé est d'apprendre la fonction $f(x)$ par le biais d'un ensemble d'apprentissage :

$$\{(x_1, \ell_1), (x_2, \ell_2), \ldots, (x_N, \ell_N)\} \subset \mathbb{R}^p \times \{-1, 1\}$$

où les ℓ_k sont les labels traduisant l'appartenance à une classe donnée, N est la taille de l'ensemble d'apprentissage et p la dimension des vecteurs d'entrée.

On verra ultérieurement que l'on peut définir, à partir de cette première formulation, des problèmes plus complexes selon que les données sont linéairement séparables ou non ou selon que l'on cherche une séparatrice linéaire ou non linéaire. La figure ci-après illustre ces différentes situations.

3.2 Formulation du problème de discrimination linéaire

Si le problème est linéairement séparable, on doit alors avoir :

$$\ell_k f(x_k) \geq 0, \quad 1 \leq k \leq N, \qquad \text{autrement dit} \qquad \ell_k(w^T x_k + b) \geq 0, \quad 1 \leq k \leq N$$

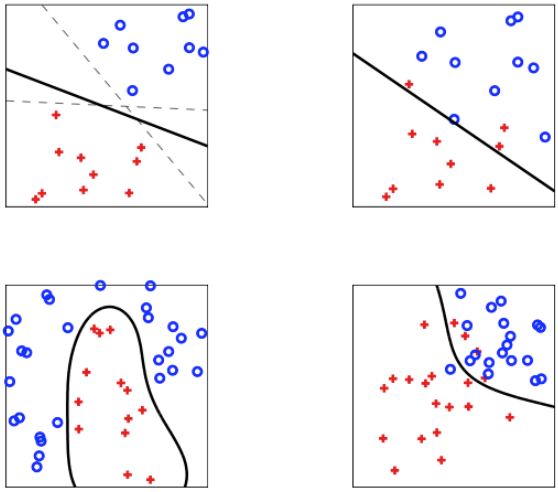


FIGURE 15 – Quatre types de problèmes de discrimination binaire (frontière de décision en noir). Séparabilité linéaire ou non linéaire

A toute fonction de décision on peut associer une frontière de décision :

$$\Delta(w, b) = \{x \in \mathbb{R}^p \mid w^T x + b = 0\}$$

Comme la fonction de décision linéaire, cette frontière de décision est définie à un terme multiplicatif près dans le sens où la frontière définie par le couple (w, b) est la même que celle engendrée par $(kw, kb), \forall k \in \mathbb{R}$. Pour garantir l'unicité de la solution, on peut soit considérer l'hyperplan standard (tel que $\|w\| = 1$) soit l'**hyperplan canonique** par rapport à un point x (tel que $w^T x + b = 1$).

Même dans ce cas simple, le choix de l'hyperplan séparateur n'est pas évident. Il existe en effet une infinité d'hyperplans séparateurs, dont les performances en apprentissage sont identiques (le risque empirique est le même), mais dont les performances en généralisation peuvent être très différentes.

Pour résoudre ce problème, on cherche à déterminer un hyperplan optimal, défini comme l'hyperplan qui maximise la **marge** entre les observations et l'hyperplan séparateur. La marge est la plus petite distance entre les observations d'apprentissage et l'hyperplan séparateur qui satisfait la condition de séparabilité : $\ell_k(w^T x_k + b) \geq 0$. La distance d'une observation x_k à l'hyperplan est donnée par sa projection orthogonale sur l'hyperplan :

$$d_k = \frac{\ell_k(w^T x_k + b)}{||w||}$$

L'hyperplan séparateur (w, b) de marge maximale est donc donné par :

$$\arg\max_{w,b} \left\{ \frac{1}{||w||} \min_k \left[\ell_k(w^T x_k + b) \right] \right\}$$

Afin de faciliter l'optimisation, on choisit de normaliser w et b, de telle manière que les observation à la marge (x^+_{marge} pour les **vecteurs supports** sur la frontière positive et x^-_{marge} pour ceux situés sur la frontière opposée) satisfassent :

$$\begin{cases} w^T x^+_{\text{marge}} + b = 1 \\ w^T x^-_{\text{marge}} + b = -1 \end{cases}$$

D'où pour toutes les observations :

$$\ell_k(w^T x_k + b) \geq 1, \quad k = 1, \ldots, N$$

Cette normalisation est appelée forme canonique de l'hyperplan ou **hyperplan canonique**. Avec cette normalisation, la marge vaut $\dfrac{1}{||w||}$ et il s'agit donc de maximiser $||w||^{-1}$.

La formulation dite **primale** des SVM s'exprime alors sous la forme d'un problème d'optimisation quadratique sous contraintes inégalités :

$$\min_{w,b} \quad \frac{1}{2}||w||^2 \quad \text{sous les contraintes} \quad \ell_k(w^T x_k + b) \geq 1$$

En effet, l'optimum de la maximisation de $||w||^{-1}$ et le même que celui de la minimisation de $||w||^2$ ($||w||$ étant positif et l'élévation au carré étant monotone sur $]0, +\infty[$). Ce problème peut se résoudre par la méthode classique des multiplicateurs de Lagrange, où le lagrangien est donné par :

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{k=1}^{N} \alpha_k \left(\ell_k(w^T x_k + b) - 1\right)$$

Le lagrangien doit être minimisé par rapport à w et b et maximisé par rapport aux paramètres de Lagrange $\alpha = (\alpha_1 \;\; \alpha_2 \;\; \ldots \;\; \alpha_N)$ (en effet, ici, l'optimum est un **point selle**).

On peut donc formuler le problème des SVM sur des données linéairement séparables de la façon suivante :

Soit $\{(x_k, \ell_k), k = 1, N\}$ un ensemble de vecteurs-observations étiquetés avec $x_k \in \mathbb{R}^p$ et $\ell_k \in \{1, -1\}$. Un séparateur à vaste marge linéaire (SVM) est un discriminateur linéaire de la forme : $D(x) = \text{signe}(w^T x + b)$ où $w \in \mathbb{R}^p$ et $b \in \mathbb{R}$ sont donnés par la résolution du problème suivant :

$$\text{Primal} \quad \begin{cases} \min\limits_{w,b} \dfrac{1}{2}||w||^2 \\ \text{sous } \ell_k(w^T x_k + b) \geq 1, \quad k = 1, \ldots, N \end{cases} \tag{13}$$

3.3 Résolution du problème de discrimination linéaire

Le problème d'optimisation sous contraintes précédent (13) est un "programme quadratique" de la forme générale :

$$\begin{cases} \min\limits_{z} \dfrac{1}{2} z^T H z + f^T z \\ \text{sous } Az \leq e \end{cases}$$

avec

$$z = \begin{pmatrix} w \\ b \end{pmatrix} \in \mathbb{R}^{p+1} \qquad f = (0, \ldots, 0)^T \in \mathbb{R}^{p+1} \qquad H = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}$$

$$A = -(\text{diag}(\ell)X \quad \ell) \qquad e = -(1, \ldots, 1)^T \in \mathbb{R}^N \qquad \ell \in \mathbb{R}^N \text{ vecteur des signes}$$

$X \in \mathbb{R}^{N \times p}$ matrice dont la ligne k est x_k^T

Ce problème est convexe puisque la matrice A est semi-définie positive. Il admet donc une solution unique (qui existe puisque le problème est linéairement séparable par hypothèse). Ce problème (dit primal) admet une **formulation duale** équivalente qui est aussi un programme quadratique.

Explicitons le lagrangien lié au problème d'optimisation précédent :

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{k=1}^{N} \alpha_k(\ell_k(w^T x_k + b) - 1)$$

où les $\alpha_i \geq 0$ sont les multiplicateurs de Lagrange associés aux contraintes.

Les conditions d'optimalité de Karush, Kuhn et Tucker permettent de caractériser la solution du problème primal (w^*, b^*) et les multiplicateurs de Lagrange α^* associés par le système d'équations suivant :

$$\begin{array}{rlll}
\text{stationarité / } w: & w^* - \sum_{k=1}^{N} \alpha_k^* \ell_k x_k & = 0 & \\
\text{stationarité / b}: & \sum_{k=1}^{N} \alpha_k^* \ell_k & = 0 & \\
\text{complémentarité}: & \alpha_k^*(\ell_k(w^{*T} x_k + b^*) - 1) & = 0 & k = 1, \ldots, N \\
\text{admissibilité primale}: & \ell_k(w^{*T} x_k + b^*) & \geq 1 & k = 1, \ldots, N \\
\text{admissibilité duale}: & \alpha_k^* & \geq 0 & k = 1, \ldots, N
\end{array}$$

Les conditions de complémentarité permettent de définir l'ensemble $\mathcal{A}$ des indices des **contraintes actives** (ou saturées) à l'optimum dont les multiplicateurs de Lagrange α_k sont strictement positifs :

$$\mathcal{A} = \{k \in [1, n] \mid \ell_k(w^{*T} x_k + b^*) = 1\}$$

Pour les autres contraintes, la condition de complémentarité implique que leur mutiplicateur de Lagrange est égal à zéro et que l'observation associée vérifie strictement l'inégalité : $\ell_j(w^{*T} x_j + b^*) > 1, \forall j \notin \mathcal{A}$.

Si l'on note $\bullet_{\mathcal{A}}$ le vecteur constitué des seules composantes de $\bullet$ indexées par $\mathcal{A}$ (ou la matrice constituée des lignes indexées par $\mathcal{A}$), la solution optimale $(w^*, b^*, \alpha_{\mathcal{A}}^*)$ vérifie le système d'équations linéaires suivant :

$$\left\{ \begin{array}{llll}
w^* & -X_{\mathcal{A}}\text{diag}(\ell_{\mathcal{A}})\alpha_{\mathcal{A}}^* & & = 0 \\
\text{diag}(\ell_{\mathcal{A}})X_{\mathcal{A}} w^* & & +b^* \ell_{\mathcal{A}} & = e_{\mathcal{A}} \\
 & -\ell_{\mathcal{A}}^T \alpha_{\mathcal{A}}^* & & = 0
\end{array} \right.$$

où $e_{\mathcal{A}}$ est un vecteur unité (constitué de 1) de la dimension adéquate. De la première égalité, on obtient donc : $w^* = \sum_{i \in \mathcal{A}} \alpha_i^* \ell_i^* x_i$.

Le vecteur w est donc une combinaison linéaire des observations x_i liées aux contraintes actives $i \in \mathcal{A}$ et pour lesquelles on a $|w^{*T} x_i + b^*| = 1$ (observations sur les frontières). Ces observations sont appelées **vecteurs supports** ; leur marge est égale à 1.

Les autres données (celles correspondant aux contraintes inactives, $i \notin \mathcal{A}$) n'interviennent pas dans le calcul (elles sont à une distance supérieure à 1 de l'hyperplan séparateur).

3.4 Formulation duale

Au problème d'optimisation initial,

$$\text{Primal} \quad \left\{ \begin{array}{ll}
\min\limits_{w,b} \dfrac{1}{2}||w||^2 & \\
\text{sous} \quad \ell_k(w^T x_k + b) \geq 1, & k = 1, \ldots, N
\end{array} \right.$$

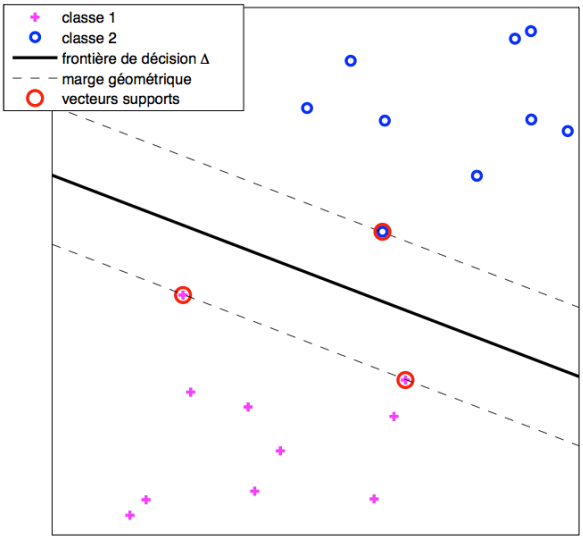


FIGURE 16 – Plan séparateur, vecteurs supports et marge

on peut associer la formulation duale suivante (voir Annexe 1) :

$$\text{Dual} \quad \begin{cases} \max\limits_{w,b,\alpha} \dfrac{1}{2}||w||^2 - \sum\limits_{k=1}^{N} \alpha_k(\ell_k(w^T x_k + b) - 1) \\ \text{sous} \quad w - \sum_{k=1}^{N} \alpha_k \ell_k x_k = 0 \\ \qquad \sum_{k=1}^{N} \alpha_k \ell_k = 0 \\ \qquad \alpha_k \geq 0 \quad k = 1, \ldots, N \end{cases}$$

ces deux problèmes d'optimisation quadratique ayant le même optimum (w^*, b^*).

L'élimination de la variable primale w permet d'obtenir la formulation suivante :

$$\text{Dual} \quad \begin{cases} \min\limits_{\alpha} \dfrac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} \alpha_i \alpha_i \ell_i \ell_j x_i^T x_j - \sum\limits_{k=1}^{N} \alpha_k \\ \text{sous} \quad \sum_{k=1}^{N} \alpha_k \ell_k = 0 \\ \qquad \alpha_k \geq 0 \quad k = 1, \ldots, N \end{cases} \tag{14}$$

que l'on peut écrire sous la forme plus compacte suivante :

$$\text{Dual} \quad \begin{cases} \min\limits_{\alpha} & \dfrac{1}{2}\alpha^T G \alpha - e^T \alpha \\ \text{sous} & \ell^T \alpha = 0 \\ & \alpha_k \geq 0, \quad k = 1, \ldots, N \end{cases} \tag{15}$$

où ℓ est le vecteur des $\ell_k, k = 1, \ldots, N$, e un vecteur de dimension N ne contenant que des 1 et G est une matrice symétrique, de dimension $N \times N$, dont l'élément générique est défini par $G_{i,j} = \ell_i \ell_j x_i^T x_j$ (G est une matrice de Gram) que l'on pourra écrire sous forme matricielle $G = \text{diag}(\ell) X X^T \text{diag}(\ell)$.

Les deux formulations primale et duale peuvent être exploitées pour résoudre le problème posé. On notera que le problème primal fait intervenir $p+1$ inconnues et N contraintes inégalités alors que le dual fait intervenir N inconnues, N contraintes égalités et N contraintes inégalités. On remarquera également que dans le problème dual, le paramètre b a disparu et qu'il faudra mettre en place une méthode pour l'estimer.

La formulation duale semblerait donc moins "intéressante", le nombre d'observations N étant en général très supérieur à la dimension p de celles-ci. Il faut cependant remarquer que la formulation duale ne fait intervenir que des **produits scalaires** $x_i^T x_j$ entre les vecteurs-observations. Cette remarque est fondamentale pour l'extension de la méthode permettant de déterminer des **séparatrices non linéaires**.

Rappelons à cet effet que la fonction de décision linéaire est de la forme $D(x) = \text{signe}(f(x))$ avec $f(x) = w^T x + b$. En injectant l'expression de la solution optimale $w^* = \sum_{k=1}^{N} \alpha_k^* \ell_k x_k$, on obtient :

$$f(x) = \sum_{k=1}^{N} \alpha_k^* \ell_k x_k^T x + b^*$$

Cependant, rappelons que de nombreux coefficients α_k^* sont nuls. Seuls les paramètres de Lagrange relatifs aux vecteurs supports sont non nuls, on a donc :

$$f(x) = \sum_{k \in \mathcal{A}} \alpha_k^* \ell_k x_k^T x + b^* \tag{16}$$

où $\mathcal{A}$ est l'ensemble des indices des vecteurs supports.

Qu'il s'agisse du problème à résoudre (14) ou de la fonction de décision (16), les deux s'appuient sur le calcul de produits scalaires entre les observations d'apprentissage ou entre un vecteur à classer et les vecteurs-observations d'apprentissage.

3.5 Cas de données non séparables linéairement – marge poreuse ou souple

L'ensemble d'apprentissage :

$$\{(x_1, \ell_1), (x_2, \ell_2), \ldots, (x_N, \ell_N)\} \subset \mathbb{R}^p \times \{-1, 1\}$$

peut ne pas être linéairement séparable. Il faut alors relâcher les contraintes en autorisant certaines observations à avoir une marge inférieure à 1 voire une marge négative. Les contraintes du problème primal initial : $\ell_k(w^T x_k + b) \geq 1$ peuvent alors être remplacées par :

$$\ell_k(w^T x_k + b) \geq 1 - \xi_k \quad \text{avec } \xi_k \geq 0$$

Il faut alors modifier le critère d'optimisation en introduisant une pénalité pour ces observations. Le problème se transforme alors en :

$$\begin{cases} \min\limits_{w,b,\xi_k} \dfrac{1}{2}||w||^2 + C \sum\limits_{k=1}^{N} \xi_k \\ \text{sous} \quad \ell_k(w^T x_k + b) \geq 1 - \xi_k, \quad k = 1, \ldots, N \\ \qquad\quad \xi_k \geq 0, \quad k = 1, \ldots, N \end{cases} \tag{17}$$

Les variables supplémentaires $\xi_k, k = 1, \ldots, N$ (appelées *slack variables* en anglais ou **variables ressorts**) sont introduites afin de relâcher les contraintes : on accepte que certaines

observations franchissent la marge (distance de l'observation à la séparatrice inférieure à 1 ou, même, soient du mauvais côté de l'hyperplan séparateur).

Pour chaque observation, l'expression $\ell_k(w^T x_k + b) \geq 1 - \xi_k$ a un sens à la condition que $\xi_k \geq 0$. Cette variable renseigne à quel point une observation (ξ_k, ℓ_k) viole la contrainte :
— si $\xi_k = 0$, l'observation respecte la contrainte ;
— si $\xi_k > 1$, l'observation est mal classée ;
— si $0 < \xi_k < 1$, l'observation est bien classée, mais elle a franchi la marge.

On détermine ces variables de façon à ce qu'elles soient le plus petites possible (ou que leur somme soit la plus petite possible puisqu'elles sont toutes positives), d'où l'ajout du terme correspondant dans le critère. Le scalaire C gère le compromis entre une marge maximale et le relâchement des contraintes.

En utilisant la formulation duale, le problème (17) peut se ré-écrire sous la forme (voir annexe A.5) :

$$\begin{cases} \min\limits_{\alpha} & \frac{1}{2}\alpha^T G \alpha - e^T \alpha \\ \text{sous} & \ell^T \alpha = 0 \\ & 0 \leq \alpha_k \leq C, \quad k = 1, \ldots, N \end{cases} \tag{18}$$

où la seule différence par rapport au problème (15) est l'apparition d'une borne supérieure sur les coefficients α_k.

3.6 Cas non linéaire - utilisation de noyaux

Dans beaucoup de situations, il est préférable de rechercher des frontières de décision non linéaires entre classes. Nous avons précédemment fait remarquer que les résultats obtenus dans le cadre linéaire s'appuient essentiellement sur le calcul de produits scalaires. Grâce à cette propriété, et de façon assez inattendue, la transposition du cas linéaire au cas non linéaire va s'effectuer sans difficulté majeure en utilisant des fonctions noyaux.

De façon générale, les méthodes à noyaux permettent de trouver des fonctions de décision non linéaires, tout en s'appuyant fondamentalement sur des méthodes linéaires. Une fonction noyau correspond à un produit scalaire dans un espace de redescription des données, souvent de grande dimension. Dans cet espace, qu'il n'est pas nécessaire de manipuler explicitement, les méthodes linéaires peuvent être mises en œuvre pour y trouver des régularités linéaires, correspondant à des régularités non linéaires dans l'espace d'origine.

L'idée générale consiste à projeter les données initiales appartenant à l'espace $\mathcal{X}$ dans un nouvel espace, appelé espace de redescription (*feature space*), $\mathcal{F}$, de dimension supérieure (voire de dimension infinie).

$$\begin{aligned} \phi : \mathcal{X} &\to \mathcal{F} \\ x &\to \phi(x) \end{aligned}$$

Ce nouvel espace doit être muni d'un produit scalaire (espace de Hilbert) de façon à pouvoir calculer, pour deux observations x et y de $\mathcal{X}$, le produit scalaire de leurs transformées $\phi^T(x)\phi(y)$. On cherche alors dans ce nouvel espace un hyperplan séparateur (à marge dure ou à marge souple) comme on l'a fait précédemment pour le cas linéaire.

L'exemple ci-dessous illustre le fait que l'immersion dans un espace de plus grande dimension d'un jeu de données non séparables linéairement peut être bénéfique. Considérons des données dans $\mathbb{R}^2$. Les 50 observations appartenant à la première classe sont repérées par des cercles rouges et les 50 observations appartenant à la seconde classe sont repérées par des croix bleues.

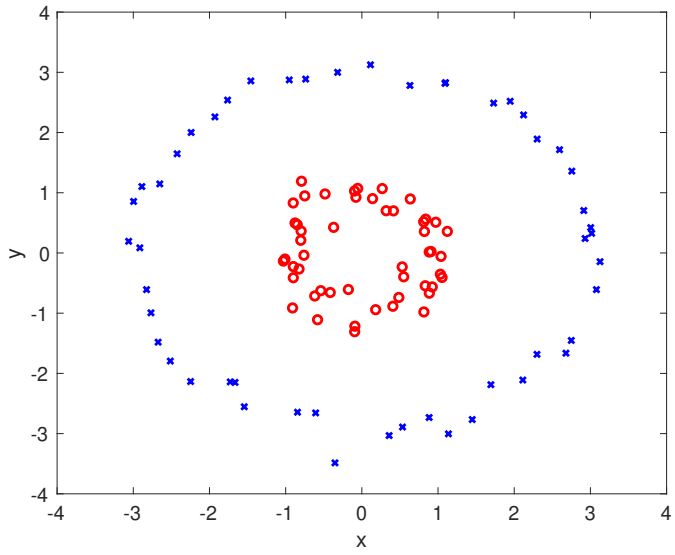


FIGURE 17 – Données dans $\mathbb{R}^2$

Ces deux classes ne sont clairement pas séparables linéairement. Projetons les dans un espace de plus grande dimension, ici $\mathbb{R}^3$, tel que :

$$\begin{aligned} \phi \quad &: \mathbb{R}^2 \quad \to \mathbb{R}^3 \\ &(x, y) \quad \to (x, y, z = x^2 + y^2) \end{aligned}$$

Si l'on "visualise" les données dans cet espace, on observe qu'elles deviennent séparables linéairement.

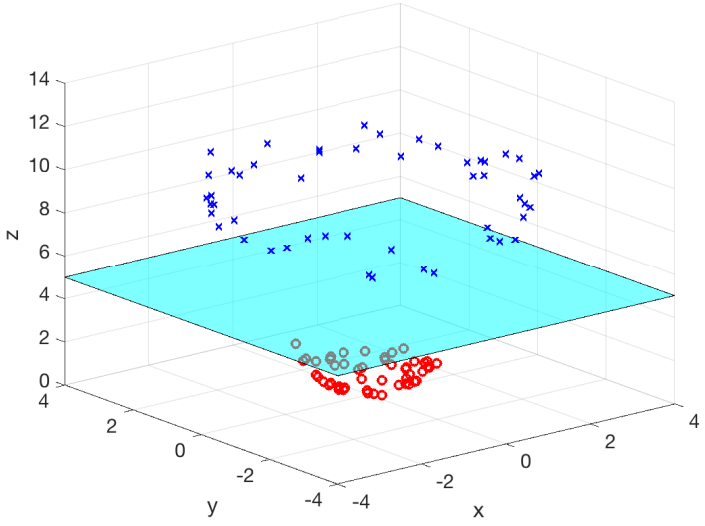


FIGURE 18 – Données dans $\mathbb{R}^3$

On peut, en particulier, le mettre en évidence en projetant les données dans le plan (y, z) par exemple :

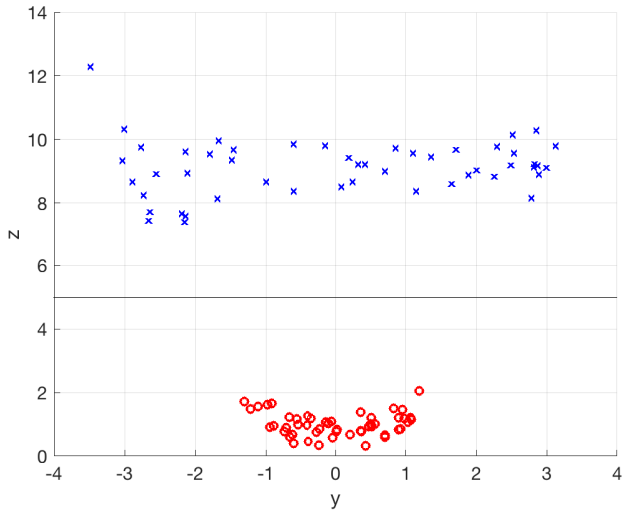


FIGURE 19 – Projection des données de $\mathbb{R}^3$ sur le plan (y, z)

Cet exemple est très fréquemment utilisé. On notera cependant qu'il est très artificiel. Avant d'envisager le traitement quelconque d'un jeu de données, il convient, lorsque c'est possible, de l'analyser. Ici la séparation peut être effectuée par une projection ... dans $\mathbb{R}$! Il suffit en effet de calculer le module de chaque vecteur-observation : $\rho = \sqrt{x^2 + y^2}$.

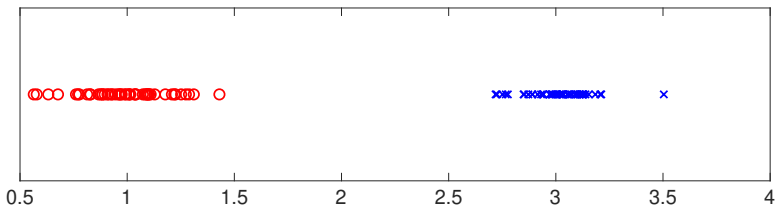


FIGURE 20 – Module des 100 observations

La projection dans un espace de grande dimension peut entraîner une augmentation importante des temps de calcul. Dans l'exemple précédent, la fonction ϕ est explicite et elle projette les données initiales dans un espace de dimension finie. On pressent intuitivement que si la séparation des données en classes est difficile dans l'espace initial, plus l'espace dans lequel on projette sera de grande dimension (voire de dimension infinie) plus on a de chance de pouvoir trouver un séparateur linéaire des différentes classes. La question que l'on peut donc se poser est : peut-on s'affranchir de la difficulté de cette projection et conserver l'avantage fourni par la grande dimension ? La réponse est oui et passe par l'utilisation de fonctions noyaux.

Un noyau κ est défini de manière générale comme une fonction de deux variables sur $\mathbb{R}$:

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\to \kappa(x, y) \end{aligned}$$

La matrice de Gram du noyau $\kappa(.,.)$ pour les observations $\{x_1 \ldots, x_i, \ldots, x_n\}$ (pour tout entier n fini) est la matrice carrée K de dimension $n \times n$ et de terme général $K_{ij} = \kappa(x_i, x_j)$.

Un noyau κ est dit positif si, pour tout entier n fini et pour toutes les suites de n observations possibles $\{x_i, i = 1, \ldots, n\}$, la matrice de Gram associée est une matrice symétrique définie positive.

Si l'on se donne une transformation $\phi(x)$ de $\mathcal{X}$ vers un espace de Hilbert $\mathcal{H}$ muni d'un produit scalaire, le noyau peut donc être défini en utilisant ce produit scalaire :

$$\kappa(x, y) = \phi^T(x)\phi(y)$$

mais, de façon plus intéressante, on démontre aussi que l'on peut associer un produit scalaire à tout noyau positif. Il existe donc de nombreux noyaux parmi lesquels on peut citer :

$$\begin{array}{ll} \text{le noyau polynomial d'ordre } p : & \kappa(x, y) = (x^T y)^p \\ \text{le noyau affine :} & \kappa(x, y) = (x^T y + \sigma)^p \\ \text{le noyau gaussien :} & \kappa(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma}\right). \end{array}$$

L'utilisation de fonctions noyaux permet ainsi de calculer implicitement un produit scalaire dans un espace de dimension éventuellement infini par un calcul n'impliquant qu'un nombre fini de termes. Par ailleurs, il apparaît que la spécification de la fonction noyau est suffisante. En effet, à l'instar de (16), la fonction de décision $f(x)$ ne va dépendre que de l'évaluation de la fonction noyau calculée pour un vecteur courant et les vecteurs supports.

Les fonctions noyaux doivent être vues comme des mesures de similarité entre deux vecteurs-observations x et y, mesure qu'il est naturel d'utiliser en induction puisqu'un a priori évident est de supposer que deux vecteurs similaires doivent être associées à des sorties similaires. De fait, les fonctions noyaux peuvent être considérées comme une généralisation des fonctions de covariance.

On peut alors formuler le cas général de l'obtention d'un séparateur à vaste marge.

Soit $\{(x_k, \ell_k), k = 1 \ldots, n\}$ un ensemble de vecteurs-observations étiquetés avec $x_k \in \mathcal{X}$ et $\ell_k \in \{1, -1\}$. Soit $\mathcal{H}$ un espace de Hilbert de noyau positif κ. Un séparateur à vaste marge (SVM) est un discriminateur de la forme : $D(x) = \text{signe}(f(x) + b)$ où $f \in \mathcal{H}$ et $b \in \mathbb{R}$ sont donnés par la résolution du problème suivant pour un $C \geq 0$ donné :

$$\left\{ \begin{array}{l} \min\limits_{f,b,\xi_k} \dfrac{1}{2}||f||^2_{\mathcal{H}} + C \displaystyle\sum_{k=1}^{N} \xi_k \\ \text{sous} \quad \ell_k(f(x_k) + b) \geq 1 - \xi_k, \quad k = 1, \ldots, N \\ \qquad \xi_k \geq 0, \quad k = 1, \ldots, N \end{array} \right. \tag{19}$$

La solution est donnée par :

$$f(x) = \sum_{k \in \mathcal{A}} \alpha_k^* \ell_k \, \kappa(x, x_k) \tag{20}$$

où $\mathcal{A}$ est l'ensemble des indices des vecteurs supports correspondant aux contraintes actives et les α_k sont solutions du programme quadratique suivant :

$$\left\{ \begin{array}{ll} \min\limits_{\alpha} & \dfrac{1}{2}\alpha^T G \alpha - e^T \alpha \\ \text{sous} & 0 \leq \alpha_k \leq C, \quad k = 1, \ldots, N \\ & \ell^T \alpha = 0 \end{array} \right. \tag{21}$$

où G est la matrice $n \times n$ de terme général $G_{ij} = \ell_i \ell_j \, \kappa(x_i, x_j)$. Le biais b a la valeur du multiplicateur de Lagrange de la contrainte d'égalité à l'optimum.

A Annexe 1 : rappel succinct d'optimisation sous contraintes

A.1 Dualité faible et forte

On considère le problème d'optimisation sous contraintes, $\mathcal{P}$, appelé problème primal, suivant :

$$\begin{array}{lll} \min\limits_{x} & f(x) & \text{Fonction à minimiser} \\ \text{sous} & h_i(x) = 0, \quad \forall i = 1, \ldots, n & n \text{ contraintes égalités} \\ & g_j(x) \leq 0, \quad \forall j = 1, \ldots, m & m \text{ contraintes inégalités} \end{array} \tag{22}$$

On suppose ici que $x \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$ est la fonction scalaire à minimiser, et que f, $h_i, i = 1, \ldots, n$ et $g_j, j = 1, \ldots, m$ sont différentiables. Pour résoudre ce problème, on introduit le lagrangien défini par :

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{n} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_i(x) \tag{23}$$

On associe, à chaque contrainte, un paramètre scalaire appelé multiplicateur de Lagrange. Pour chaque contrainte égalité $h_i(x) = 0$, on associe $\lambda_i \in \mathbb{R}$ et pour chaque contrainte inégalité $g_j(x) \leq 0$, on associe $\mu_j \in \mathbb{R}_+$ (c-à-d $\mu_j \geq 0$). L'introduction du lagrangien permet de transformer un problème avec contraintes en un problème sans contrainte avec des variables inconnues supplémentaires λ_i et μ_j. Ces variables jouent le rôle de coefficients de pénalisation de la fonction à optimiser. Elles sont appelées variables duales du problème primal.

Les conditions d'optimalité de la solution sont données par les conditions de Karush-Kuhn-Tucker (KKT) [6].

$$\begin{array}{lll} \text{Stationnarité} & \nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0 \quad \text{c'est-à-dire} & \\ & \nabla_x f(x^*) + \sum_{i=1}^{n} \lambda_i^* \nabla_x h_i(x^*) + \sum_{j=1}^{m} \mu_j^* \nabla_x g_i(x^*) = 0 & \\ \text{Admissibilité primale} & h_i(x^*) = 0, \quad \forall i = 1, \ldots, n & \\ & g_j(x^*) \leq 0, \quad \forall j = 1, \ldots, m & \\ \text{Admissibilité duale} & \mu_j^* \geq 0, \quad \forall j = 1, \ldots, m & \\ \text{Complémentarité} & \mu_j^* g_j(x^*) = 0, \quad \forall j = 1, \ldots, m & \end{array} \tag{24}$$

On peut définir, à partir du lagrangien, la fonction duale $q(\lambda, \mu) = \min_x \mathcal{L}(x, \lambda, \mu)$. Le théorème de la dualité faible établit que la valeur optimale du problème $\mathcal{P}$: $f^* = f(x^*) = \min_x f(x)$ avec $h_i(x^*) = 0, \forall i$ et $g_j(x^*) \leq 0, \forall j$ est telle que :

$$q(\lambda, \mu) \leq f^* \tag{25}$$

Une façon de rendre cette borne inférieure la plus proche possible de la solution optimale f^* est de maximiser $q(\lambda, \mu)$ par rapport aux variables duales λ et μ. On définit alors un second problème d'optimisation $\mathcal{D}$ appelé problème dual :

$$\begin{array}{ll} \max\limits_{\lambda, \mu} & q(\lambda, \mu) \\ \text{sous} & \mu_j \geq 0, \quad \forall j = 1, \ldots, m \end{array} \tag{26}$$

6. Les valeurs optimales (celles qui minimisent le critère), sont notées ici avec une *. Ultérieurement, on pourra utiliser la même notation, lorsque ce n'est pas ambigu, pour les valeurs courantes et optimales des paramètres de Lagrange.
William Karush (dans sa thèse de master en 1939, non publiée) puis Harold W. Kuhn et Albert W. Tucker (Kuhn et Tucker, 1951) ont développé un ensemble de conditions nécessaires à l'optimalité d'un problème d'optimisation sous contraintes.

Si f est convexe, h_i affines et g_j convexes, on a une dualité forte et l'équation (25) devient une égalité. Dans ce cas, les deux problèmes primal et dual ont le même optimum.

Le problème dual est parfois plus aisé à résoudre que le problème primal et sa résolution permet d'obtenir la solution du problème primal. En effet, la résolution du problème dual fournit les valeurs optimales des multiplicateurs de Lagrange.

A.2 Exemple élémentaire

A titre d'exemple, examinons le problème suivant :

$$\begin{array}{ll} \min\limits_{\theta \in \mathbb{R}^2} f(\theta) = & \frac{1}{2}(\theta_1^2 + \theta_2^2) \\ \text{sous} & \theta_1 - 2\theta_2 + 2 \leq 0 \end{array} \tag{27}$$

Le lagrangien s'écrit :

$$\mathcal{L}(\theta, \mu) = \frac{1}{2}(\theta_1^2 + \theta_2^2) + \mu(\theta_1 - 2\theta_2 + 2), \quad \mu \geq 0$$

Les conditions de stationnarité du lagrangien par rapport à θ s'écrivent :

$$\nabla_\theta \mathcal{L}(\theta, \mu) = 0$$

c'est-à-dire :

$$\begin{cases} \theta_1 = -\mu \\ \theta_2 = 2\mu \end{cases} \tag{28}$$

En remplaçant (28) dans la fonction duale $q(\mu) = \min_\theta \mathcal{L}(\theta, \mu)$, on obtient :

$$q(\mu) = \frac{1}{2}(\mu^2 + 4\mu^2) + \mu(-\mu - 4\mu + 2) \tag{29}$$

$$= -\frac{5}{2}\mu^2 + 2\mu \tag{30}$$

On résout alors le problème dual en maximisant, par rapport à μ, $q(\mu)$ avec $\mu \geq 0$. On a :

$$\nabla_\mu q(\mu) = 0 \Rightarrow -5\mu + 2 = 0$$

d'où :

$$\mu = \frac{2}{5}$$

En remplaçant cette valeur dans (28), on obtient :

$$\theta^* = \begin{pmatrix} -\frac{2}{5} & \frac{4}{5} \end{pmatrix}^T$$

Les fonctions objectifs de $\mathcal{P}$, $f(\theta^*)$ et de $\mathcal{D}$, $q(\mu)$, évaluées à leur optimum, sont bien égales : $f(\theta^*) = \frac{1}{2}((-\frac{2}{5})^2 + \frac{4}{5}^2) = \frac{2}{5}$ et $q(\mu) = -\frac{5}{2}\left(\frac{2}{5}\right)^2 + 2\frac{2}{5} = \frac{2}{5}$.

A.3 Dualité de Wolfe

Considérons un problème n'impliquant que des contraintes inégalités :

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{sous} \quad & g_j(x) \leq 0, \quad \forall j = 1, \ldots, m \end{aligned} \tag{31}$$

Le problème dual correspondant s'écrit :

$$\begin{aligned} \max_{\mu} \quad & \min_{x} \left(f(x) + \sum_{j=1}^{m} \mu_j g_j(x) \right) \\ \text{sous} \quad & \mu_j \geq 0, \quad \forall j = 1, \ldots, m \end{aligned} \tag{32}$$

La fonction objectif de ce problème dual est le lagrangien et son minimum par rapport à x interviendra lorsque le gradient de la fonction objectif sera nul. On peut donc formuler le problème d'optimisation selon :

$$\begin{aligned} \max_{\mu, x} \quad & f(x) + \sum_{j=1}^{m} \mu_j g_j(x) \\ \text{sous} \quad & \nabla_x f(x) + \sum_{j=1}^{m} \mu_j \nabla_x g_j(x) = 0 \\ & \mu_i \geq 0, \quad \forall i = 1, \ldots, m \end{aligned} \tag{33}$$

Le problème (33) est appelé dual de Wolfe ; il concerne la maximisation du lagrangien sous les contraintes KKT.

A.4 Le cas particulier des SVM

Le problème primal s'écrit sous la forme :

$$\text{Primal} \quad \begin{cases} \min\limits_{w,b} \dfrac{1}{2}||w||^2 \\ \text{sous } \ell_k(w^T x_k + b) \geq 1, \quad k = 1, \ldots, N \end{cases}$$

Le lagrangien associé s'écrit [7] :

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{k=1}^{N} \alpha_k \left(\ell_k(w^T x_k + b) - 1 \right)$$

Pour formuler le dual de Wolfe, il faut écrire les conditions de stationnarité du Lagrangien par rapport à w et par rapport à b. On a :

$$\frac{\partial \mathcal{L}(w, b, \alpha)}{\partial w} = w - \sum_{k=1}^{N} \alpha_k \ell_k x_k \quad \text{et} \quad \frac{\partial \mathcal{L}(w, b, \alpha)}{\partial b} = \sum_{k=1}^{N} \alpha_k \ell_k$$

7. Attention, dans le cas général, on a des contraintes de type $g_j(x) \leq 0$ avec des paramètres de Lagrange $\mu_j \geq 0$. Ici, il faut se ramener à ce cas et écrire : $-(\ell_k(w^T x_k + b) - 1) \leq 0$

Le dual de Wolfe du problème primal s'écrit donc :

$$\text{Dual de Wolfe} \quad \begin{cases} \max\limits_{w,b,\alpha} & \frac{1}{2}\|w\|^2 - \sum\limits_{k=1}^{N} \alpha_k \left(\ell_k(w^T x_k + b) - 1\right) \\ \text{sous} & \alpha_k \geq 0, \quad k = 1, \ldots, N \\ & w - \sum\limits_{k=1}^{N} \alpha_k \ell_k x_k = 0 \\ & \sum\limits_{k=1}^{N} \alpha_k \ell_k = 0 \end{cases}$$

On peut également le ré-écrire en remplaçant w par sa valeur ($w = \sum_{k=1}^{N} \alpha_k \ell_k x_k$) et en éliminant b. On a en effet :

$$\begin{aligned} \mathcal{L}(w, b, \alpha) &= \frac{1}{2}||w||^2 - \sum_{k=1}^{N} \alpha_k \left(\ell_k(w^T x_k + b) - 1\right) \\ &= \frac{1}{2} \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_j \alpha_i \ell_j \ell_i x_j^T x_i}_{w^T w} - \sum_{i=1}^{N} \alpha_i \ell_i \underbrace{\sum_{j=1}^{N} \alpha_j \ell_j x_j^T}_{w^T} x_i - b \underbrace{\sum_{i=1}^{N} \alpha_i \ell_i}_{=0} + \sum_{i=1}^{N} \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_j \alpha_i \ell_j \ell_i x_j^T x_i + \sum_{i=1}^{N} \alpha_i \end{aligned}$$

d'où :

$$\text{Dual de Wolfe} \quad \begin{cases} \max\limits_{\alpha} & -\frac{1}{2} \sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} \alpha_j \alpha_i \ell_j \ell_i x_j^T x_i + \sum\limits_{i=1}^{N} \alpha_i \\ \text{sous} & \alpha_k \geq 0, \quad k = 1, \ldots, N \\ & \sum\limits_{k=1}^{N} \alpha_k \ell_k = 0 \end{cases}$$

En définissant une matrice symétrique G, de dimension $N \times N$, dont l'élément générique est défini par $G_{i,j} = \ell_j \ell_i x_j^T x_i$ (G est une matrice de Gram) que l'on pourra écrire sous forme matricielle $G = \text{diag}(\ell) X X^T \text{diag}(\ell)$, le dual de Wolfe s'écrit comme un problème d'optimisation quadratique sous contraintes :

$$\mathcal{D} \quad \begin{cases} \min\limits_{\alpha} & \frac{1}{2}\alpha^T G \alpha - e^T \alpha \\ \text{sous} & 0 \leq \alpha_k, \quad k = 1, \ldots, N \\ & \ell^T \alpha = 0 \end{cases}$$

où ℓ est le vecteur des $\ell_k, k = 1, \ldots, N$ et e un vecteur de dimension N ne contenant que des 1.

Sous cette forme, le problème d'optimisation que l'on doit résoudre ne fait intervenir les données d'entrée $x_k, k = 1, \ldots, N$ que sous la forme de leurs produits scalaires $x_j^T x_i, i, j = 1, \ldots, N$ intervenant dans la matrice G. Cette remarque importante sera mise à profit lors de la transposition de la méthode au cas non linéaire.

On notera qu'en résolvant le problème dual $\mathcal{D}$, on détermine les paramètres de Lagrange α du problème primal, mais on a fait disparaître le paramètre b. Pour l'estimer, on peut écrire le dual du dual – ou bi-dual – (qui correspond au problème primal).

Le lagrangien associé à $\mathcal{D}$ s'écrit :

$$\mathcal{L}(\alpha, \beta, \gamma) = \frac{1}{2}\alpha^T G\alpha - e^T\alpha + \beta\ell^T\alpha - \gamma^T\alpha$$

avec $\gamma_i \geq 0$ et son gradient par rapport à α :

$$\nabla_\alpha \mathcal{L}(\alpha, \beta, \gamma) = G\alpha - e + \beta\ell - \gamma$$

Le dual de Wolfe de $\mathcal{D}$ s'écrit donc :

$$\begin{cases} \max\limits_{\beta,\gamma} & \frac{1}{2}\alpha^T G\alpha - e^T\alpha + \beta\ell^T\alpha - \gamma^T\alpha \\ \text{sous} & G\alpha - e + \beta\ell - \gamma = 0 \\ & 0 \leq \gamma_k, \quad k = 1, \ldots, N \end{cases}$$

Ce problème peut être simplifié ; en effet, en pré-multipliant la contrainte égalité par α^T, on a :

$$\alpha^T G\alpha - \alpha^T e + \beta\alpha^T\ell - \alpha^T\gamma = 0$$

or, tous les termes de l'égalité précédente étant scalaires, on a aussi :

$$\alpha^T G\alpha - e^T\alpha + \beta\ell^T\alpha - \gamma^T\alpha = 0$$

On en déduit que la fonction objectif du problème dual peut s'écrire :

$$\frac{1}{2}\alpha^T G\alpha - e^T\alpha + \beta\ell^T\alpha - \gamma^T\alpha = \underbrace{\alpha^T G\alpha - e^T\alpha + \beta\ell^T\alpha - \gamma^T\alpha}_{=0} - \frac{1}{2}\alpha^T G\alpha$$

Finalement, le problème "bi-dual" s'écrit :

$$\begin{cases} \max\limits_{\beta,\gamma} & -\frac{1}{2}\alpha^T G\alpha \\ \text{sous} & G\alpha - e + \beta\ell - \gamma = 0 \\ & 0 \leq \gamma_k, \quad k = 1, \ldots, N \end{cases}$$

or ce problème doit correspondre au problème primal initial. On a vu précédemment que $w = \sum_{k=1}^{N} \alpha_k \ell_k x_k$ que l'on peut aussi écrire en utilisant des notations matricielles :

$$\underbrace{w}_{\in \mathbb{R}^{p\times 1}} = X^T \mathrm{diag}(\ell)\alpha = \underbrace{\begin{pmatrix} x_1 & \ldots & x_N \end{pmatrix}}_{\in \mathbb{R}^{p\times N}} \underbrace{\begin{pmatrix} \ell_1 & & \\ & \ddots & \\ & & \ell_N \end{pmatrix}}_{\in \mathbb{R}^{N\times N}} \underbrace{\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix}}_{\in \mathbb{R}^{N\times 1}}$$

On a ainsi :

$$-\frac{1}{2}\|w\|^2 = -\frac{1}{2}w^T w = -\frac{1}{2}\alpha^T \mathrm{diag}(\ell) X X^T \mathrm{diag}(\ell)\alpha = -\frac{1}{2}\alpha^T G\alpha$$

et l'on a bien $\max_{\beta,\gamma} -\frac{1}{2}\alpha^T G\alpha = \min_w \frac{1}{2}\|w\|^2$.

On peut alors remarquer que l'on a $G\alpha = \text{diag}(\ell)X \underbrace{X^T \text{diag}(\ell)\alpha}_{w} = \text{diag}(\ell)Xw$. La première contrainte du problème précédent s'écrit donc :

$$G\alpha - e + \beta\ell - \gamma = 0 \Leftrightarrow \text{diag}(\ell)Xw - e + \beta\ell - \gamma = 0$$

ou

$$G\alpha - e + \beta\ell - \gamma = 0 \Leftrightarrow \text{diag}(\ell)Xw + \beta\ell - \gamma = e$$

Comme on doit également vérifier $0 \leq \gamma_k, k = 1, \ldots, N$, on a donc :

$$G\alpha - e + \beta\ell - \gamma = 0 \Leftrightarrow \text{diag}(\ell)Xw + \beta\ell \geq e$$

Prises individuellement, ligne par ligne, ces contraintes s'écrivent :

$$\ell_k(x_k^T w + \beta) \geq 1$$

On peut donc comparer les deux formulations du problème primal :

$$\left\{ \begin{array}{l} \min\limits_{w,b} \frac{1}{2}||w||^2 \\ \text{sous } \ell_k(w^T x_k + b) \geq 1, \quad k = 1, \ldots, N \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} \max\limits_{\beta,\gamma} & -\frac{1}{2}\alpha^T G\alpha \\ \text{sous} & G\alpha - e + \beta\ell - \gamma = 0 \\ & 0 \leq \gamma_k \quad k = 1, \ldots, N \end{array} \right.$$

et conclure que le paramètre b correspond à la valeur du multiplicateur de Lagrange de la contrainte égalité du problème dual.

A.5 Le cas particulier des SVM à marge souple

Le problème primal s'écrit sous la forme :

$$\text{Primal} \quad \left\{ \begin{array}{ll} \min\limits_{w,b} & \frac{1}{2}||w||^2 + C\sum\limits_{k=1}^{N} \xi_k \\ \text{sous} & \ell_k(w^T x_k + b) \geq 1 - \xi_k, \quad k = 1, \ldots, N \\ & \xi_k \geq 0, \quad k = 1, \ldots, N \end{array} \right.$$

Le lagrangien associé s'écrit :

$$\mathcal{L}(w, b, \alpha, \beta) = \frac{1}{2}||w||^2 + C\sum_{k=1}^{N} \xi_k - \sum_{k=1}^{N} \alpha_k \left(\ell_k(w^T x_k + b) - 1 + \xi_k\right) - \sum_{k=1}^{N} \beta_k \xi_k$$

avec des multiplicateurs de Lagrange $\alpha_k \geq 0$ et $\beta_k \geq 0$. Le calcul des dérivées partielles du lagrangien par rapport à w et b est identique au cas précédent. La dérivée partielle par rapport à la variable d'écart ξ_k s'écrit :

$$\frac{\partial \mathcal{L}(w, b, \alpha, \beta)}{\partial \xi_k} = C - \alpha_k - \beta_k$$

La condition de stationnarité $C - \alpha_k - \beta_k = 0$ conduit donc, en tenant compte de la positivité de β_k, à $\alpha_k \leq C$.

Comme précédemment, on peut écrire le dual de Wolfe dont le critère est le lagrangien précédent. On observe qu'il diffère du précédent de la quantité :

$$C\sum_{k=1}^{N}\xi_k - \sum_{k=1}^{N}\alpha_k\xi_k - \sum_{k=1}^{N}\beta_k\xi_k$$

Dans cette somme, le facteur multiplicatif de ξ_k est $C-\alpha_k-\beta_k$ qui doit être nul. Le critère à optimiser est donc le même que dans le cas des marges "dures". L'unique différence est l'apparition d'une borne supérieure sur les $\alpha_k, k=1,\ldots,N$.

$$\mathcal{D}\quad\begin{cases}\min\limits_{\alpha} & \frac{1}{2}\alpha^T G\alpha - e^T\alpha\\ \text{sous} & 0\leq\alpha_k\leq C,\quad k=1,\ldots,N\\ & \ell^T\alpha = 0\end{cases}$$

B Annexe 2 : matrice de Gram pour un noyau gaussien

B.1 Première approche

La présentation des calculs est inspirée de *Creating a radial basis function kernel matrix in matlab*[8].

On considère une matrice de données $X \in \mathbb{R}^{N\times p}$ constituée de vecteurs x_i pour $i = 1, \ldots, N$:

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}$$

On cherche à calculer la matrice de Gram dans le cas de l'utilisation d'un noyau Gaussien, soit :

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

La technique s'appuie sur le fait que l'on veut calculer la matrice de terme générique $K_{ij} = \kappa(x_i, x_j) = f(\|x_i - x_j\|^2)$ de façon efficace. Les calculs matriciels sont basés sur des produits scalaires (des multiplications de vecteurs) et pas sur la norme de la différence de deux vecteurs. Si l'on ne veut pas utiliser de boucles (et c'est préférable en Matlab), il faut trouver comment exprimer $\|x_i - x_j\|^2$ en utilisant des opérations matricielles. On a ainsi :

$$\begin{aligned} \|x_i - x_j\|^2 &= (x_i - x_j)^T(x_i - x_j) \\ &= x_i^T x_i - x_i^T x_j - x_j^T x_i + x_j^T x_j \\ &= ||x_i||^2 - 2x_i^T x_j + ||x_j||^2 \end{aligned}$$

On peut tout d'abord calculer le vecteur $nms \in \mathbb{R}^{1\times N}$ des normes au carré de chaque vecteur (*normsquare*) :

$$nms = \begin{pmatrix} \|x_1\|^2 & \|x_2\|^2 & \ldots & \|x_N\|^2 \end{pmatrix}$$

```
nms = sum(X'.^2);
```

On peut ensuite commencer à créer la matrice de Gram de dimension $N \times N$ en utilisant le produit de nms par un vecteur de 1 :

$$\begin{pmatrix} \|x_1\|^2 \\ \|x_2\|^2 \\ \vdots \\ \|x_N\|^2 \end{pmatrix} \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} = \begin{pmatrix} \|x_1\|^2 & \|x_1\|^2 & \ldots & \|x_1\|^2 \\ \|x_2\|^2 & \|x_2\|^2 & \ldots & \|x_2\|^2 \\ \vdots & \vdots & \vdots & \vdots \\ \|x_N\|^2 & \|x_N\|^2 & \ldots & \|x_N\|^2 \end{pmatrix}$$

```
nms'*ones(1,N)
```

qui est le vecteur des $\|x_i\|^2$ et, de façon analogue, le vecteur des $\|x_j\|^2$ s'obtient avec le produit suivant :

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} \|x_1\|^2 & \|x_2\|^2 & \ldots & \|x_N\|^2 \end{pmatrix} = \begin{pmatrix} \|x_1\|^2 & \|x_2\|^2 & \ldots & \|x_N\|^2 \\ \|x_1\|^2 & \|x_2\|^2 & \ldots & \|x_N\|^2 \\ \vdots & \vdots & \vdots & \vdots \\ \|x_1\|^2 & \|x_2\|^2 & \ldots & \|x_N\|^2 \end{pmatrix}$$

8. https://stackoverflow.com/questions/37362258/creating-a-radial-basis-function-kernel-matrix-in-matlab/4

`ones(N,1)*nms`

On peut aussi calculer :

$$2XX^T = 2\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}\begin{pmatrix} x_1 & x_2 & \dots & x_N \end{pmatrix} = \begin{pmatrix} x_1^T x_1 & x_1^T x_2 & \dots & x_1^T x_N \\ x_2^T x_1 & x_2^T x_2 & \dots & x_2^T x_N \\ \vdots & \vdots & \vdots & \vdots \\ x_N^T x_1 & x_N^T x_2 & \dots & x_N^T x_N \end{pmatrix}$$

et finalement calculer la matrice de Gram selon la décomposition écrite précédemment :

```
K = exp(-(nms'*ones(1,N) + ones(N,1)*nms - 2*X*X')/(2*sigma^2));
```

Le calcul de la matrice de Gram de l'ensemble des vecteurs $x_i, i = 1, \dots, N$ de la matrice X s'effectue donc selon :

```
% Calcul de la matrice de Gram dans le cas d'un noyau gaussien
% X : matrice de données, sigma : dispersion du noyau
[N,p] = size(X);
nms = sum(X'.^2);
K = exp(-(nms'*ones(1,N) + ones(N,1)*nms - 2*X*X')/(2*sigma^2));
```

Lors de l'évaluation des résultats de la classification, on peut également être appelé à calculer une pseudo-matrice de Gram K construite sur la base de deux matrices :

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} \in \mathbb{R}^{N\times p}, \qquad Y = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_M^T \end{pmatrix} \in \mathbb{R}^{M\times p}$$

telle que $K = [K_{i,j}]$ avec $K_{i,j} = \kappa(x_i, y_j) = \exp\left(-\frac{\|x_i - y_j\|^2}{2\sigma^2}\right)$, $i = 1, \dots, N$ et $j = 1, \dots, M$.

Le calcul de cette matrice s'effectue selon le programme suivant :

```
N = size(X,1);
M = size(Y,1);
nms1 = sum(X'.^2);
nms2 = sum(Y'.^2);
K = exp(-(nms1'*ones(1,M) + ones(N,1)*nms2 - 2*X*Y')/(2*sigma^2));
```

B.2 Seconde approche (identique en termes de méthode de calcul)

A partir de ML_Toolbox [9] : `knGauss`.

On considère les deux matrices (attention, transposées des précédentes)

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_N \end{pmatrix} \in \mathbb{R}^{p\times N}, \qquad Y = \begin{pmatrix} y_1 & y_2 & \dots & y_M \end{pmatrix} \in \mathbb{R}^{p\times M}$$

On calcule le vecteur des produits scalaires des vecteurs colonnes par eux-mêmes (à l'aide de la fonction `dot`.

9. https://github.com/epfl-lasa/ML_toolbox dans ML_toolbox/methods/clustering/knkmeans/

L'instruction `dot(X,X,1)'` donne :

$$a = \begin{pmatrix} x_1^T x_1 \\ x_2^T x_2 \\ \dots \\ x_N^T x_N \end{pmatrix} = \begin{pmatrix} \|x_1\|^2 \\ \|x_2\|^2 \\ \dots \\ \|x_N\|^2 \end{pmatrix}$$

et l'instruction `dot(Y,Y,1)`

$$b = \begin{pmatrix} y_1^T y_1 & y_2^T y_2 & \dots & y_M^T y_M \end{pmatrix} = \begin{pmatrix} \|y_1\|^2 & \|y_2\|^2 & \dots & \|y_M\|^2 \end{pmatrix}$$

L'utilisation de la fonction Matlab `bsxfun` (*Binary Singleton Expansion Function*) avec la fonction `@plus` (`bsxfun(@plus,dot(X,X,1)',dot(Y,Y,1))`) permet de générer la matrice :

$$\begin{pmatrix} \|x_1\|^2 + \|y_1\|^2 & \|x_1\|^2 + \|y_2\|^2 & \dots & \|x_1\|^2 + \|y_M\|^2 \\ \|x_2\|^2 + \|y_1\|^2 & \|x_2\|^2 + \|y_2\|^2 & \dots & \|x_2\|^2 + \|y_M\|^2 \\ \vdots & \vdots & \vdots & \vdots \\ \|x_N\|^2 + \|y_1\|^2 & \|x_N\|^2 + \|y_2\|^2 & \dots & \|x_N\|^2 + \|y_M\|^2 \end{pmatrix}$$

Il suffit ensuite d'y retrancher les doubles produits $2(X^T * Y)$:

```
D = bsxfun(@plus,dot(X,X,1)',dot(Y,Y,1))-2*(X'*Y);
```

pour obtenir :

$$D = \begin{pmatrix} \|x_1\|^2 + \|y_1\|^2 - 2x_1^T y_1 & \|x_1\|^2 + \|y_2\|^2 - 2x_1^T y_2 & \dots & \|x_1\|^2 + \|y_M\|^2 - 2x_1^T y_M \\ \|x_2\|^2 + \|y_1\|^2 - 2x_2^T y_1 & \|x_2\|^2 + \|y_2\|^2 - 2x_2^T y_2 & \dots & \|x_2\|^2 + \|y_M\|^2 - 2x_2^T y_M \\ \vdots & \vdots & \vdots & \vdots \\ \|x_N\|^2 + \|y_1\|^2 - 2x_N^T y_1 & \|x_N\|^2 + \|y_2\|^2 - 2x_N^T y_2 & \dots & \|x_N\|^2 + \|y_M\|^2 - 2x_N^T y_M \end{pmatrix}$$

Il reste ensuite à calculer l'exponentielle :

```
K = exp(D/(-2*s^2));
```

B.3 Code Matlab de calcul de la matrice de Gram pour un noyau gaussien

```
function K = kernelgauss(X,sigma,Y)
%KERNELGAUSS  Compute the Gram matrix with Gaussian kernel
%
%   KERNELGAUSS(X,sigma) Compute the matrix with generic element
%   K(i,j)=exp(-|| x_i - x_j ||^2/(2*sigma^2))
%   The matrix K is square with dimension the number of rows of X
%
%   KERNELGAUSS(X,sigma,Y) Compute the matrix with generic element
%   K(i,j)=exp(-|| x_i - y_j ||^2/(2*sigma^2))
%   The matrix K has as many rows as X and as many columns as the number of
%   rows of Y

% Didier Maquin 02/2019

if nargin < 3
    Y=X;
end;
[N1,p1]=size(X);
[N2,p2]=size(Y);

if p1 ~= p2
    error('Error: The number of columns of the first argument must be...
     identical as the number of columns of the third one');
end

nms1 = sum(X'.^2);
nms2 = sum(Y'.^2);
K = exp(-(nms1'*ones(1,N2) + ones(N1,1)*nms2 - 2*X*Y')/(2*sigma^2));

% end kernelgauss
```

B.4 Code Matlab alternatif

```
function K = knGauss(X, Y, s)
% Gaussian (RBF) kernel K = exp(-|x-y|/(2s));
% Input:
%   X: d x nx data matrix
%   Y: d x ny data matrix
%   s: sigma of gaussian
% Ouput:
%   K: nx x ny kernel matrix
% Written by Mo Chen (sth4nth@gmail.com).
if nargin < 3
    s = 1;
end

if nargin < 2 || isempty(Y)
    K = ones(1,size(X,2));            % norm in kernel space
else
    D = bsxfun(@plus,dot(X,X,1)',dot(Y,Y,1))-2*(X'*Y);
    K = exp(D/(-2*s^2));
end
```

La première ligne de commentaire devrait sans doute être écrite :

```
% Gaussian (RBF) kernel K = exp(-||x-y||^2/(2s^2));
```

C Annexe 3 : propriété du noyau gaussien

Montrons que l'utilisation d'un noyau Gaussien correspond à la projection des observations dans un espace de dimension infinie ; ce qui signifie que l'on ne pourra pas analyser les $\phi(x_i)$. Pour simplifier les notations des calculs suivants, considérons deux vecteurs d'observation que l'on note $x \in \mathbb{R}^m$ et $y \in \mathbb{R}^m$, on doit avoir :

$$\begin{aligned}\phi^T(x)\phi(y) = \kappa(x,y) &= \exp\left(-\frac{\|x-y\|^2}{2}\right)\\ \kappa(x,y) &= \exp\left(-\frac{\|x\|^2}{2}\right)\exp\left(-\frac{\|y\|^2}{2}\right)\exp\left(x^T y\right)\\ &= \exp\left(-\frac{\|x\|^2}{2}\right)\exp\left(-\frac{\|y\|^2}{2}\right)\sum_{k=0}^{\infty}\frac{(x^T y)^k}{k!}\end{aligned}$$

Considérons le premier terme du membre de droite. On a $e^a = \left(e^{\frac{a}{k}}\right)^k$ d'où :

$$\exp\left(-\frac{\|x\|^2}{2}\right) = \exp\left(-\frac{\|x\|^2}{2k}\right)^k$$

et $\frac{1}{k!} = \frac{1}{\sqrt{k!}}\frac{1}{\sqrt{k!}} = \left(\frac{1}{\sqrt{k!}^{1/k}}\frac{1}{\sqrt{k!}^{1/k}}\right)^k$, d'où :

$$k(x,y) = \sum_{k=0}^{\infty}\left(\frac{\exp\left(-\frac{\|x\|^2}{2k}\right)}{\sqrt{k!}^{1/k}}\frac{\exp\left(-\frac{\|y\|^2}{2k}\right)}{\sqrt{k!}^{1/k}}x^T y\right)^k$$

Il faut maintenant développer $(x^T y)^k$; on a :

$$(x^T y)^k = (x_1 y_1 + x_2 y_2 + \ldots + x_m y_m)^k$$

On peut utiliser la formule du multinôme de Newton [10] :

$$(a_1 + a_2 + \cdots + a_m)^k = \sum_{n_1+n_2+\ldots+n_m=k}\binom{k}{n_1, n_2, \ldots, n_m}a_1^{n_1}a_2^{n_2}\ldots a_m^{n_m}$$

La somme porte sur toutes les combinaisons d'indices entiers naturels $n_1, \ldots, n_m$ tels que $n_1 + n_2 + \cdots + n_m = k$ certains d'entre eux pouvant être nuls. On peut alors écrire :

$$\kappa(x,y) = \sum_{k=0}^{\infty}\sum_{\sum_i n_i = k}\frac{\exp\left(-\frac{\|x\|^2}{2k}\right)}{\sqrt{k!}^{1/k}}\binom{k}{n_1, \ldots, n_m}x_1^{n_1}\ldots x_m^{n_m}\frac{\exp\left(-\frac{\|y\|^2}{2k}\right)}{\sqrt{k!}^{1/k}}\binom{k}{n_1, \ldots, n_m}y_1^{n_1}\ldots y_m^{n_m}$$

On peut alors définir un vecteur $\phi(x)$ de dimension infinie, tels que sa composante de rang k, $\phi_k(x)$ soit :

$$\phi_k(x) = \frac{\exp\left(-\frac{\|x\|^2}{2k}\right)}{\sqrt{k!}^{1/k}}\binom{k}{n_1, \ldots, n_m}x_1^{n_1}\ldots x_m^{n_m} \quad / \sum_i n_i = k$$

et l'on a alors :

$$\phi^T(x)\phi(y) = \kappa(x,y)$$

10. https://fr.wikipedia.org/wiki/Formule_du_multinôme_de_Newton)

Références

[1] Carraro L., Badea A. *Notions sur l'analyse discriminante*. Polycopié d'analyse discriminante, Ecole des Mines de Saint-Etienne. http://carraro.fr/documents/regression/07_08_AnalyseDiscriminante.pdf

[2] Cornillon P.A. *Analyse discriminante linéaire (au sens de Fisher ou LDA)*. Polycopié d'analyse données, département "Mathématiques Appliquées et Sciences Sociales", Université de Rennes 2.

[3] Cornuéjols A., Miclet L., Barra V. *Apprentissage artificiel. Deep learning, concepts et algorithmes* 3ème édition, Eyrolles, mai 2018.

[4] Lebart L., Morineau A., Fénelon J.P. *Traitement des données statistiques - méthodes et programmes*. Bordas, Paris, 1979.

[5] Martin A. *L'analyse de données*. Polycopié d'analyse données, ENSIETA, septembre 2004. http://www.arnaud.martin.free.fr/Doc/polyAD.pdf

[6] Wikistat *Machines à vecteurs supports*. Projet Wikistat, avril 2019. http://wikistat.fr/pdf/st-m-app-svm.pdf

D Analyse discriminante sur les iris de Fisher

D.1 Programme Matlab®

```
% Necessite la boîte à outils stats
% Utilisation du tracé à l'aide de gplotmatrix et boxplot

% Chargement du fichier de données
load fisheriris

% Tracés divers
figure(1);
gplotmatrix(meas,meas,species)

figure(2);
for i=1:4
     h=subplot(2,2,i)
     boxplot(h,meas(:,i),species)
end

% Matrice de variance-covariance totale des données
T=cov(meas);

% Matrices de variance-covariance de chaque classe
cov1=cov(meas(1:50,:));
cov2=cov(meas(51:100,:));
cov3=cov(meas(101:150,:));

% Calcul de la variance inter-classe par différence
B=T-(cov1+cov2+cov3)/3;

% Indicateurs statistiques élémentaires
t1=sqrt(diag(T));
t2=sqrt(diag(cov1));
t3=sqrt(diag(cov2));
t4=sqrt(diag(cov3));
t5=t2./t1;
t6=t3./t1;
t7=t4./t1;
t8=mean([t5 t6 t7],2);
t=[t1 t2 t3 t4 t5 t6 t7 t8];
disp('      s       s_set     s_ver     s_vir    s_set/s   s_ver/s   s_vir/s    moyenne')
disp(t)

% Calcul de la variance des centres de gravité
m1=mean(meas(1:50,:));
m2=mean(meas(51:100,:));
m3=mean(meas(101:150,:));
B2=cov([mean1 ; mean2 ; mean3]);
rank(B2) % Rang égal au nombre de classes - 1
```

```
% B2 et B différent car cov est un estimateur non biaisé
% en remplaçant cov(x) par cov(x,1) estimateur biaisé, on a égalité

% Décomposition en valeurs propres et vecteurs propres
[V,D]=eig(inv(T)*B)

% Rangement par ordre décroissant
[Dmax,ordre]=sort(diag(D),'descend')
V=V(:,ordre)

% Sélection des deux premiers vecteurs (1er plan factoriel)
VV=V(:,1:2);

% Projection des données dans le premier plan factoriel
Proj=meas*VV;

% Tracé avec marqueur selon les classes
figure(2)
gplotmatrix(Proj(:,1),Proj(:,2),species)

% Extension de la matrice de données avec les termes quadratiques
 measnew=[...
     meas ...
     meas(:,1).*meas(:,2)...
     meas(:,1).*meas(:,3)...
     meas(:,1).*meas(:,4)...
     meas(:,2).*meas(:,3)...
     meas(:,2).*meas(:,4)...
     meas(:,3).*meas(:,4)...
     meas(:,1).^2 ...
     meas(:,2).^2 ...
     meas(:,3).^2 ...
     meas(:,4).^2]

 % Traitement
 T=cov(measnew);
 cov1=cov(measnew(1:50,:));
 cov2=cov(measnew(51:100,:));
 cov3=cov(measnew(101:150,:));
 B=T-(cov1+cov2+cov3)/3;
 [V,D]=eig(inv(T)*B)
 [Dmax,ordre]=sort(diag(D),'descend')
 V=V(:,ordre)
 VV=V(:,1:2);
 Proj=measnew*VV;
 figure(3)
 gplotmatrix(Proj(:,1),Proj(:,2),species)
```

E SVM à noyau gaussien

E.1 Programme Matlab®

```
clear
close all
set(0,'defaultAxesFontSize',14)
rand('seed',2);

%% Génération des données
N = 50;
r = 0.55+0.4*rand(N,1);
theta = rand(N,1)*pi/2;
X1 = [r.*cos(theta) r.*sin(theta)];
%
r = 0.65*rand(N,1);
theta = rand(N,1)*pi/2;
X2 = [r.*cos(theta) r.*sin(theta)];

X = [X1;X2];
l = [ones(N,1) ; -ones(N,1)];
n = 2*N;

figure(1);
h1 = plot(X(l==1,1),X(l==1,2),'+r'); hold on
set(h1,'LineWidth',2);
h2 = plot(X(l==-1,1),X(l==-1,2),'db');
set(h2,'LineWidth',2);
theta2 = 0:0.01:pi/2;
plot(0.6*cos(theta2),0.6*sin(theta2),'--')
title('Façon de générer les données')
axis([0 1 0 1]);
legend('Classe 1 : 0.55 < r < 1', 'Classe 2 : r < 0.65')

%print('SVM_1','-depsc','-r0')

%% Visualisation des données d'entrée
figure(2);
h1 = plot(X(l==1,1),X(l==1,2),'+r'); hold on
set(h1,'LineWidth',2);
h2 = plot(X(l==-1,1),X(l==-1,2),'db');
set(h2,'LineWidth',2);
title('Données initiales')
axis([0 1 0 1]);

% print('SVM_2','-depsc','-r0')

%% Résolution du dual avec CVX

C = 10000; % Facteur de pénalisation des variables ressorts
sigma = 3; % Portée du noyau 0.08 ou 0.09
```

```
K = kernelgauss(X,sigma);
G = (l*l').*K;
e = ones(n,1);

cvx_begin
   cvx_precision best
   cvx_quiet(true)
   variable alph(n)
   dual variables de dp dC
   minimize( 1/2*alph'*G*alph - e'*alph )
   subject to
      de : l'*alph == 0;
      dp : alph >= 0;
      dC : alph <= C;
cvx_end

pos = find(alph > 1e-4); % Indices des vecteurs suppports
alpha_supp = alph(pos);
% b est le paramètre de Lagrange de la contrainte égalité au signe près
b = -de;

%% Visualisation de la classification

[xtest1 xtest2] = meshgrid([0:.01:1],[0:0.01:1]); % Grille d'évaluation

nn = length(xtest1); nn2=nn*nn;
% Crée une matrice de nn2 observations de R^2
Xgrid = [reshape(xtest1,nn*nn,1) reshape(xtest2,nn*nn,1)];
% Evalue une pseudo matrice de Gram entre ces observations et les vecteurs
% supports
Kgrid = kernelgauss(Xgrid,sigma,X(pos,:));
% Evalue la fonction de décision pour les points de la grille
ypred = Kgrid*(l(pos).*alpha_supp) + b;
% Recrée une matrice de dimension nn x nn dont chaque élément est la valeur
% de la fonction de décision
ypred = reshape(ypred,nn,nn);

figure(3);
hold on
[cc,hh] = contour(xtest1,xtest2,ypred,[0 0],'k');
box on
clabel(cc,hh); % Trace la frontière de décision
set(hh,'LineWidth',2);
[cc,hh] = contour(xtest1,xtest2,ypred,[-1 -1],'k');
clabel(cc,hh); % Trace la marge inf
set(hh,'LineWidth',.5);
[cc,hh] = contour(xtest1,xtest2,ypred,[1 1],'k');
clabel(cc,hh); % Trace la marge sup
set(hh,'LineWidth',.5);
```

```
h1 = plot(X(l==1,1),X(l==1,2),'+r');
set(h1,'LineWidth',2);
h2 = plot(X(l==-1,1),X(l==-1,2),'db');
set(h2,'LineWidth',2);
Xsup = X(pos,:);
h3 = plot(Xsup(:,1),Xsup(:,2),'ok');
set(h3,'LineWidth',2);
label = ['Résultat de la classification avec sigma = ' num2str(sigma)];
title(label)
legend('0 : Séparatrice','-1 : Marge inf',' 1 : Marge sup','Classe 1',...
       'Classe 2','Supports')
axis([0 1 0 1]);
% print('SVM_3','-depsc','-r0')
```

E.2 Résultat de l'éxécution

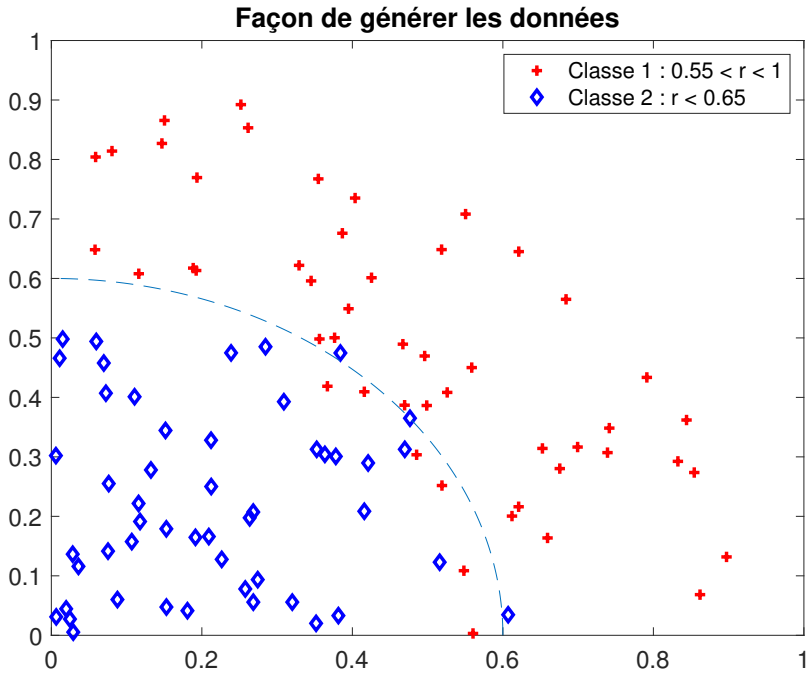


FIGURE 21 – Les données de la classe n° 1 sont dans le quadrant positif dans une couronne dont le cercle intérieur a comme diamètre 0.55 et comme diamètre extérieur 1. Les données de la classe n° 2 sont dans ce même quadrant et dans un disque de diamètre égal à 0.65. Ce choix permet l'"interpénétration" des observations (elles ne sont pas séparables par l'arc de cercle de rayon 0.6).

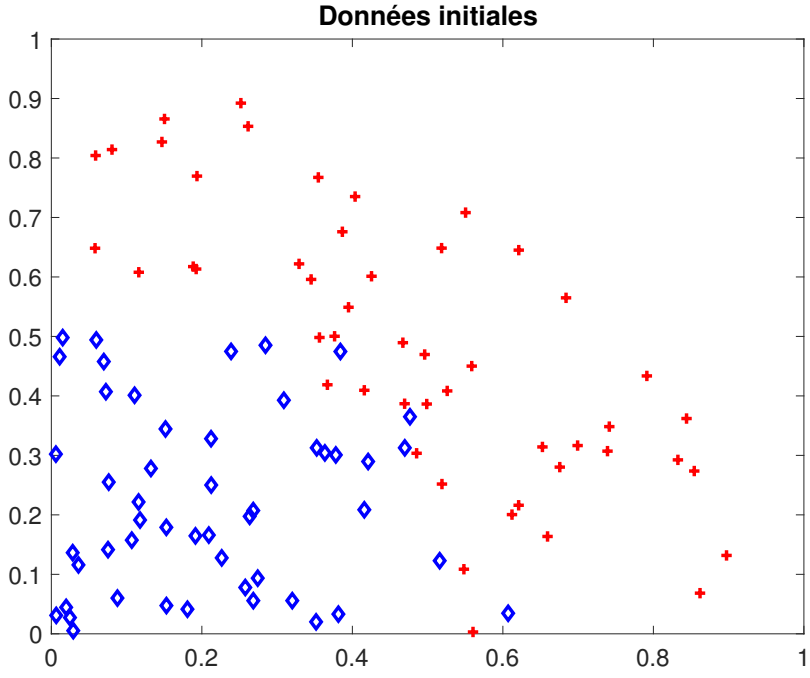


FIGURE 22 – Données initiales (sans indication sur la façon dont elles ont été générées)

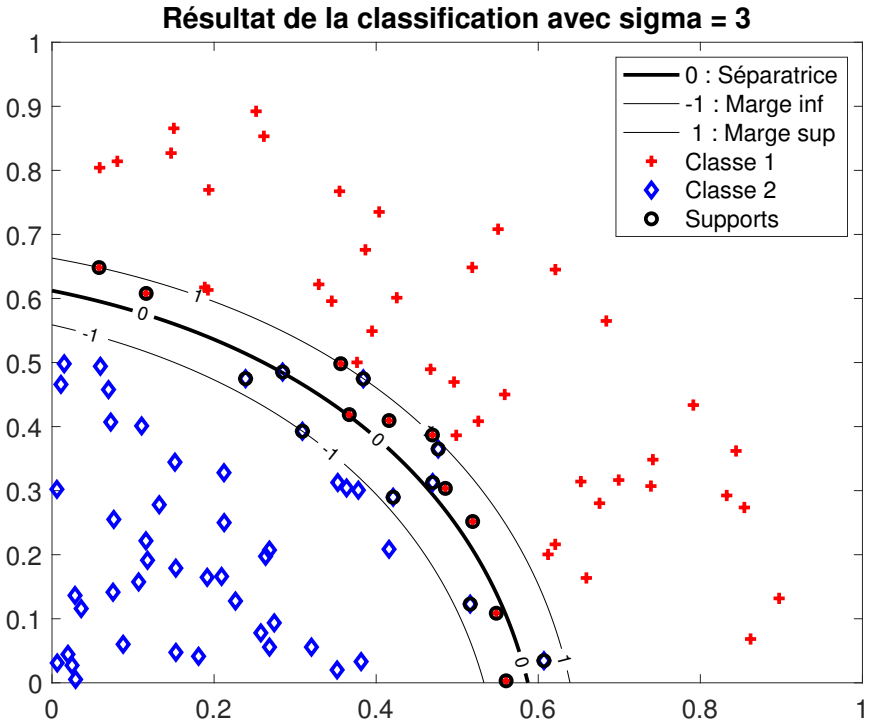


FIGURE 23 – Résultat de la classification : on retrouve quasiment ici, avec ces réglages, la séparatrice utilisée pour la création du jeu de données

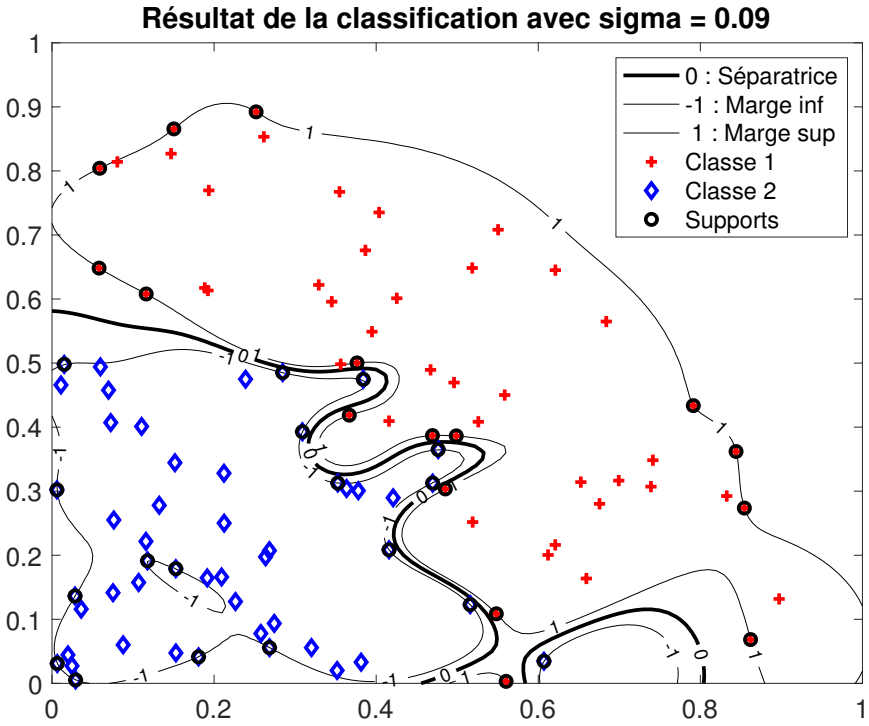


FIGURE 24 – Résultat de la classification : situation de sur-apprentissage. Les données des classes 1 et 2 sont parfaitement séparées mais la capacité de généralisation de ce classifieur est médiocre

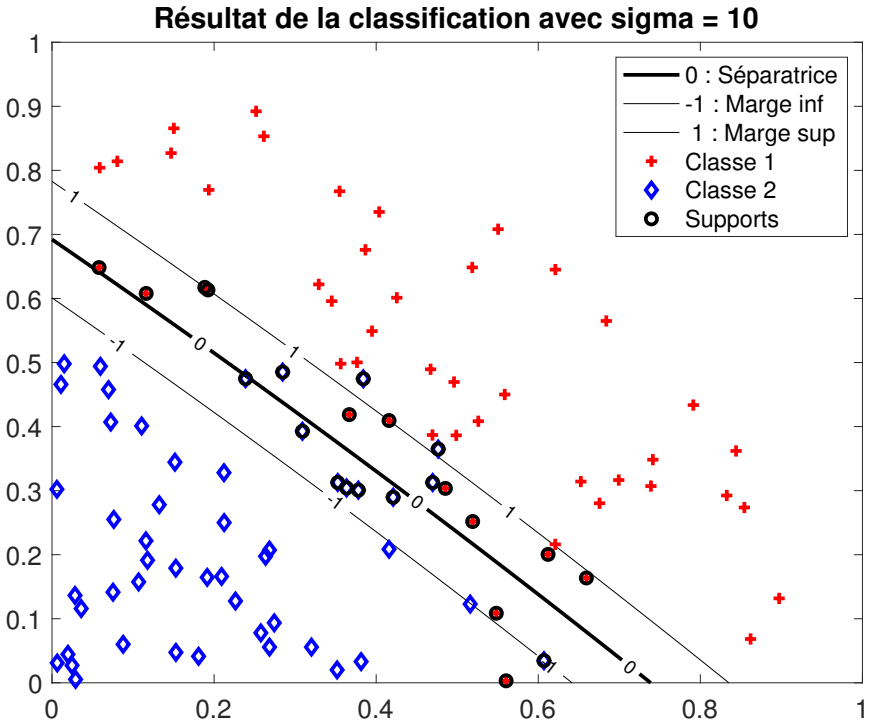


FIGURE 25 – Résultat de la classification : en augmentant la portée du noyau gaussien, on tend vers un classifieur linéaire