Méthodes d'analyse de données

Eléments de classification de données

Didier Maquin

Ecole Nationale Supérieure d'Electricité et de Mécanique

Octobre 2020

Plan

Plan

Plan

Plan

Les iris de Fisher

Les données dites "iris de Fisher" sont issues d'une étude du botaniste Anderson et ont été utilisées en 1937 par le célèbre statisticien Sir Ronald Fisher.

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.

Les iris de Fisher

Les données dites "iris de Fisher" sont issues d'une étude du botaniste Anderson et ont été utilisées en 1937 par le célèbre statisticien Sir Ronald Fisher.

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.

Iris *Setosa*

Iris *Versicolor*

Iris *Virginica*

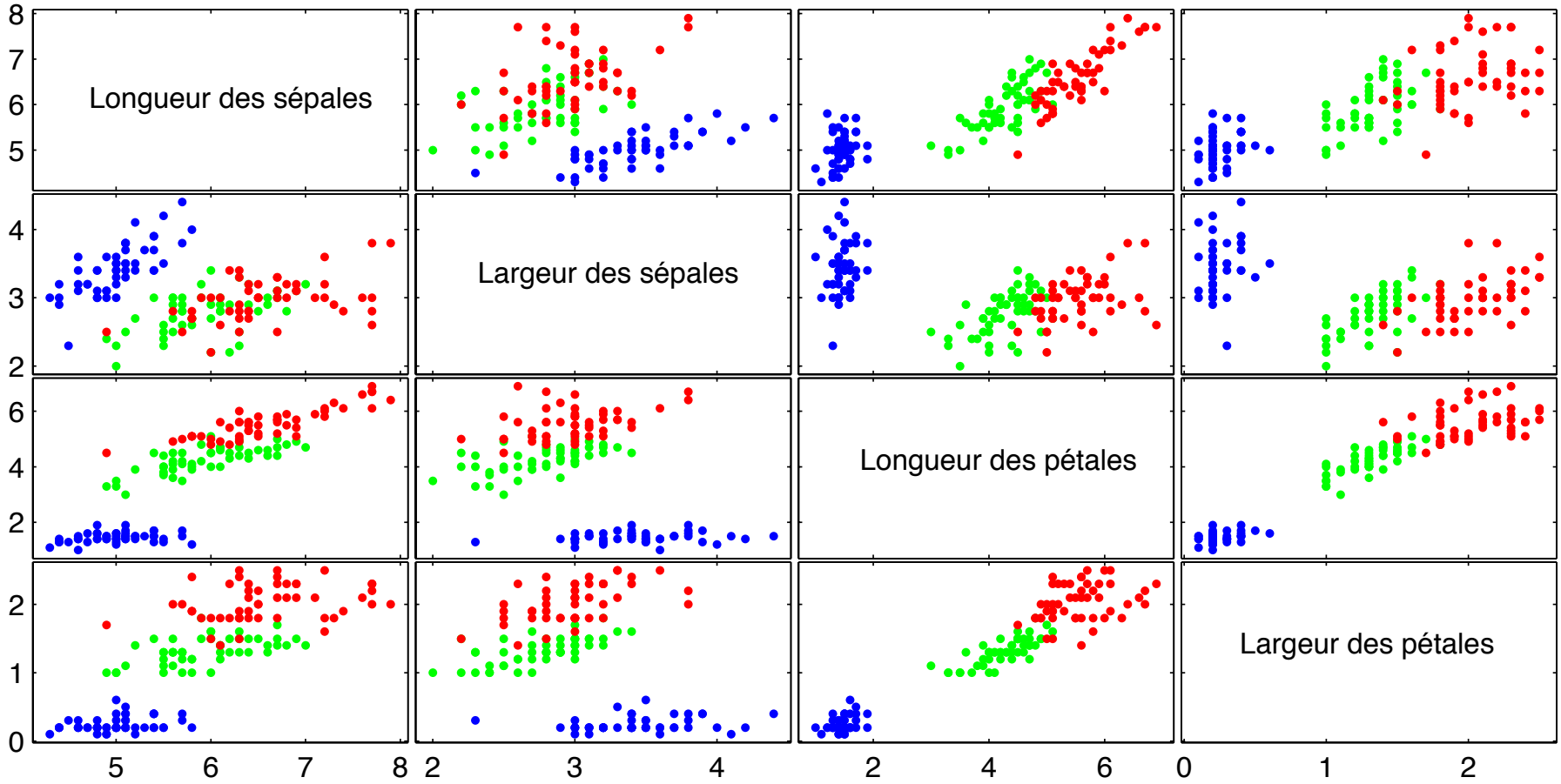
Les iris de Fisher

Deux questions peuvent être posées :

- **Analyse discriminante descriptive** : quelles variables, quels groupes de variables, quels sous-espaces discriminent-ils au mieux les 3 espèces d'iris ?

- **Analyse discriminante décisionnelle** : comment affecter une nouvelle fleur à une espèce en connaissant les valeurs des 4 variables quantitatives qui la décrivent.

Analyse discriminante



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 :

- 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 $\{i, Y(i) = y_h\}$.

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 :

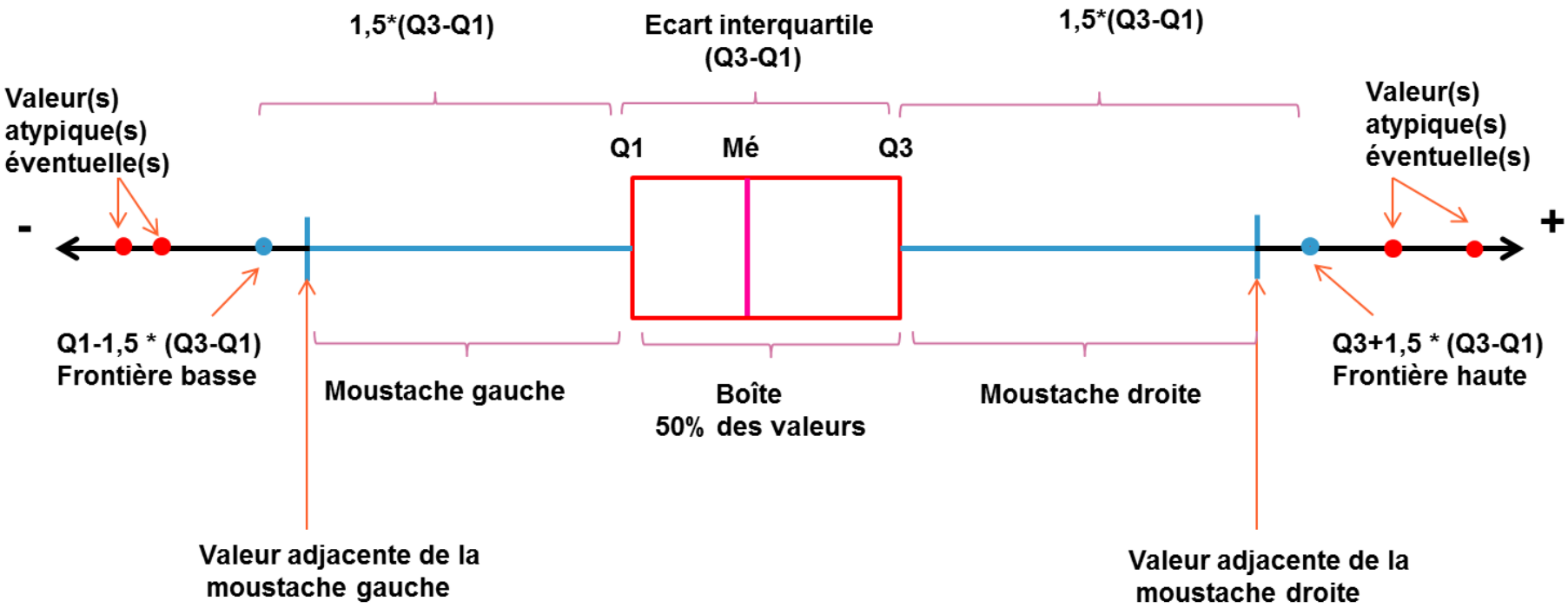
- p variables X_j, le plus souvent quantitatives ;
- l'appartenance à un groupe qui se traduit par une variable qualitative Y possédant m modalités $y_h, 1 \leq h \leq m$

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

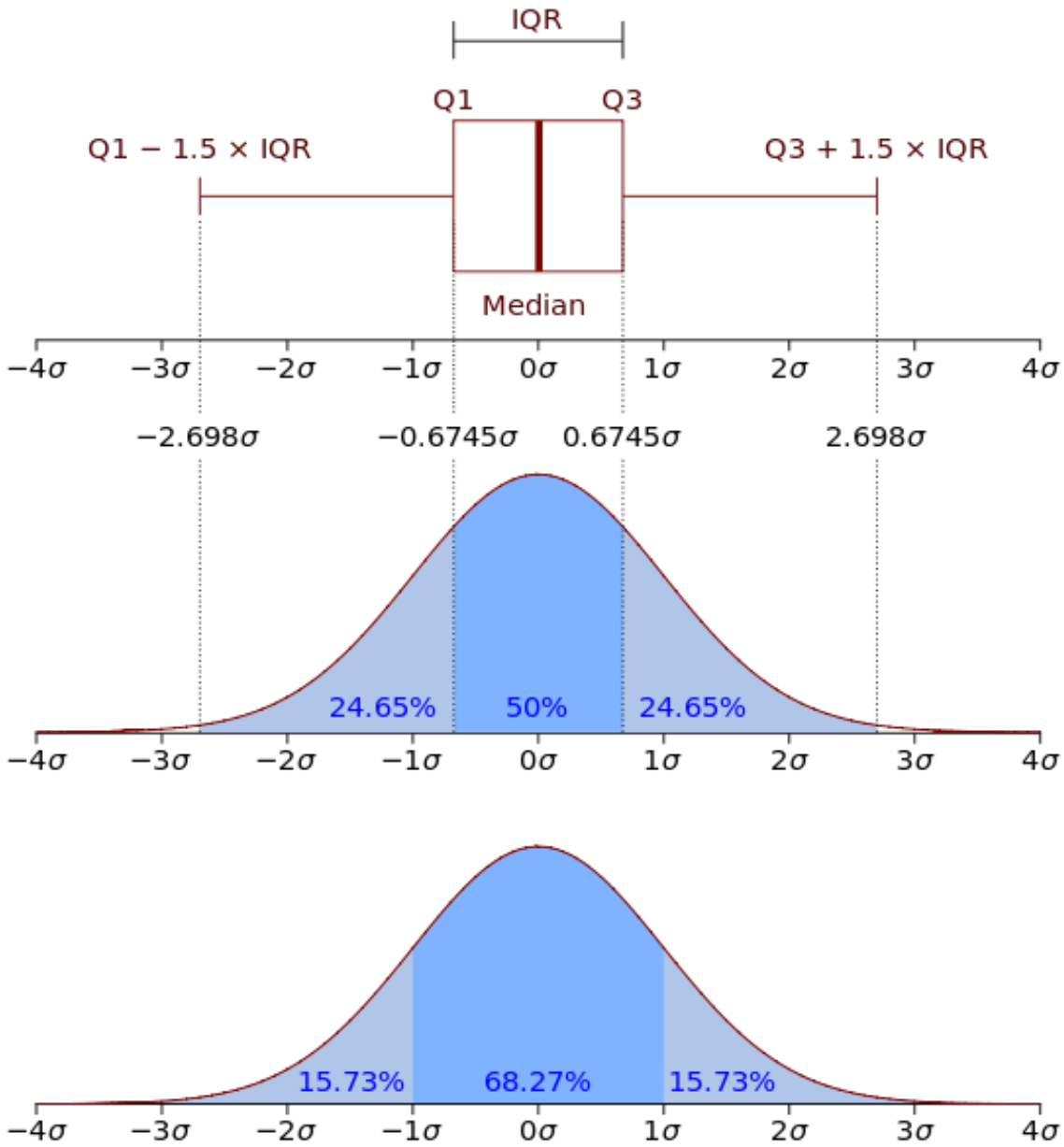
Indicateurs statistiques

	Longueur sépale	Largeur sépale	Longueur pétale	Largeur pétale
Minimum	4.30	2.00	1.00	0.10
1^{er} 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
$3^{ème}$ quartile	6.40	3.30	5.10	1.80
Maximum	7.90	4.40	6.90	2.50

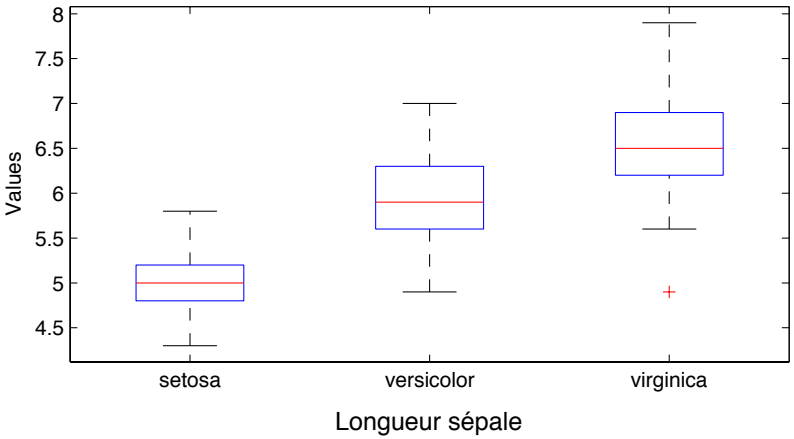
Rappel : boîtes à moustaches ou *boxplot*



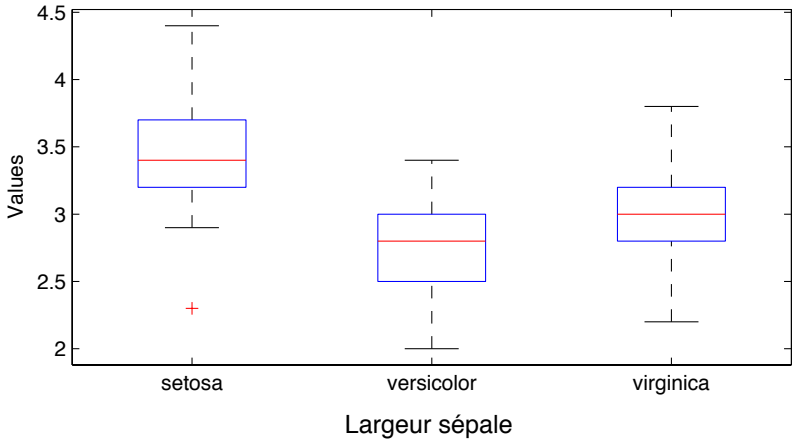
Rappel : boîtes à moustaches ou *boxplot*



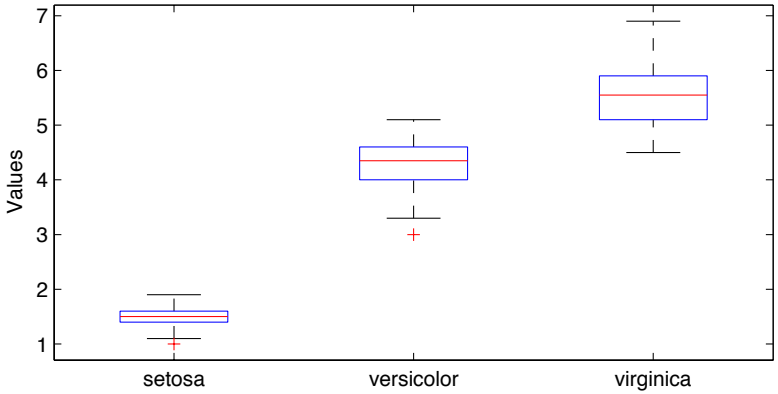
Analyse discriminante



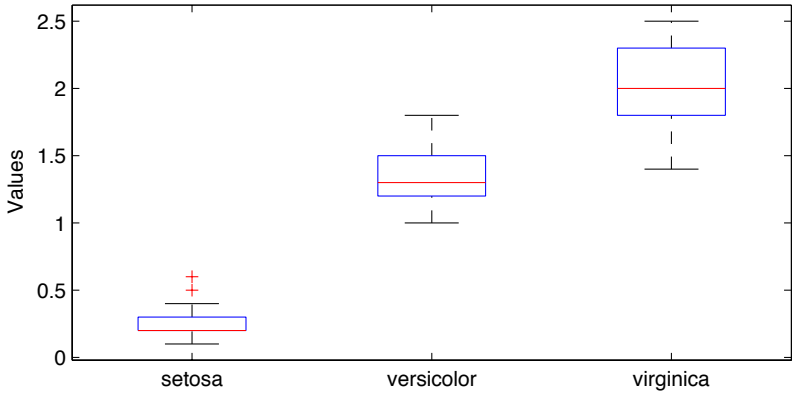
Longueur sépale



Largeur sépale



Longueur pétale



Largeur pétale

Pouvoir discriminant et variance

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

σ désigne l'écart-type de la variable considérée
σ_{set} désigne l'écart-type de la variable observée pour l'espèce *setosa*
σ_{ver} désigne l'écart-type de la variable observée pour l'espèce *versicolor*
σ_{vir} désigne l'écart-type de la variable observée pour l'espèce *virginica*
mean représente la moyenne des 3 rapports σ_{set}/σ, σ_{ver}/σ et σ_{vir}/σ.

Pouvoir discriminant et variance

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

σ désigne l'écart-type de la variable considérée
σ_{set} désigne l'écart-type de la variable observée pour l'espèce *setosa*
σ_{ver} désigne l'écart-type de la variable observée pour l'espèce *versicolor*
σ_{vir} désigne l'écart-type de la variable observée pour l'espèce *virginica*
mean représente la moyenne des 3 rapports σ_{set}/σ, σ_{ver}/σ et σ_{vir}/σ.

- le pouvoir discriminant d'une variable est lié à la façon dont la loi de la variable considérée est concentrée
- plus la quantité *mean* est faible, plus la variable est utile pour discriminer les espèces

Modèle probabiliste

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 $(1 \leq h \leq m)$.

Modèle probabiliste

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 $(1 \leq h \leq m)$.

Théorème de la variance conditionnelle

$$\text{Var}(X_j) = \text{Esp}(\text{Var}(X_j \mid Y)) + \text{Var}(\text{Esp}(X_j \mid Y))$$

Modèle probabiliste

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 $(1 \leq h \leq m)$.

Théorème de la variance conditionnelle

$$\text{Var}(X_j) = \text{Esp}(\text{Var}(X_j \mid Y)) + \text{Var}(\text{Esp}(X_j \mid Y))$$

Expressions de l'espérance et de la variance conditionnelles

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

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

Modèle probabiliste

Considérons la première variable X_1 la longueur sépale.

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

$$\text{Esp}(\text{Var}(\text{longueur sépale} \mid Y)) = \sigma^2_{set} P(Y = setosa) + \sigma^2_{ver} P(Y = versicolor) + \sigma^2_{vir} P(Y = virginica)$$

du fait que les 3 espèces comportent chacune 50 individus :

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

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

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

L'indicateur *mean* est égal, aux carrés près, au quotient $\text{Esp}(\text{Var}(X \mid Y)/\text{Var}(X)$. Ce rapport est compris entre 0 et 1 et la variable X sera d'autant plus discriminante pour Y qu'il est petit.

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

- $\text{Var}(X) = \mathbf{T}$: matrice de covariance empirique totale
- $\text{Esp}(\text{Var}(X \mid Y)) = \mathbf{W}$: matrice de covariance intra-classes
- $\text{Var}(\text{Esp}(X \mid Y)) = \mathbf{B}$: matrice de covariance inter-classes

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

- $\text{Var}(X) = \mathbf{T}$: matrice de covariance empirique totale
- $\text{Esp}(\text{Var}(X \mid Y)) = \mathbf{W}$: matrice de covariance intra-classes
- $\text{Var}(\text{Esp}(X \mid Y)) = \mathbf{B}$: matrice de covariance inter-classes

Indices de Sobol

$\text{Var}(\text{Esp}(X \mid Y))/\text{Var}(X)$ est appelé indice de Sobol. Il est, comme l'indice *mean*, compris entre 0 et 1, mais plus il est proche de 1, plus la variable X est discriminante.

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

- $\text{Var}(X) = \mathbf{T}$: matrice de covariance empirique totale
- $\text{Esp}(\text{Var}(X \mid Y)) = \mathbf{W}$: matrice de covariance intra-classes
- $\text{Var}(\text{Esp}(X \mid Y)) = \mathbf{B}$: matrice de covariance inter-classes

Indices de Sobol

$\text{Var}(\text{Esp}(X \mid Y))/\text{Var}(X)$ est appelé indice de Sobol. Il est, comme l'indice *mean*, compris entre 0 et 1, mais plus il est proche de 1, plus la variable X est discriminante.

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

Table – Indices de Sobol pour les données d'iris

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

- $g = \text{Esp}(X)$: centre de gravité du nuage complet
- $g^h = \text{Esp}(X \mid Y = y_h)$: centre de gravité du groupe G_h

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

- $g = \text{Esp}(X)$: centre de gravité du nuage complet
- $g^h = \text{Esp}(X \mid Y = y_h)$: centre de gravité du groupe G_h

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

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

- $g = \text{Esp}(X)$: centre de gravité du nuage complet
- $g^h = \text{Esp}(X \mid Y = y_h)$: centre de gravité du groupe G_h

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

On a :

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

et donc :

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

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

- $g = \text{Esp}(X)$: centre de gravité du nuage complet
- $g^h = \text{Esp}(X \mid Y = y_h)$: centre de gravité du groupe G_h

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

On a :

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

et donc :

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

Notons, pour $i \in G_h$, $p(i \mid Y = y_h)$ la probabilité conditionnelle $p_i \mid P(Y = y_h)$:

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

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

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

$\text{Esp}(\text{Var}(X \mid Y)) = \mathbf{W}$: moyenne, pondérée par leur probabilité, des matrices de covariance dans chaque groupe G_h.

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

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

$\text{Esp}(\text{Var}(X \mid Y)) = \mathbf{W}$: moyenne, pondérée par leur probabilité, des matrices de covariance dans chaque groupe G_h.

Le vecteur $\text{Esp}(X \mid Y)$ prend m valeurs $g^1, \ldots, g^m$ avec comme probabilités respectives $P(Y = y_1), \ldots, P(Y = y_m)$. Son espérance vaut $g = \text{Esp}(X)$ et sa matrice de covariance se réduit donc à :

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

$\text{Var}(\text{Esp}(X \mid Y)) = \mathbf{B}$: matrice de covariance du nuage des centres de gravité $g^1, \ldots, g^m$ affectés de leur probabilité.

Modèle probabiliste

$$\text{Var}(X) = \text{Esp}(\text{Var}(X \mid Y)) + \text{Var}(\text{Esp}(X \mid Y))$$

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

$\text{Esp}(\text{Var}(X \mid Y)) = \mathbf{W}$: moyenne, pondérée par leur probabilité, des matrices de covariance dans chaque groupe G_h.

Le vecteur $\text{Esp}(X \mid Y)$ prend m valeurs $g^1, \ldots, g^m$ avec comme probabilités respectives $P(Y = y_1), \ldots, P(Y = y_m)$. Son espérance vaut $g = \text{Esp}(X)$ et sa matrice de covariance se réduit donc à :

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

$\text{Var}(\text{Esp}(X \mid Y)) = \mathbf{B}$: matrice de covariance du nuage des centres de gravité $g^1, \ldots, g^m$ affectés de leur probabilité.

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

$$\textit{Total} = \textit{Within} + \textit{Between}$$

Maximisation de l'indice de Sobol

On cherche à identifier parmi toutes les combinaisons linéaires de variables celle qui a l'indice de Sobol le plus important. Soit $\beta = (\beta_1, \ldots, \beta_p)^T$ le vecteur des coefficients de la combinaison linéaire :

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

Maximisation de l'indice de Sobol

On cherche à identifier parmi toutes les combinaisons linéaires de variables celle qui a l'indice de Sobol le plus important. Soit $\beta = (\beta_1, \ldots, \beta_p)^T$ le vecteur des coefficients de la combinaison linéaire :

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

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

$$\text{Var}(\text{Esp}(\beta^T X \mid Y)) = \text{Var}(\beta^T \text{Esp}(X \mid Y)) = \beta^T \text{Var}(\text{Esp}(X \mid Y))\beta = \beta^T \mathbf{B} \beta$$

$$\text{Var}(\beta^T X) = \beta^T \text{Var}(X)\beta = \beta^T \mathbf{T} \beta$$

Maximisation de l'indice de Sobol

On cherche à identifier parmi toutes les combinaisons linéaires de variables celle qui a l'indice de Sobol le plus important. Soit $\beta = (\beta_1, \ldots, \beta_p)^T$ le vecteur des coefficients de la combinaison linéaire :

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

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

$$\text{Var}(\text{Esp}(\beta^T X \mid Y)) = \text{Var}(\beta^T \text{Esp}(X \mid Y)) = \beta^T \text{Var}(\text{Esp}(X \mid Y))\beta = \beta^T \mathbf{B} \beta$$

$$\text{Var}(\beta^T X) = \beta^T \text{Var}(X)\beta = \beta^T \mathbf{T} \beta$$

On en déduit que l'indice de Sobol vaut :

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

Maximisation de l'indice de Sobol

On en déduit que l'indice de Sobol vaut :

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

Maximisation de l'indice de Sobol

On en déduit que l'indice de Sobol vaut :

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

Sa maximisation conduit à :

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

soit :

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

ou encore :

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

Maximisation de l'indice de Sobol

On en déduit que l'indice de Sobol vaut :

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

Sa maximisation conduit à :

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

soit :

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

ou encore :

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

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 $\mathbf{T}^{-1}\mathbf{B}$.

Axes factoriels

Le calcul des vecteurs propres et valeurs propres de la matrice $\mathbf{T}^{-1}\mathbf{B}$ donne :

$\beta_1 = \begin{pmatrix} 0.2087 & 0.3862 & -0.5540 & -0.7074 \end{pmatrix}^T$ avec $S(\beta_1) = 0.9695$ et

$\beta_2 = \begin{pmatrix} 0.0065 & 0.5866 & -0.2526 & -0.7695 \end{pmatrix}^T$ avec $S(\beta_2) = 0.2114$.

Axes factoriels

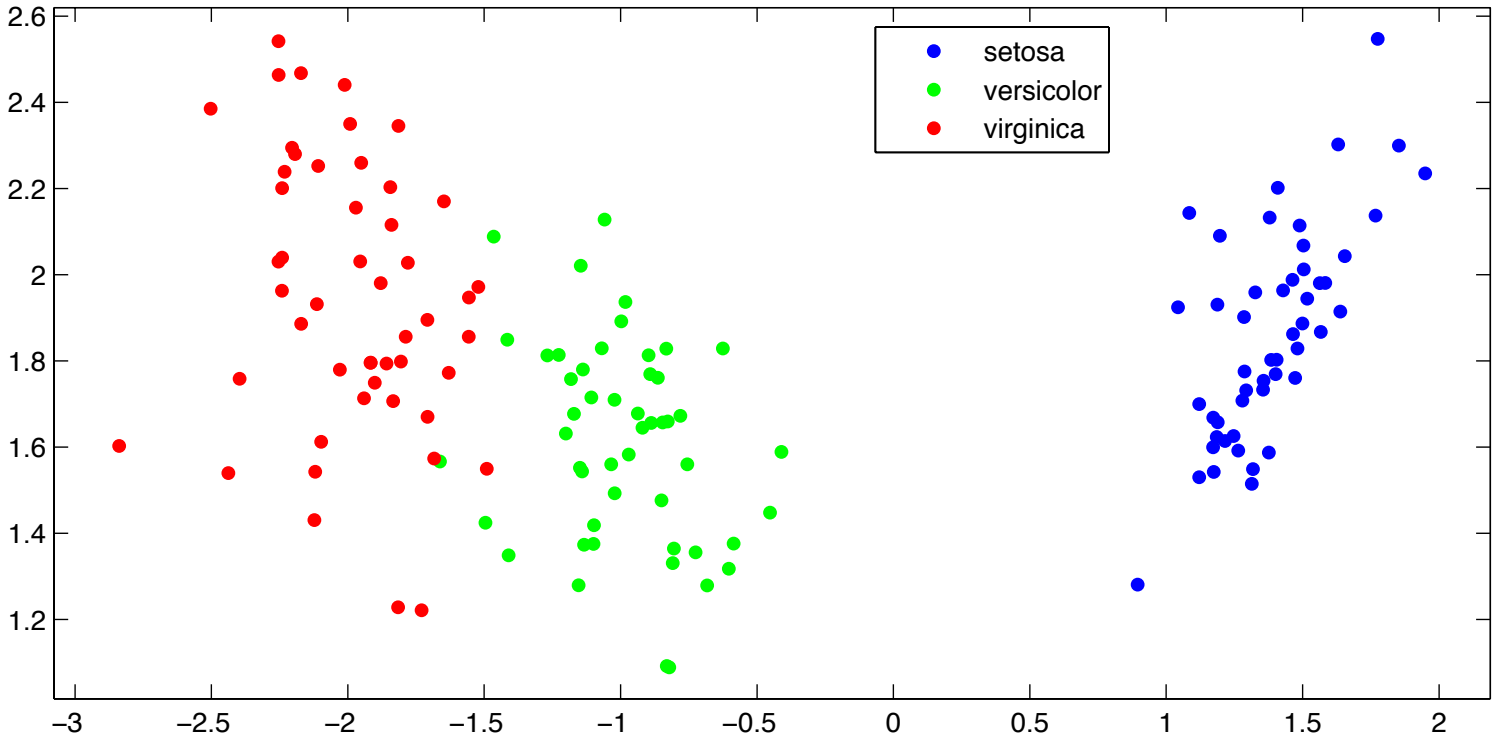
Le calcul des vecteurs propres et valeurs propres de la matrice $\mathbf{T}^{-1}\mathbf{B}$ donne :

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

Projection sur l'espace des variables discriminantes



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*.

Extension de la matrice de données

Notons X_{aug} le vecteur :

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

Ce vecteur appartient à l'espace quadratique associé aux variables initiales.

Extension de la matrice de données

Notons X_{aug} le vecteur :

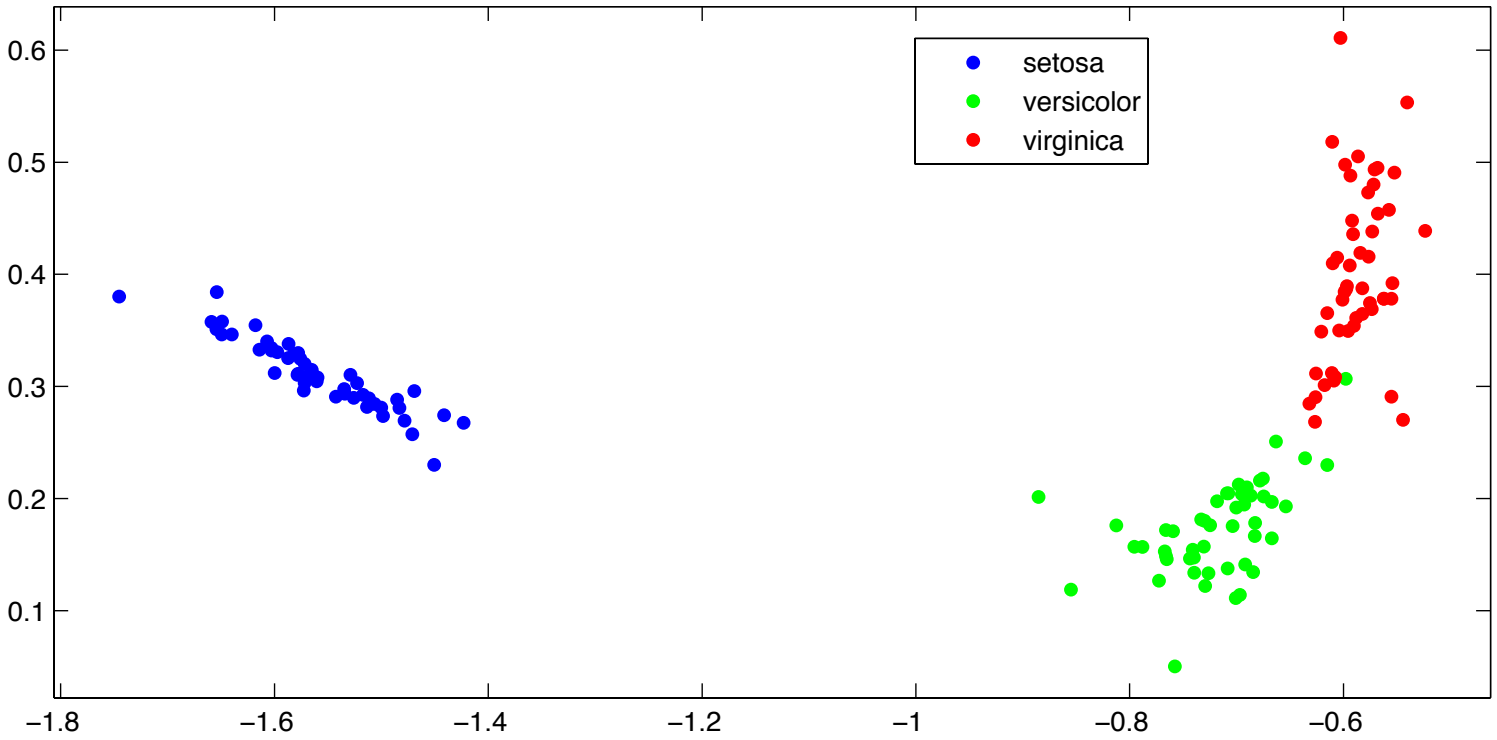
$$X_{aug} = \begin{pmatrix} X_1 & X_2 & \dots & X_p & X_1X_2 & X_1X_3 & \dots & X_{p-1}X_p & X_1^2 \dots & X_p^2 \end{pmatrix}^T$$

Ce vecteur appartient à l'espace quadratique associé aux variables initiales.

Les deux valeurs propres les plus grandes (correspondant aux indices de Sobol) sont $S(\beta_1) = 0.9864$ et $S(\beta_2) = 0.7466$.

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

Projection sur l'espace des variables discriminantes



Analyse discriminante décisionnelle

Objet

Etant 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 ?

Objet

Etant 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 ?

Démarche

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.

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

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

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

Objet

Etant 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 ?

Démarche

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.

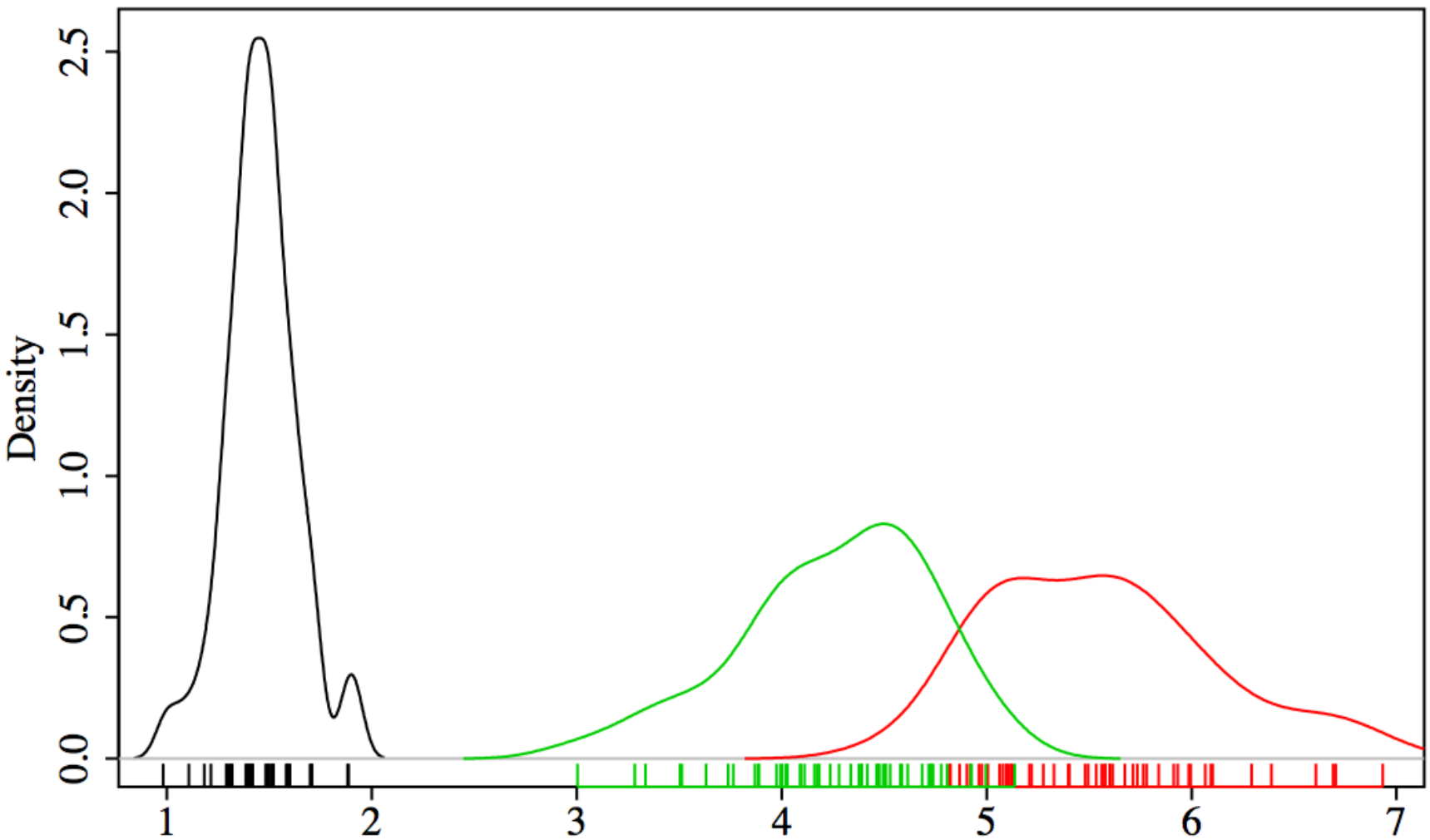
$$\|x - g^h\|^2_{\mathbf{W}^{-1}} = (x - g^h)^T \mathbf{W}^{-1} (x - g^h)$$

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

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

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 $l_h = l_k, h \neq k$.

Approche probabiliste



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

Hypothèse

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 :

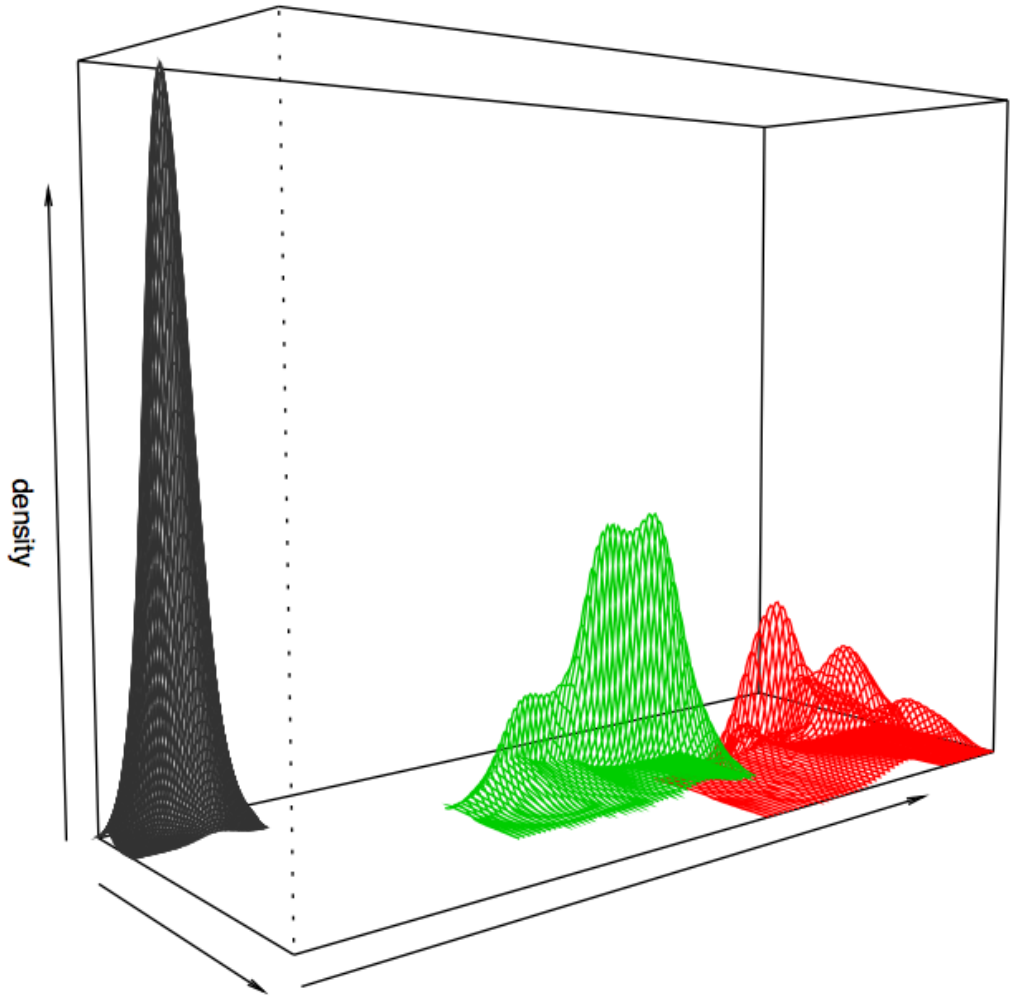
$$f(X, Y = setosa) = \frac{1}{\sqrt{2\pi}\sigma_{set}} \exp\left(-\frac{1}{2\sigma_{set}^2}(X - \mu_{set})^2\right)$$

de même pour la seconde et la troisième espèce :

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

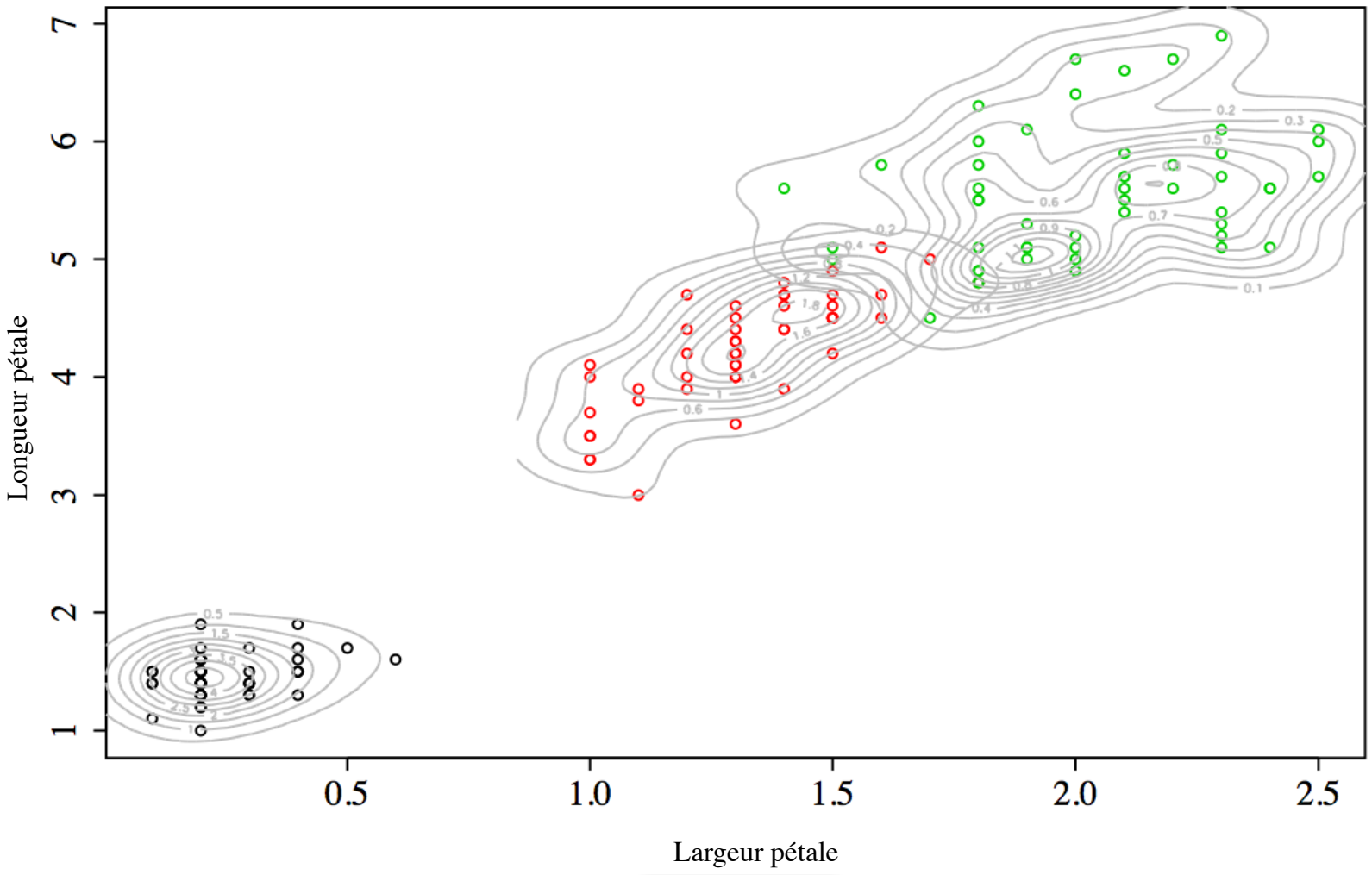
$$f(X, Y = virginica) = \frac{1}{\sqrt{2\pi}\sigma_{vir}} \exp\left(-\frac{1}{2\sigma_{vir}^2}(X - \mu_{vir})^2\right)$$

Approche probabiliste



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

Approche probabiliste



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

Théorème de Bayes

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

Théorème de Bayes

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

Pour une nouvelle observation $X = x_0$:

$$P(Y = setosa \mid X = x_0) = \frac{f(x_0 \mid Y = setosa)P(Y = setosa)}{f(x_0)}$$

$$P(Y = versicolor \mid X = x_0) = \frac{f(x_0 \mid Y = versicolor)P(Y = versicolor)}{f(x_0)}$$

$$P(Y = virginica \mid X = x_0) = \frac{f(x_0 \mid Y = virginica)P(Y = virginica)}{f(x_0)}$$

Théorème de Bayes

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

Pour une nouvelle observation $X = x_0$:

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

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

Description

La probabilité conditionnelle de l'appartenance au groupe G_j sachant x_0 s'écrit :

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

Description

La probabilité conditionnelle de l'appartenance au groupe G_j sachant x_0 s'écrit :

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

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

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 : $f(X \mid Y = y_j) \approx \mathcal{N}(\mu_j, \Sigma)$ (hypothèse d'homoscédasticité).

Description

La probabilité conditionnelle de l'appartenance au groupe G_j sachant x_0 s'écrit :

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

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

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 : $f(X \mid Y = y_j) \approx \mathcal{N}(\mu_j, \Sigma)$ (hypothèse d'homoscédasticité).

Probabilité d'affectation

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

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 $\Sigma_j \in \mathbb{R}^{p \times p}$.

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 $\Sigma_j \in \mathbb{R}^{p \times p}$.

Centre de gravité des groupes :

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

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 $\Sigma_j \in \mathbb{R}^{p \times p}$.

Centre de gravité des groupes :

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

Variances (discriminante quadratique)

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

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 $\Sigma_j \in \mathbb{R}^{p \times p}$.

Centre de gravité des groupes :

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

Variances (discriminante quadratique)

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

Variance (discriminante linéaire)

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

Interprétation géométrique

Considérons $x_0 \in \mathbb{R}^2$, un point du plan. Les frontières des classes sont définies par :

$$f(x_0 \mid Y = y_j) = f(x_0 \mid Y = y_j')$$

Interprétation géométrique

Considérons $x_0 \in \mathbb{R}^2$, un point du plan. Les frontières des classes sont définies par :

$$f(x_0 \mid Y = y_j) = f(x_0 \mid Y = y_j')$$

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

c'est-à-dire :

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

Interprétation géométrique

Considérons $x_0 \in \mathbb{R}^2$, un point du plan. Les frontières des classes sont définies par :

$$f(x_0 \mid Y = y_j) = f(x_0 \mid Y = y_j')$$

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

c'est-à-dire :

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

On obtient une équation quadratique en x et y qui permet de dire qu'une frontière sera de la forme d'une conique.

Interprétation géométrique

Considérons $x_0 \in \mathbb{R}^2$, un point du plan. Les frontières des classes sont définies par :

$$f(x_0 \mid Y = y_j) = f(x_0 \mid Y = y_j')$$

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

c'est-à-dire :

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

On obtient une équation quadratique en x et y qui permet de dire qu'une frontière sera de la forme d'une conique.

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

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

On obtient l'équation d'une droite.

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

Supposons que $m = 3$ et que $\Sigma = \Sigma_1 = \Sigma_2 = \Sigma_3 = I_2$. Les observations suivent toutes des lois normales $\mathcal{N}(\mu_j, I_2)$. La frontière entre le groupe 1 et le groupe 2 est donc :

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

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

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

Supposons que $m = 3$ et que $\Sigma = \Sigma_1 = \Sigma_2 = \Sigma_3 = I_2$. Les observations suivent toutes des lois normales $\mathcal{N}(\mu_j, I_2)$. La frontière entre le groupe 1 et le groupe 2 est donc :

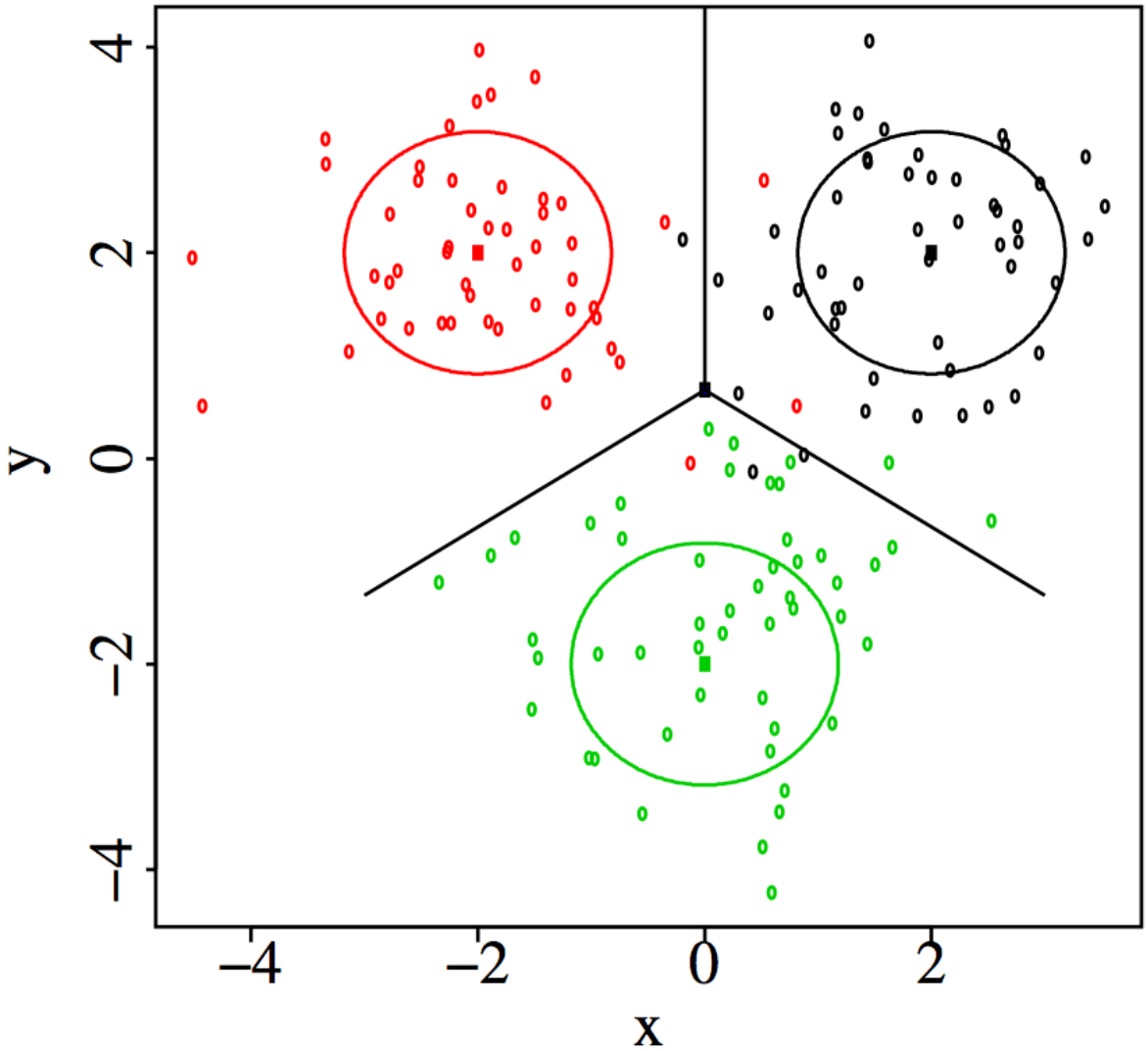
$$x_0^T(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T(\mu_1 - \mu_2) = 0$$

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

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 :

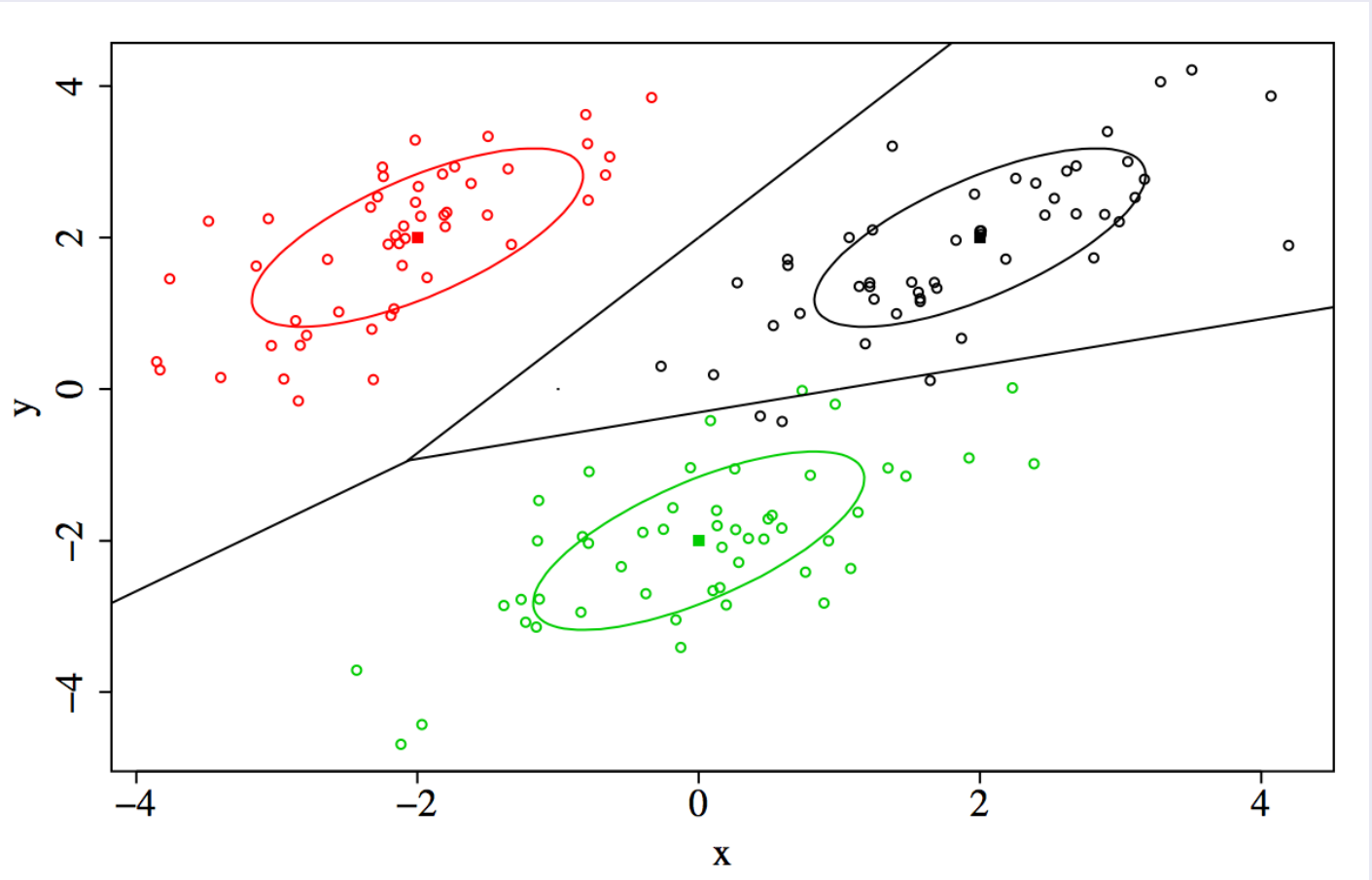
$$< \overrightarrow{G_{12}M}, \overrightarrow{G_2G_1} >= 0$$

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).


4
2
0
−2
−4
y
−4
−2
0
2
x

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

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.



Plan

Classification automatique

Objet

Produire des groupements d'objets ou d'individus décrits par un certain nombre de variables ou de caractères

Classification automatique

Objet

Produire des groupements d'objets ou d'individus décrits par un certain nombre de variables ou de caractères

Résultats attendus

Partitions ou hiérarchie de partitions

- arbres (au sens de la théorie des graphes) où les sommets seront les objets à classer
- classes empiétantes
- zones à forte densité

Démarche

Les techniques de classification font appel à une démarche algorithmique. Il existe diverses familles d'algorithmes :

- algorithmes ascendants (ou encore agglomératifs) qui procèdent à la construction des classes par agglomération successive des objets
- algorithmes descendants (ou encore divisifs) qui procèdent par dichotomies successives de l'ensemble des objets
- algorithmes conduisant directement à des partitions comme les méthodes d'aggrégation autours de centres mobiles

Présentation

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

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 [a]. 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 ;

a. 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

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 [a]. 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.

---

a. 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

Calcul de distance

Si x, y et z sont trois objets et si x et y sont regroupés en un seul élément noté h, examinons la façon dont on peut définir la distance de ce groupement à z

Calcul de distance

Si x, y et z sont trois objets et si x et y sont regroupés en un seul élément noté h, examinons la façon dont on peut définir la distance de ce groupement à z

Distance du *saut minimal* (*single linkage*)

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

Calcul de distance

Si x, y et z sont trois objets et si x et y sont regroupés en un seul élément noté h, examinons la façon dont on peut définir la distance de ce groupement à z

Distance du *saut minimal* (*single linkage*)

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

Distance du *saut maximal* (*complete linkage*)

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

Calcul de distance

Si x, y et z sont trois objets et si x et y sont regroupés en un seul élément noté h, examinons la façon dont on peut définir la distance de ce groupement à z

Distance du *saut minimal* (*single linkage*)

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

Distance du *saut maximal* (*complete linkage*)

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

Distance moyenne

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

Distance moyenne généralisée

Si 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

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

Algorithme

On désignera par élément soit les objets à classer eux-mêmes, soit les regroupements d'objets générés par l'algorithme.

Algorithme

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) ;

Algorithme

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 aggrège en un nouvel élément ;

Algorithme

On désignera par élément soit les objets à classer eux-mêmes, soit les regroupements d'objets générés par l'algorithme.

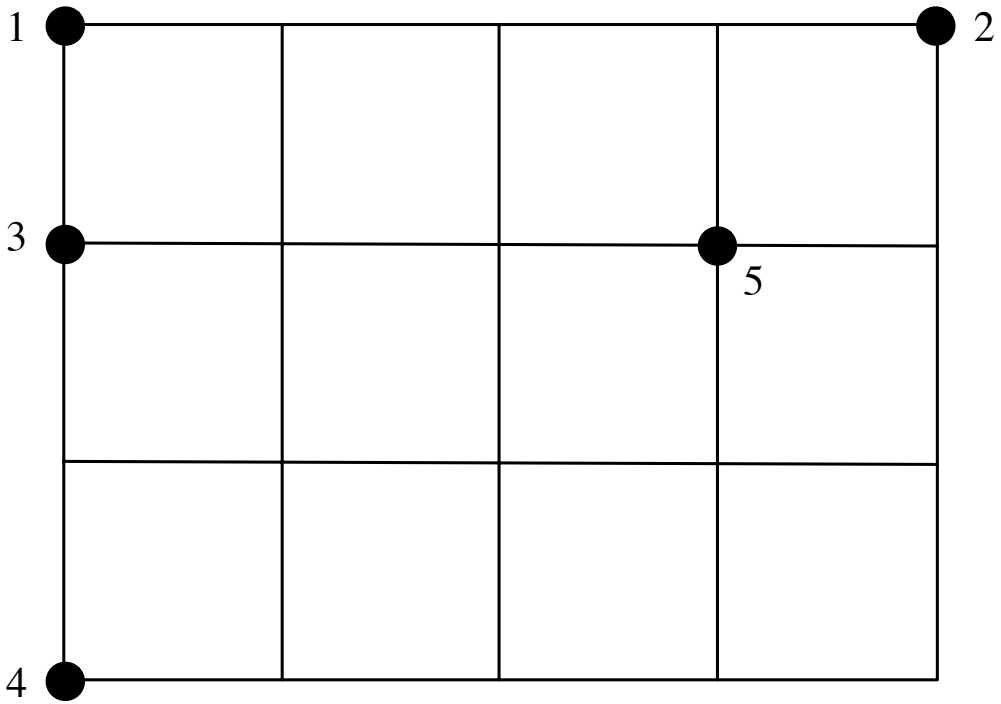
1. à l'étape 1, il y a n éléments à classer (qui sont les n objets) ;
2. on cherche les deux éléments les plus proches que l'on aggrège en un nouvel élément ;
3. on calcule les distances entre le nouvel élément et les éléments restants. On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec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

Algorithme

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 aggrège 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.

Classification ascendante hiérarchique



	(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

Distance = carré de la distance euclidienne

Exemple

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

a)

Exemple

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) ;

	(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)

Exemple

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) ;

	(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)

Exemple

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)

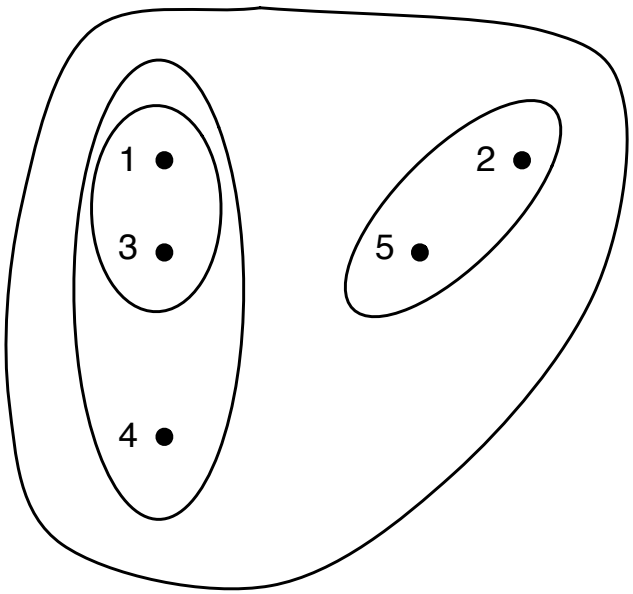
	(4)	(6)	(7)
(4)	0	4	13
(6)	4	0	9
(7)	13	9	0

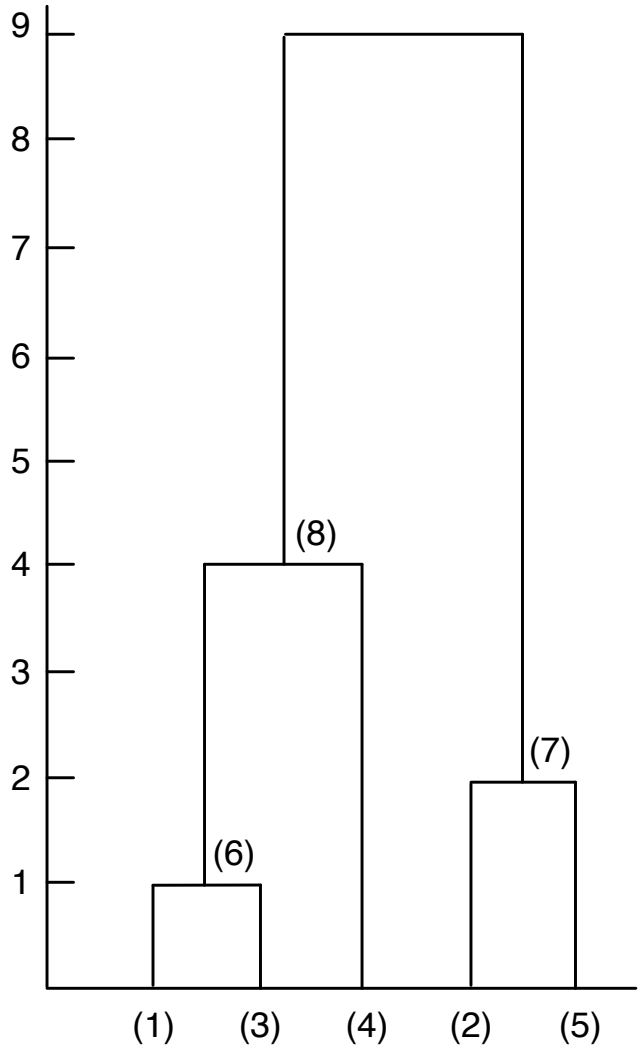
b)

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

c)

Regroupements obtenus – hiérarchie indicée





Hiérarchie indicée : 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 $v(h) < v(h')$.

Distance

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)

Distance

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)

Notion d'ultramétrique

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

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

Classification ascendante hiérarchique

Distance

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)

Notion d'ultramétrique

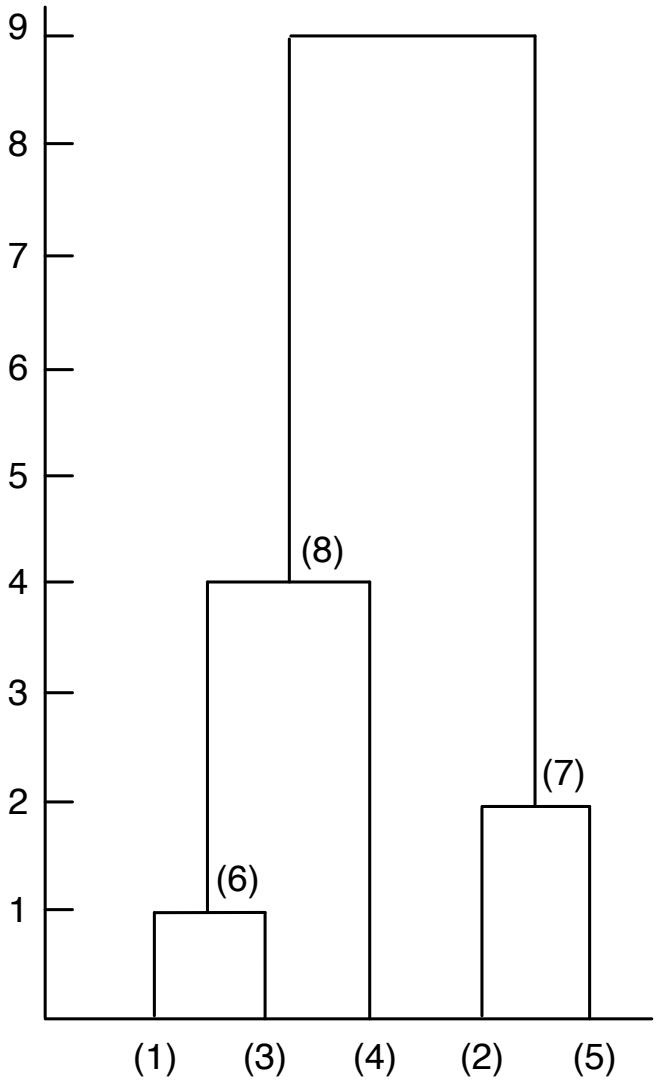
Cette distance sera dite ultramétrique si elle vérifie la condition suivante, plus forte que l'inégalité triangulaire :

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

Equivalence

Il est équivalent de munir l'ensemble E d'une ultramétrique ou de définir une hiérarchie indicée de parties de cet ensemble.

Classification ascendante hiérarchique



	(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

Hiérarchie indicée $\rightarrow$ distance

Montrons que

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

- Deux parties de la hiérarchie H sont soit disjointes, soit liées par une relation d'inclusion.

Hiérarchie indicée $\rightarrow$ distance

Montrons que

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

- 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 $d(x, z)$).

Hiérarchie indicée $\rightarrow$ distance

Montrons que

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

- 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. $h(y, z) \subset h(x, z)$).

Hiérarchie indicée $\rightarrow$ distance

Montrons que

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

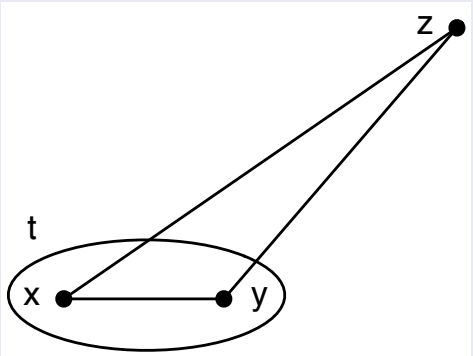
- Deux parties de la hiérarchie H sont soit disjointes, soit liées par une relation d'inclusion.
- Appelons $h(x,z)$ la plus petite partie de H contenant x et z (dont l'indice est $d(x,z)$).
- Puisque $h(x,z)$ et $h(y,z)$ ne sont pas disjointes, on a, par exemple, $h(x,z) \subset h(y,z)$ (resp. $h(y,z)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

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

ce qui établit l'inégalité.

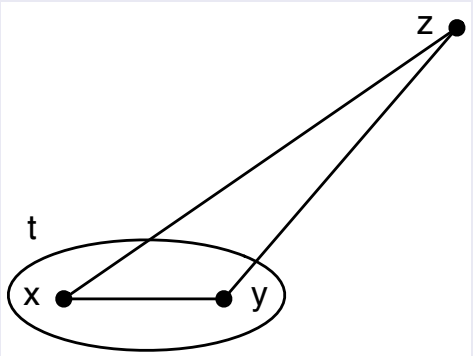
Distance → hiérarchie indicée

Si x et y sont agrégés en t, il faut en principe calculer les distances au nouvel élément t



Distance → hiérarchie indicée

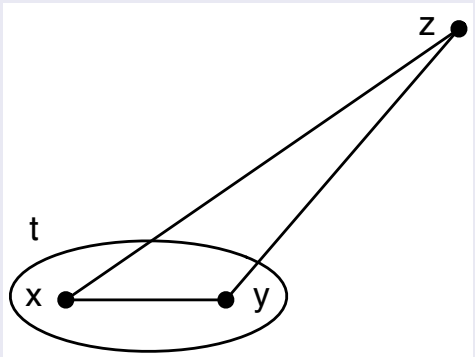
Si x et y sont agrégés en t, il faut en principe calculer les distances au nouvel élément t



- On a obligatoirement $d(z,x) \geq d(x,y)$ et $d(z,y) \geq d(x,y)$ sinon (z,x) ou (z,y) auraient été agrégés à la place de (x,y).

Distance → hiérarchie indicée

Si x et y sont agrégés en t, il faut en principe calculer les distances au nouvel élément t

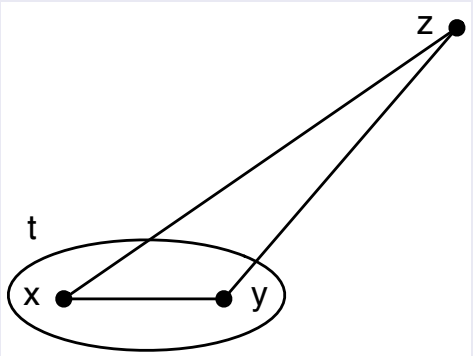


- On a obligatoirement $d(z,x) \geq d(x,y)$ et $d(z,y) \geq d(x,y)$ sinon (z,x) ou (z,y) auraient été agrégés à la place de (x,y).
- Pour une ultramétrique, cela implique que $d(z,x) = d(z,y)$, ce que l'on peut exprimer de façon imagée en disant que pour une ultramétrique, tous les triangles sont isocèles, avec le plus petit côté pour base.

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

Distance → hiérarchie indicée

Si x et y sont agrégés en t, il faut en principe calculer les distances au nouvel élément t



- On a obligatoirement $d(z,x) \geq d(x,y)$ et $d(z,y) \geq d(x,y)$ sinon (z,x) ou (z,y) auraient été agrégés à la place de (x,y).
- Pour une ultramétrique, cela 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.

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

- De la même façon :

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

Agrégation autour de centres mobiles

Position du problème

- soit à partitionner un ensemble I de n individu caractérisés par p paramètres ou variables

Agrégation autour de centres mobiles

Position du problème

- soit à partitionner un ensemble I de n individu caractérisés par p paramètres ou variables
- l'espace $\mathbb{R}^p$ supportant les n points-individus est muni d'une distance appropriée notée d (fréquemment la distance euclidienne)

Agrégation autour de centres mobiles

Position du problème

- soit à partitionner un ensemble I de n individu caractérisés par p paramètres ou variables
- 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.

Agrégation autour de centres mobiles

Position du problème

- soit à partitionner un ensemble I de n individu caractérisés par p paramètres ou variables
- 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.

Algorithme

Agrégation autour de centres mobiles

Position du problème

- soit à partitionner un ensemble I de n individu caractérisés par p paramètres ou variables
- 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.

Algorithme

- Etape 0 : $k = 0$; déterminer q centres provisoires de classes.
  Centres $\{C_1^0, C_2^0, \ldots, C_q^0\} \rightarrow$ partition P^0 de I en q classes $\{I_1^0, I_2^0, \ldots, I_q^0\}$.
  L'individu $i \in I_k^0$ 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).

Agrégation autour de centres mobiles

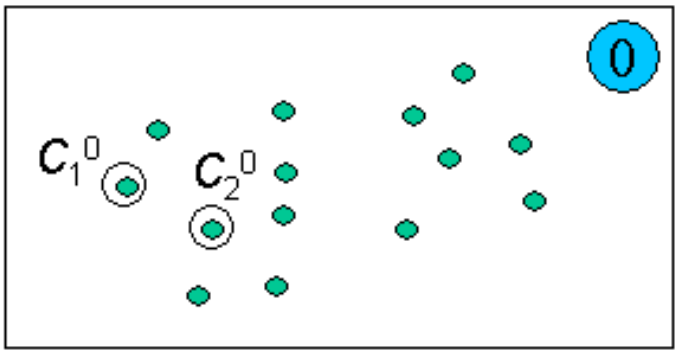
Position du problème

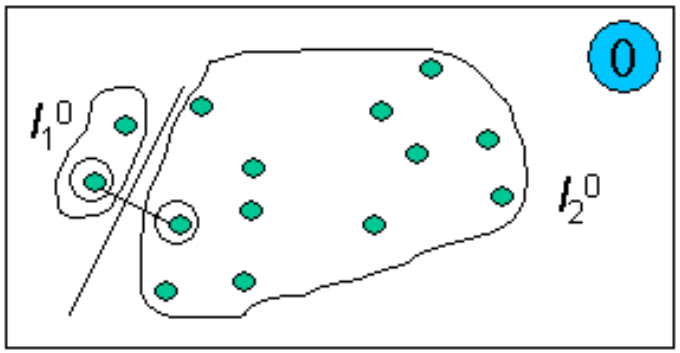
- soit à partitionner un ensemble I de n individu caractérisés par p paramètres ou variables
- l'espace $\mathbb{R}^p$ supportant les n points-individus est muni d'une distance appropriée notée d (fréquemment la distance euclidienne)
- on désire constituer au maximum q classes.

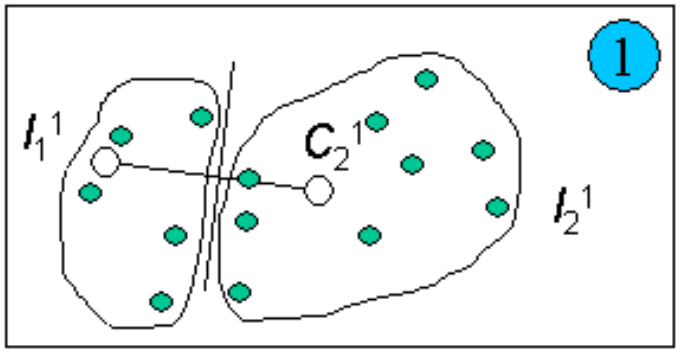
Algorithme

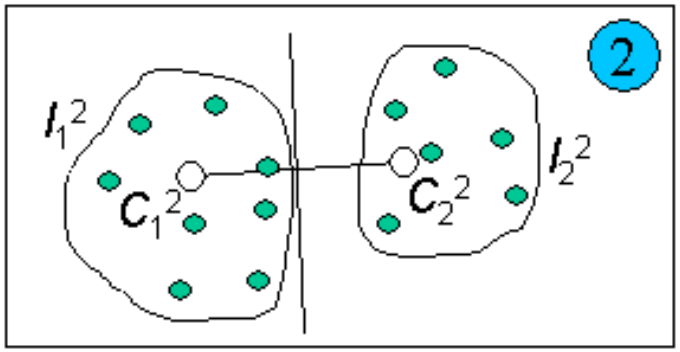
- Etape 0 : $k = 0$; déterminer q centres provisoires de classes.
  Centres $\{C_1^0, C_2^0, \ldots, C_q^0\} \rightarrow$ partition P^0 de I en q classes $\{I_1^0, I_2^0, \ldots, I_q^0\}$.
  L'individu $i \in I_k^0$ si le point i est plus proche de C_k^0 que de tous les autres centres.
  Les classes 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 $\{C_k^1, C_2^k, \ldots, C_q^k\}$ 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.

Agrégation autour de centres mobiles









Algorithmes similaires

Algorithmes similaires

- Nuées dynamiques
  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.

Algorithmes similaires

- Nuées dynamiques
  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.

- *k-means* ou k-moyennes
  Après avoir choisi une première fois les centres mobiles, on 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.

Plan

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 $(-1, +1)$.

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.

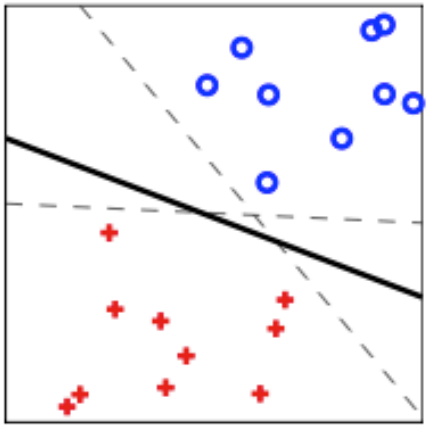
Introduction

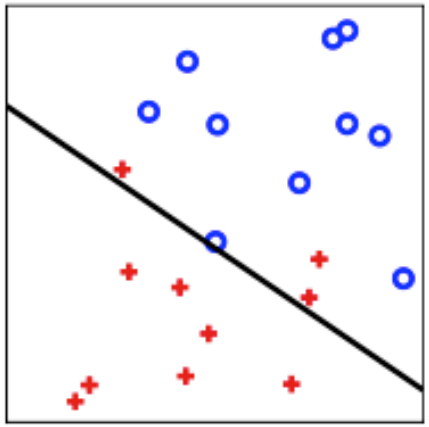
Les *Support Vector Machines* ou, en français, Séparateurs à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire $(-1, +1)$.

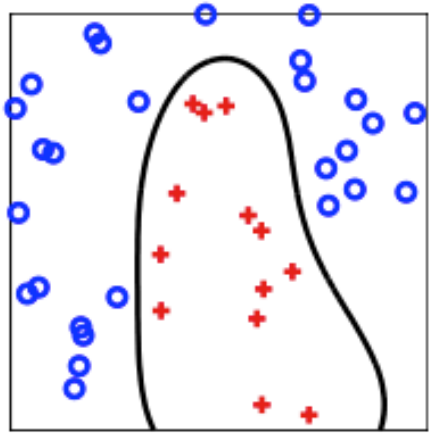
Ils sont basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité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

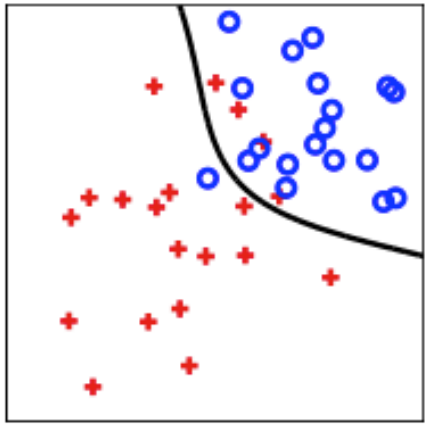
Cette approche découle directement des travaux de Vapnik en théorie de l'apprentissage à partir de 1995. Elle s'est focalisée sur les propriétés de généralisation (ou prévision) d'un modèle en contrôlant sa complexité. L'autre idée directrice de Vapnik est d'éviter de substituer à l'objectif initial : 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.

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









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

Le problème de discrimination linéaire

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.

Le problème de discrimination linéaire

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.

Le problème de discrimination linéaire

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.

Le problème de discrimination linéaire

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

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

Le problème de discrimination linéaire

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

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

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

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

Le problème de discrimination linéaire

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

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

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 $w^T x + b = 1$).

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

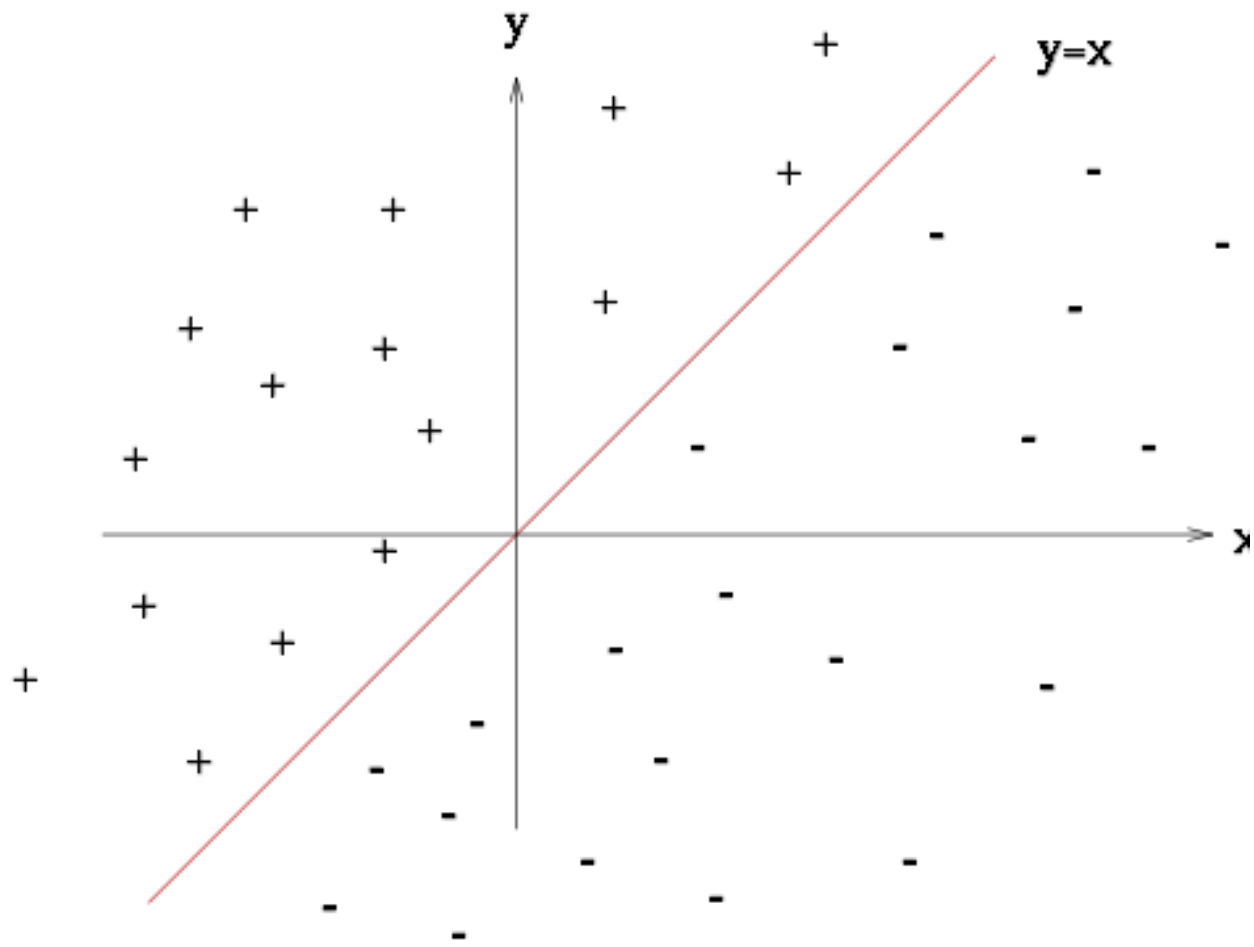


Figure – Exemple d'un problème de discrimination à deux classes, avec une séparatrice linéaire : la droite d'équation $y = x$. Le problème est linéairement séparable.

Marge maximale

On se place désormais dans le cas où le problème est linéairement séparable. 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.

Marge maximale

On se place désormais dans le cas où le problème est linéairement séparable. 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.

Marge maximale

On se place désormais dans le cas où le problème est linéairement séparable. 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 :

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

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

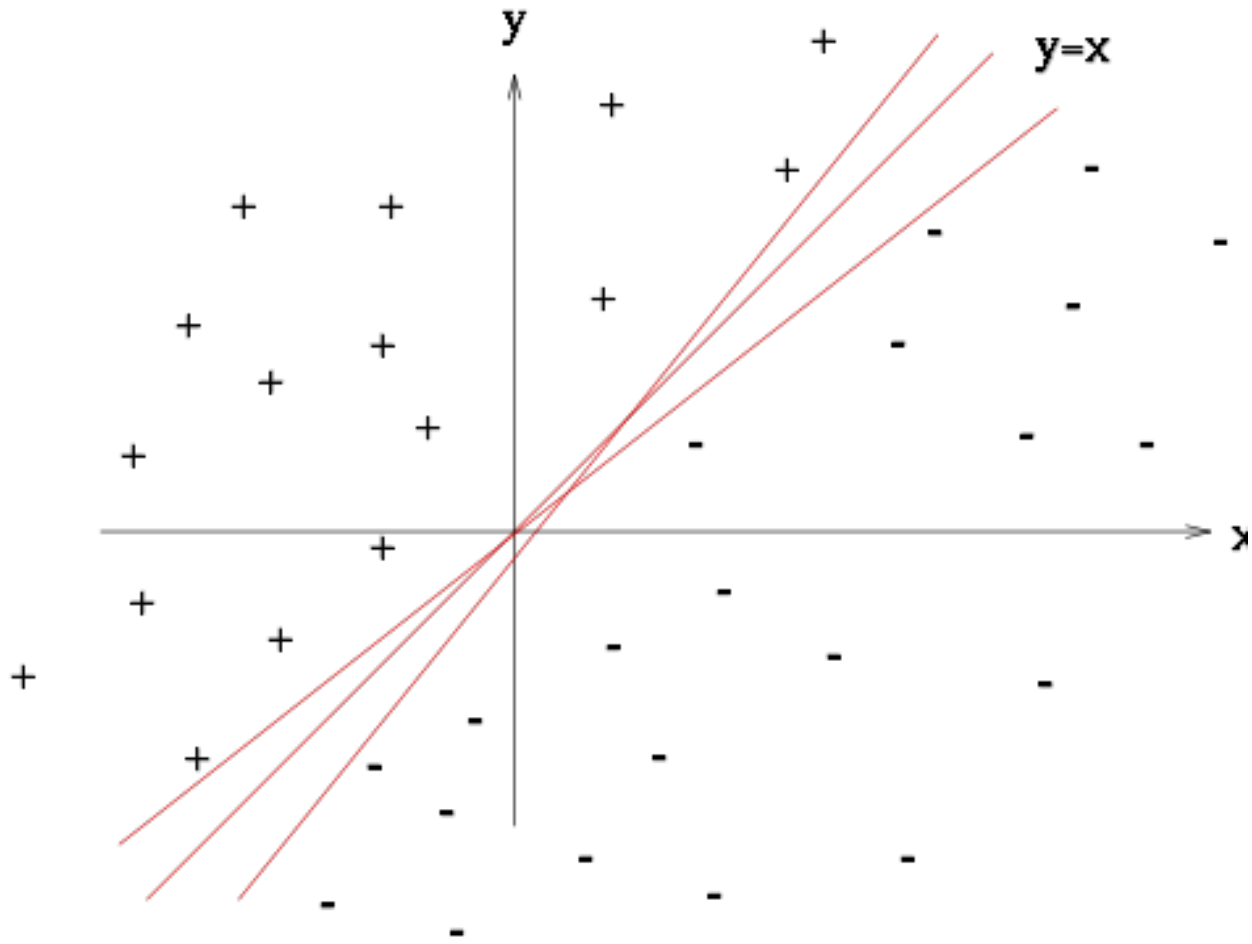


Figure – Pour un ensemble de points linéairement séparables, il existe une infinité d'hyperplans séparateurs.

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

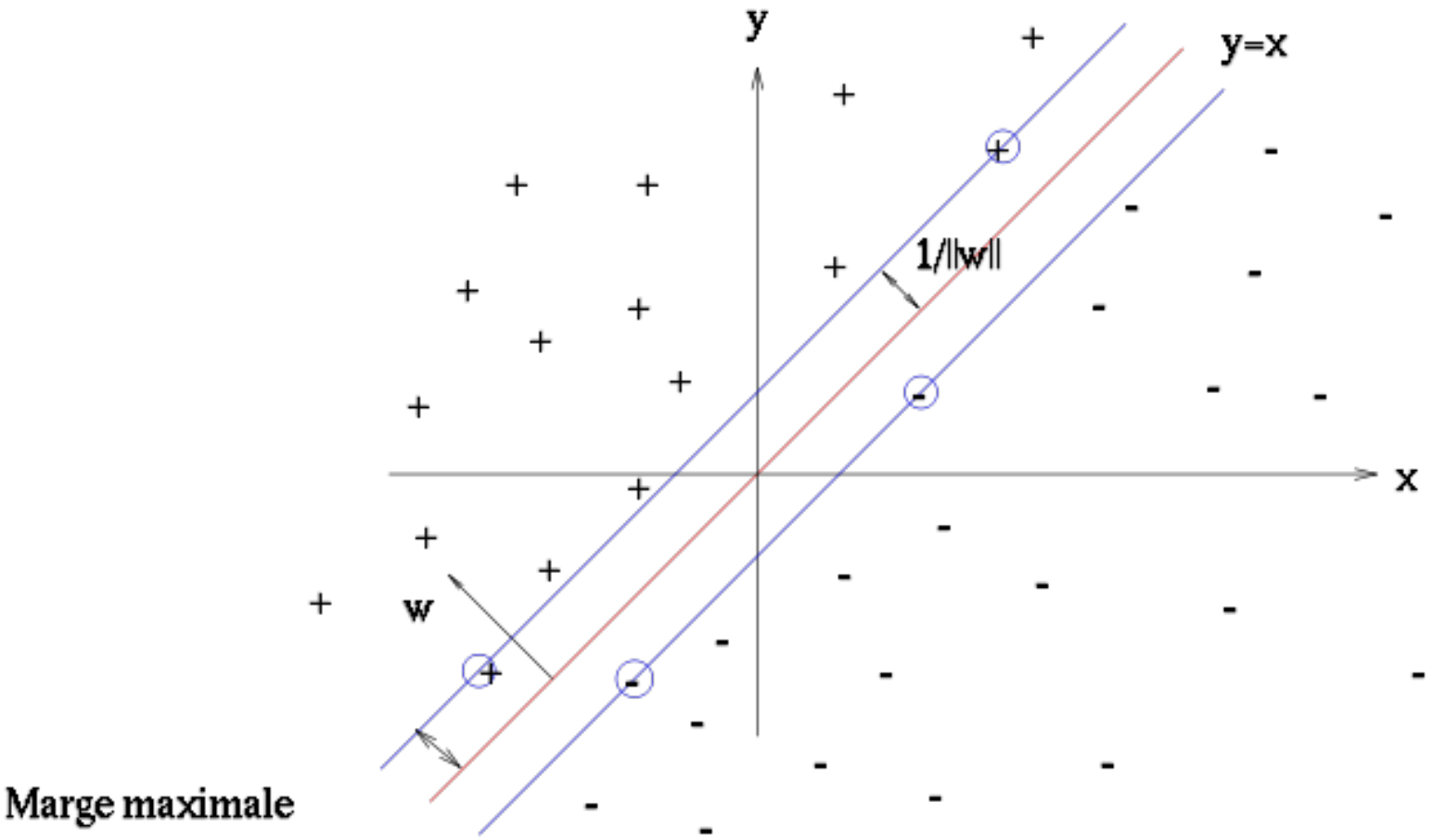


Figure – L'hyperplan optimal (en rouge) avec la marge maximale. Les échantillons entourés sont des vecteurs supports.

Hyperplan séparateur optimal

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

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

Hyperplan séparateur optimal

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.

Hyperplan séparateur optimal

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 $\frac{1}{||w||}$ et il s'agit donc de maximiser $||w||^{-1}$.

Hyperplan séparateur optimal

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

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

Hyperplan séparateur optimal

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 $]0, +\infty[$).

Hyperplan séparateur optimal

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 à $\alpha = (\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_N)$ (en effet, ici, l'optimum est un point selle).

SVM sur des données linéairement séparables

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 :

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

Résolution

Le problème d'optimisation sous contraintes précédent 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

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

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

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

Résolution

Le problème d'optimisation sous contraintes précédent 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

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

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

Conditions d'optimalité et vecteurs supports

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.

Conditions d'optimalité et vecteurs supports

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 :

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

Conditions d'optimalité et vecteurs supports

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 :

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

Conditions d'optimalité et vecteurs supports

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 : $\ell_j({w^*}^T x_j + b^*) > 1, \forall j \notin \mathcal{A}$.

Conditions d'optimalité et vecteurs supports

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 : $\ell_j({w^*}^T x_j + b^*) > 1, \forall j \notin \mathcal{A}$.

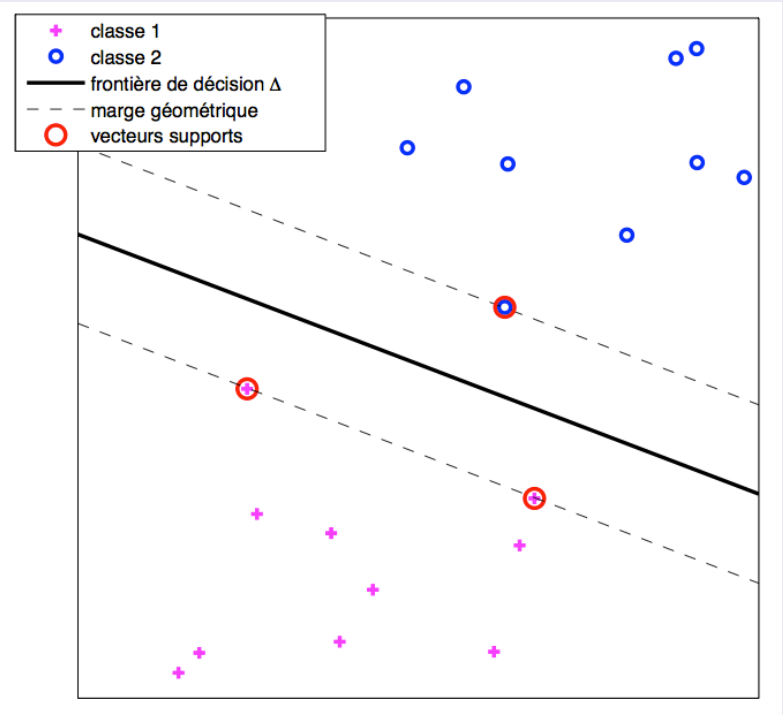
Si l'on note $\bullet_{\mathcal{A}}$ le vecteur constitué des seules composantes de $\bullet$ indexées par $\mathcal{A}$, la solution optimale $(w^*, b^*, \alpha^*_{\mathcal{A}})$ vérifie le système d'équations linéaires suivant :

$$\begin{cases} w^* & -X_{\mathcal{A}}\text{diag}(\ell_{\mathcal{A}})\alpha^*_{\mathcal{A}} & & = 0 \\ \text{diag}(\ell_{\mathcal{A}})X_{\mathcal{A}}w^* & & b^*\ell_{\mathcal{A}} & = e_{\mathcal{A}} \\ & -\ell^T_{\mathcal{A}}\alpha^*_{\mathcal{A}} & & = 0 \end{cases}$$

De la première égalité, on obtient donc : $w^* = \sum_{i \in \mathcal{A}} \alpha^*_i \ell^*_i x_i$

Conditions d'optimalité et vecteurs supports

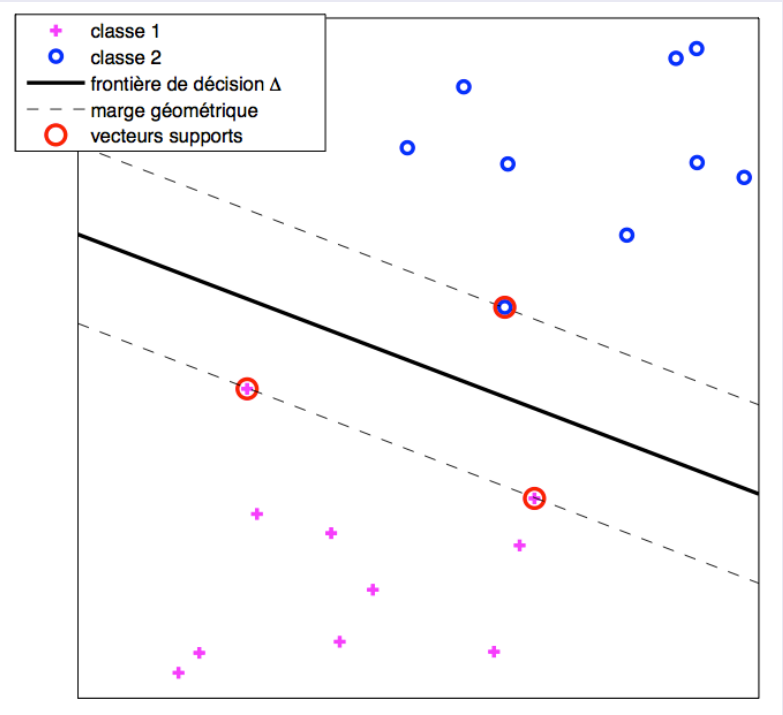
Le vecteur w est donc une combinaison linéaire des observations x_i liées aux contraintes actives $i \in \mathcal{A}$ et pour lesquelles on a $|w^{*T} x_i + b^*| = 1$ (observations sur les frontières). Ces observations sont appelées vecteurs supports ; leur marge est égale à 1).



Conditions d'optimalité et vecteurs supports

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).



Formulation duale

Au problème d'optimisation initial,

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

Formulation duale

Au problème d'optimisation initial,

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

on peut associer la formulation duale suivante :

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

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

Formulation duale

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

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

Formulation duale

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

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

Remarques (avantages – inconvénients)

Les deux formulations primales et duales 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.

Remarques (avantages – inconvénients)

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.

Fonction de décision

Rappelons 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 :

$$f(x) = \sum_{k \in \mathcal{A}} \alpha_k^* \ell_k x_k^T x + b^*, \qquad \text{où } \mathcal{A} \text{ est l'ensemble des indices des vecteurs supports.}$$

Cas de données non séparables linéairement

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.

Cas de données non séparables linéairement

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 :

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

Cas de données non séparables linéairement

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 :

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

Séparateur à marge poreuse

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.

Références

Ces planches sont très largement inspirées des références suivantes :

- 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
- 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. http://www.sites.univ-rennes2.fr/laboratoire-statistique/PAC/doc/score.pdf
- Lebart L., Morineau A., Fénelon J.P. *Traitement des données statistiques - méthodes et programmes*. Bordas, Paris, 1979.
- Martin A. *L'analyse de données*. Polycopié d'analyse données, ENSIETA, septembre 2004. http://www.arnaud.martin.free.fr/Doc/polyAD.pdf
- WikiStat. *Machines à vecteurs supports*. Document de cours. http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-svm.pdf

Qui suis-je ?

Didier Maquin

Professeur d'Automatique
Université de Lorraine

Ecole Nationale Supérieure d'Electricité et de Mécanique

Ecole Nationale Supérieure des Mines de Nancy

Centre de Recherche en Automatique de Nancy

Contact : didier.maquin@univ-lorraine.fr
Localisation : bureau 116 jaune

Plus de détails ?

Site personnel : http://www.cran.univ-lorraine.fr/didier.maquin

Site du laboratoire : http://www.cran.univ-lorraine.fr