Réduction de la dimensionnalité de données

Principe de la réduction de dimensions

- Dans les applications de régression ou de classification, on utilise des données d'entrée pour l'apprentissage.

- Cependant, il peut exister plusieurs raisons pour réduire le nombre de dimensions (de caractéristiques) des données d'entrée dans une phase de prétraitement.

  ☞ La complexité des algorithmes d'apprentissage dépend du nombre d'attributs D et le nombre de données N.

  ☞ Lorsqu'on détermine qu'une caractéristique (attribut) n'est pas utile pour notre apprentissage, on s'épargne le temps de le calculer ou de le mesurer sur un objet.

Principe de la réduction de dimensions

- Lorsque les données peuvent être expliquées (en terme de classification ou de régression) avec peu d'attributs (observés ou cachées) .

  $\Rightarrow$ on permet une meilleure généralisation du processus d'apprentissage.

- Permet de décrire le processus de génération de données et d'extraction de connaissances de manière plus simple.

- Lorsque des données peuvent être représentées par quelques dimensions sans perte d'information, on peut les visualiser et les analyser de manière plus facile.

Principe de la réduction de dimensions

- Formellement, le problème est d'apprendre un fonction f(x) telle que

$$f: \mathbb{R}^D \rightarrow \mathbb{R}^M$$

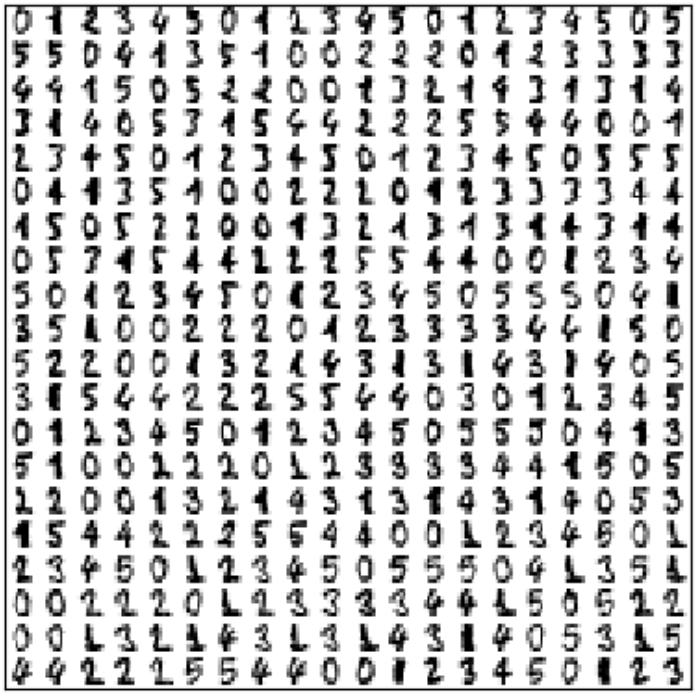
où la dimensionnalité $M < D$ (c'est un hyper-paramètre)

- Applications:
  - Visualisation des données
  - Limiter le sur-apprentissage

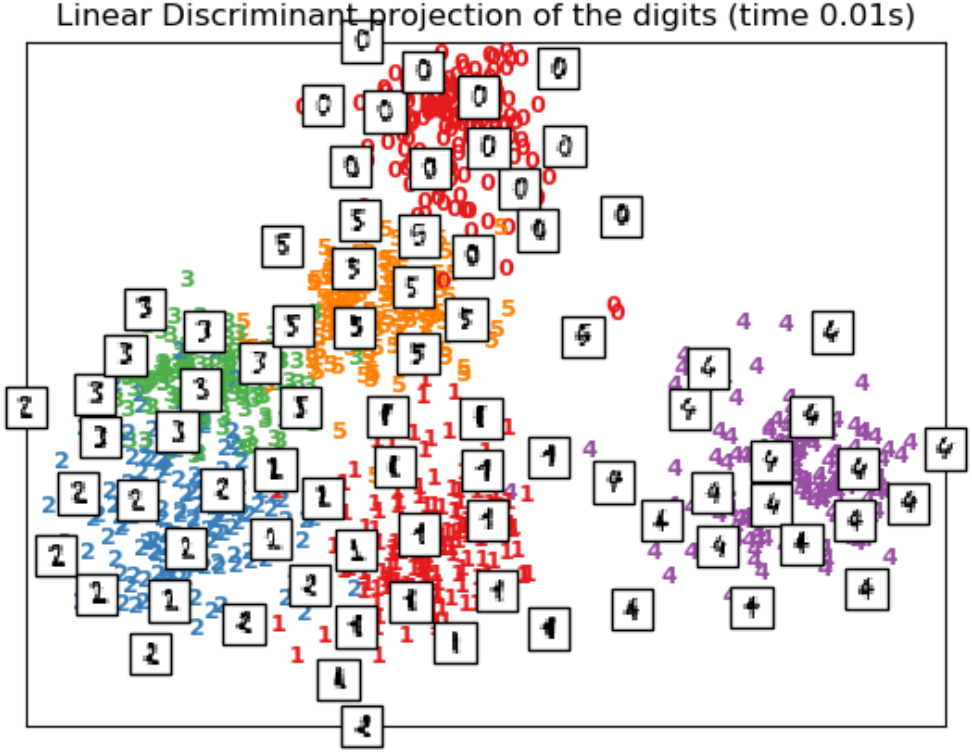
Visualisation des données

- Visualisation des données

A selection from the 64-dimensional digits dataset



Linear Discriminant projection of the digits (time 0.01s)



https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html

Méthodes de réduction de dimensions

- Il existe deux approches principales pour réduire la dimension des données.

☞ **Sélection d'attributs:**

On s'intéresse à sélectionner M attributs parmi D qui donneront l'information requise pour l'apprentissage ($M < D$). On doit donc rejeter $(D - M)$ attributs.

☞ **Extraction d'attributs:**

On s'intéresse à trouver un nouveau ensemble de M attributs qui donneront l'information requise pour l'apprentissage. Ces méthodes peuvent être supervisées ou non supervisées.

Sélection d’attributs

Principe de sélection d'attributs

- On s'intéresse à trouver le meilleur sous-ensemble (minimal) d'attributs pour construire une régression/classification précise.

- Une fois un sous-ensemble d'attributs est sectionné, on rejette les autres attributs non sélectionnés.

- En ayant D attributs au départ, Il existera $(2^D - 1)$ sous-ensembles possibles d'attributs qu'on peut former (**ex.** avec $D = 10,$ on aura 2^{10} sous-ensembles).

- En ayant une fonction d'erreur, on ne peut tester rapidement toutes les combinaisons que si D est relativement petit.

Méthodes de sélection d'attributs

- Il existe deux méthodes pour la sélection d'attributs:

☞ **Sélection en avant d'attributs (i.e. par ajout).**

On commence par un ensemble vide d'attributs et on ajoute à chaque étape l'attribut qui décroit le plus l'erreur de validation. On arrête quand l'erreur devient stable ou croissante.

☞ **Sélection en arrière d'attributs (i.e. par élimination).**

On commence par un ensemble de tous les attributs et on élimine à chaque étape l'attribut qui décroit le plus l'erreur de validation. On arrête quand l'erreur devient stable ou croissante.

Sélection en avant d'attributs

- Soit $\mathcal{A}= \{x_1, x_2, \cdots, x_D\}$ l'ensemble des D attributs dont nous disposons au départ pour les données, et soit S les attributs sélectionnés.

- Soit $E(S)$ l'erreur de validation produite en utilisant S.

- La sélection en avant d'attributs commence par $S = \emptyset$. À chaque étape, pour chaque attribut x_d, on fait l'apprentissage et le test de validation pour obtenir: $E(S \cup \mathrm{x}_d)$.

- On choisit alors le meilleur attribut à ajouter à S, tel que:

$$\mathrm{x}_j = argmin_d(E(S \cup \mathrm{x}_d))$$

Sélection en arrière d'attributs

- On arrête le processus de sélection lorsque l'erreur E ne décroit plus ou bien ne décroit pas significativement.

- Cet algorithme de sélection d'attributs s'appelle: enroulement (wrapper), car le modèle de classification ou de régression est utilisé comme une routine pour la validation.

Remarque:

- La sélection en arrière d'attributs se fait par le même principe que la sélection en avant, en mettant l'ensemble de départ $S=\mathcal{A}$ et en remplaçant $S \cup \mathrm{x}_d$ par $S-\mathrm{x}_d$.

Etude d'un exemple

Exemple:

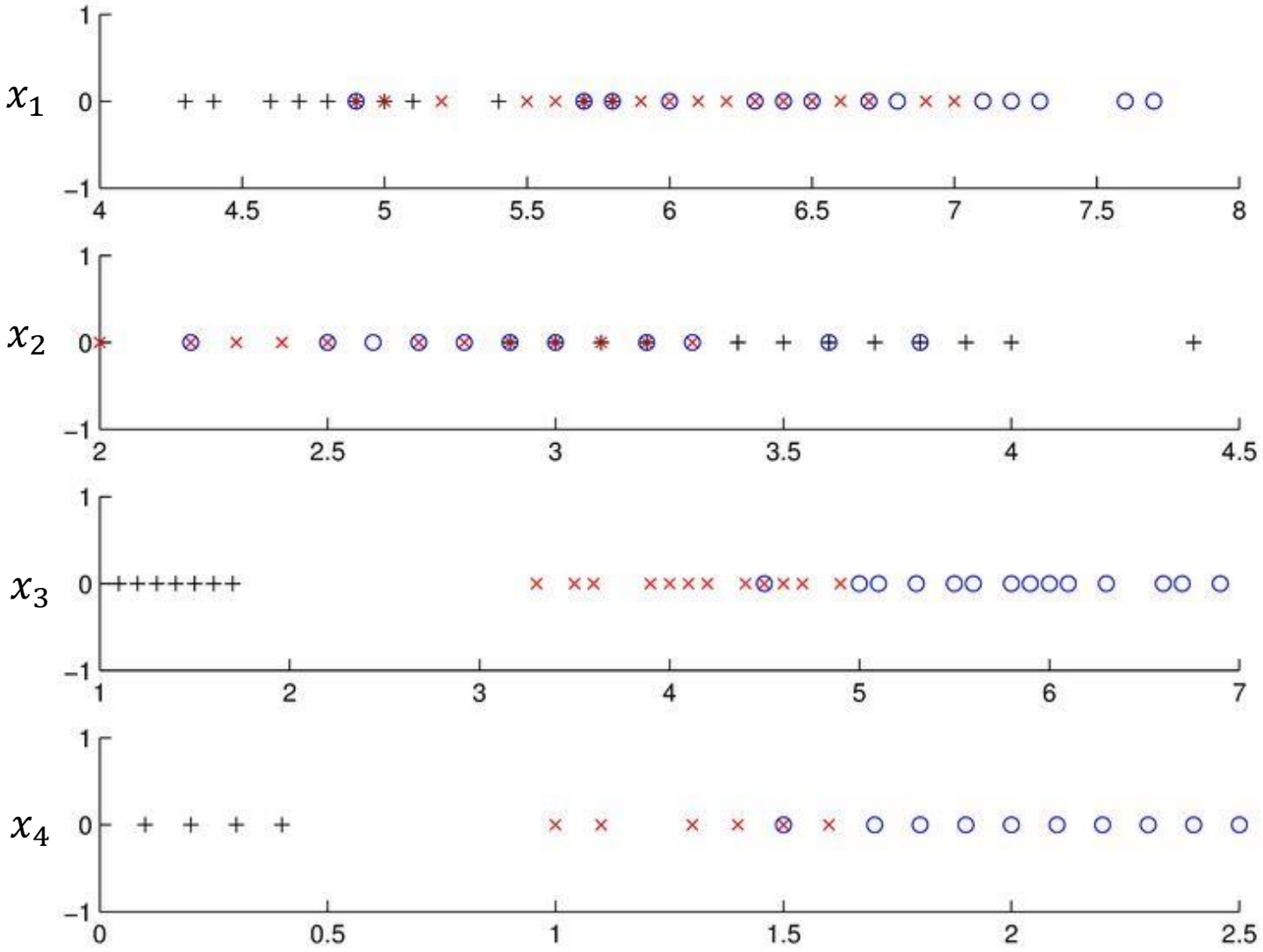
- Pour les données IRIS, nous avons $N = 150$ données $D = 4$ attributs $\{x_1, x_2, x_3$ et $x_4\}$.

- On utilise la méthode K-moyenne pour la classification.

- En utilisant un attribut à la fois, les précisions obtenues pour les attributs x_1, x_2, x_3 et x_4 sont, respectivement, 0.76, 0.57, 0.92 et 0.94.

☞ Nous sélectionnons alors x_4 comme 1er attribut.

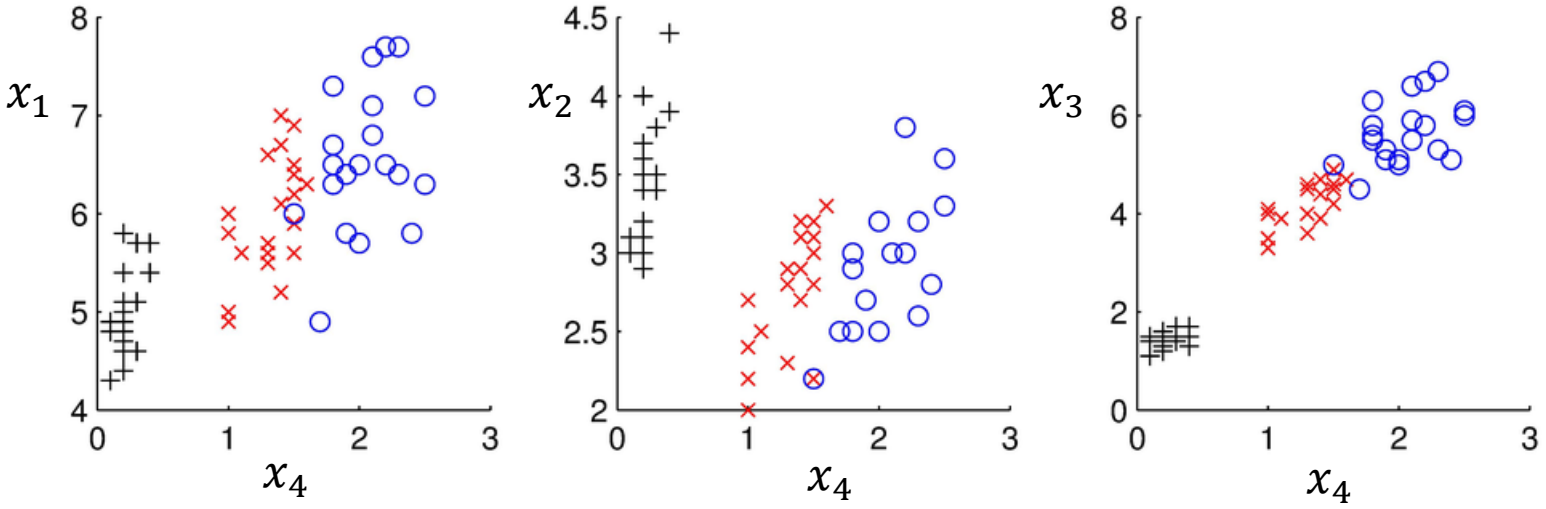
Etude d'un exemple



Etude d'un exemple

- En ajoutant un 2e attribut, les précisions obtenues avec x_1, x_2, x_3 sont, respectivement, 0.87, 0.92 et 0,92.

☞ Nous sélectionnons alors x_3 comme 2e attribut.



Conclusions sur la sélection d'attributs

- En ajoutant un 3e attribut, les précisions obtenues avec les x_1, x_2 sont égales à 0.94.

☞ On ne rajoute pas de 3e attribut.

Remarques

- La sélection d'attributs est supervisée car les sorties de la régression/classification sont utilisées pour calculer l'erreur de validation.

- Pour certaines applications (ex. reconnaissance faciale), la sélection individuelle des attributs (ex. pixels) n'est pas très discriminative pour la classification.

Extraction d’attributs:
Analyse à composantes principales (ACP)

Principe d'extraction d'attributs

- Plusieurs techniques existent pour l'extraction de nouveaux attributs dans un ensemble de données.
- Les méthodes de projection sont parmi les plus importantes pour l'extraction d'attributs.
- Ces méthodes essayent d'extraire M attributs $(M < D)$ de l'espace original à D dimensions, de sorte qu'il y a une perte minimale d'information sur la structure des données.
- Les méthodes les plus populaires sont l'analyse à composantes principales (ACP), l'analyse discriminante (AD), l'analyse factorielle (AF), analyse de corrélation (AC), etc.

Analyse à composantes principales (ACP)

- Est une méthode non supervisée (c.-à-d. n'utilise aucune information de sortie).
- L'information importante des données à garder est leur variance, de sorte qu'une fois projetées, les données auront le maximum de variance gardée.
- Soit x un vecteur aléatoire dans mon espace de départ $\boldsymbol{\mathcal{X}}$.
- Soit $\boldsymbol{w}_{(1)}$ une direction de projection telles que $\|\boldsymbol{w}_{(1)}\| = 1$.
- Soit $z_1 = \boldsymbol{w}_{(1)}{}^T x$, une variable produite par la projection de x sur la ligne de direction $\boldsymbol{w}_{(1)}$.

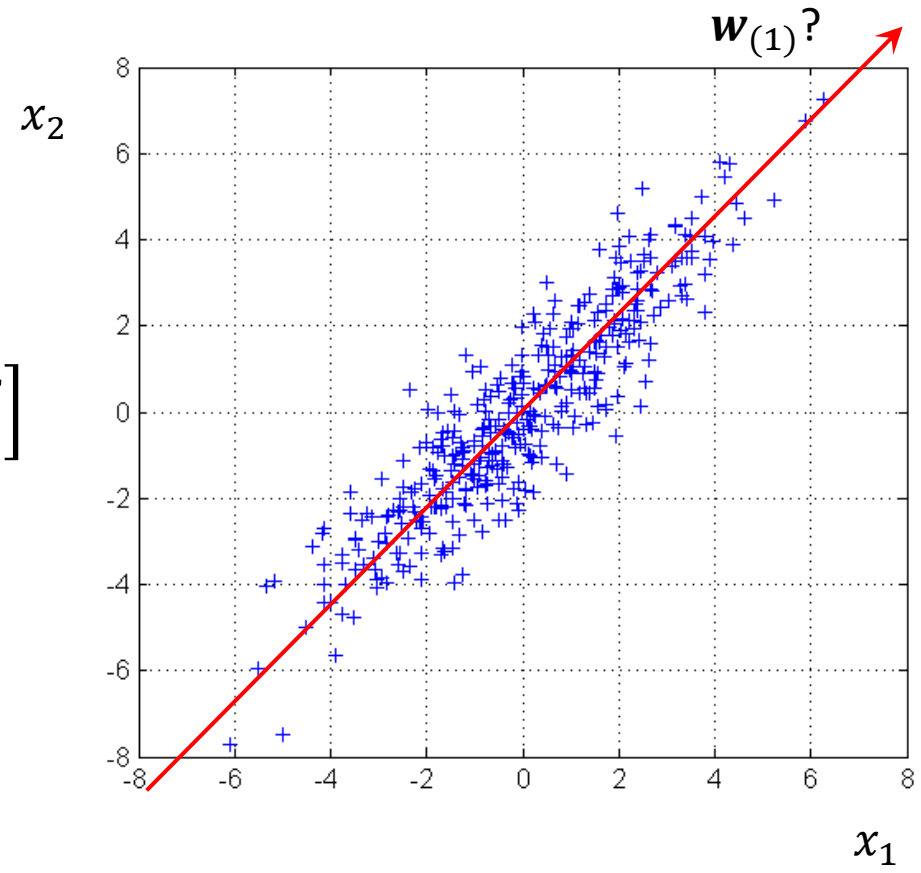
Analyse à composantes principales (ACP)

Exemple:

Quelle est la direction de projection qui maximise la variance?

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 4 \\ 4 & 5 \end{bmatrix}$$



Analyse à composantes principales (ACP)

- Soit $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ la moyenne et la matrice de covariance de x.
- On peut démontrer que $var(z_1) = \boldsymbol{w}_{(1)}{}^T \boldsymbol{\Sigma}\, \boldsymbol{w}_{(1)}$.
- On cherche alors $\boldsymbol{w}_{(1)}$ tel que $var(z_1)$ soit maximisée et que $\|\boldsymbol{w}_{(1)}\| = 1$. La formulation de cette recherche est faite par:

$$max_{\boldsymbol{w}_{(1)}} \left\{ \boldsymbol{w}_{(1)}{}^T \boldsymbol{\Sigma}\, \boldsymbol{w}_{(1)} - \lambda\left(\boldsymbol{w}_{(1)}{}^T \boldsymbol{w}_{(1)} - 1\right)\right\}$$

où λ est le multiplicateur de Lagrange pour garder la contrainte $\|\boldsymbol{w}_{(1)}\| = 1$ satisfaite pendant la maximisation.

Analyse à composantes principales (ACP)

- On prenant égale à 0 la dérivée de la fonction par rapport à $\boldsymbol{w}_{(1)}$, on obtient:

$$2\boldsymbol{\Sigma}\,\boldsymbol{w}_{(1)} - 2\lambda\boldsymbol{w}_{(1)} = 0 \qquad \Rightarrow \qquad \boldsymbol{\Sigma}\,\boldsymbol{w}_{(1)} = \lambda\boldsymbol{w}_{(1)}$$

- On remarque que $\boldsymbol{w}_{(1)}$ correspond au vecteur propre de la matrice $\boldsymbol{\Sigma}$ et λ est une valeur propre de la même matrice.

- Puisque $\|\boldsymbol{w}_{(1)}\| = 1,$ on remarque aussi que:

$$\begin{aligned}\boldsymbol{w}_{(1)}{}^{T}(\boldsymbol{\Sigma}\,\boldsymbol{w}_{(1)}) &= \boldsymbol{w}_{(1)}{}^{T}\,(\lambda\boldsymbol{w}_{(1)}).\\ &= \lambda(\boldsymbol{w}_{(1)}{}^{T}\,\boldsymbol{w}_{(1)})\\ &= \lambda\end{aligned}$$

Analyse à composantes principales (ACP)

- Donc, la 1ère direction de projection maximisant la variance est celle correspondant à la plus grande valeur propre de $\boldsymbol{\Sigma}$.

- Soit $\boldsymbol{w}_{(2)}$ la 2e direction de projection maximisant la variance $var(z_2) = \boldsymbol{w}_{(2)}{}^T \boldsymbol{\Sigma}\, \boldsymbol{w}_{(2)}$, telle que $\|\boldsymbol{w}_{(2)}\| = 1$ et $\boldsymbol{w}_{(1)}{}^T \boldsymbol{w}_{(2)} = 0$.

- On peut procéder par le même raisonnement pour trouver $\boldsymbol{w}_{(2)}$. On peut démontrer que:

  ☞ La 2e direction $\boldsymbol{w}_{(2)}$ correspond à la 2e plus grande valeur propre de $\boldsymbol{\Sigma}$.

- On peut suivre la même procédure pour extraire les M premières directions principales de projection maximisant la variance des données projetées.

Analyse à composantes principales (ACP)

- Soit $\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}, \ldots, \boldsymbol{w}_{(M)}$ les M premières directions de projection. Soit $\lambda_1, \lambda_2, \ldots, \lambda_M$ les valeurs propres correspondants à ces directions.

- On appelle ces directions les composantes principales (CP): $\boldsymbol{w}_{(1)}$ 1ère composante, $\boldsymbol{w}_{(2)}$ 2e composante, etc.

- En général, les premières composantes expliquent la plus grande variance des données.

- Par exemple, si on veut garder 90% de la variance des données, on peut garder les M CPs telles que:

Analyse à composantes principales (ACP)

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_M}{\sum_{d=1}^{D} \lambda_d} \approx 0.9$$

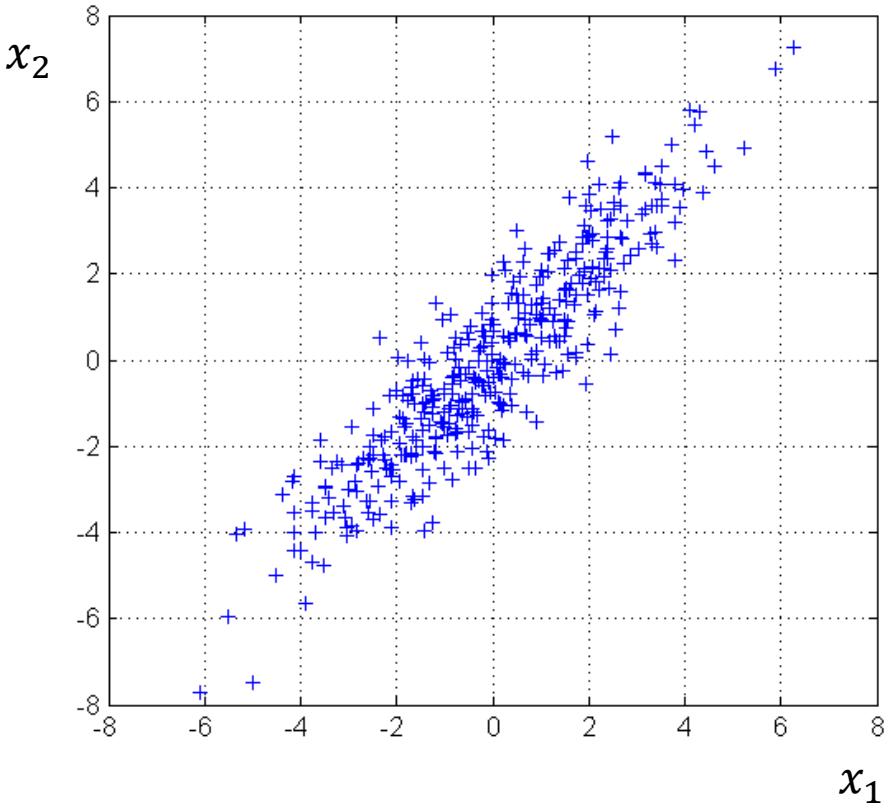
- Le graphe de Scree décrit le pourcentage de variance expliquée en fonction du nombre de CPs gardées.
- Soit $\boldsymbol{W}$ une matrice contenant les M premières composantes dans ces colonnes.
- On définit $\mathrm{z} = \boldsymbol{W}^{\boldsymbol{T}}(x - \boldsymbol{\mu})$ comme étant une transformation de l'espace de dimension D à un espace de dimension M qui préserve la variance des données.

Analyse à composantes principales (ACP)

Exemple: Soit un ensemble de 400 données dont la moyenne et la matrice de covariance sont données comme suit:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 4 \\ 4 & 5 \end{bmatrix}$$



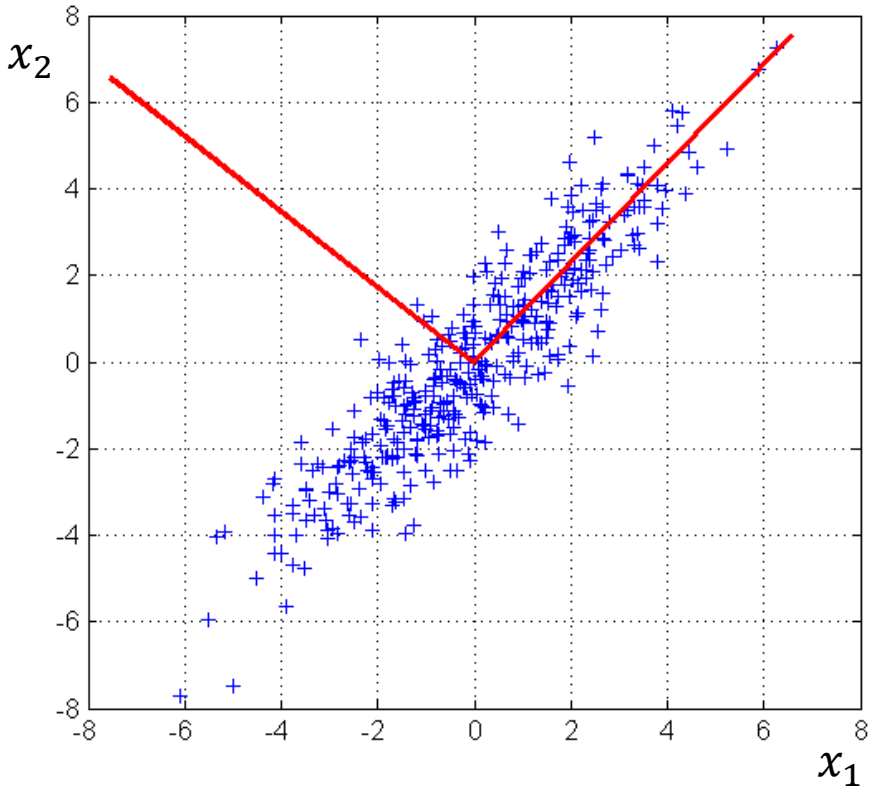
Analyse à composantes principales (ACP)

L'ACP donne les directions principales comme suit:

$$\mathbf{W} = \begin{bmatrix} \boldsymbol{w}_{(1)} & \boldsymbol{w}_{(2)} \\ \downarrow & \downarrow \\ 0.67 & -0.74 \\ 0.74 & 0.67 \end{bmatrix}$$

$\lambda_1 = 8.76.$

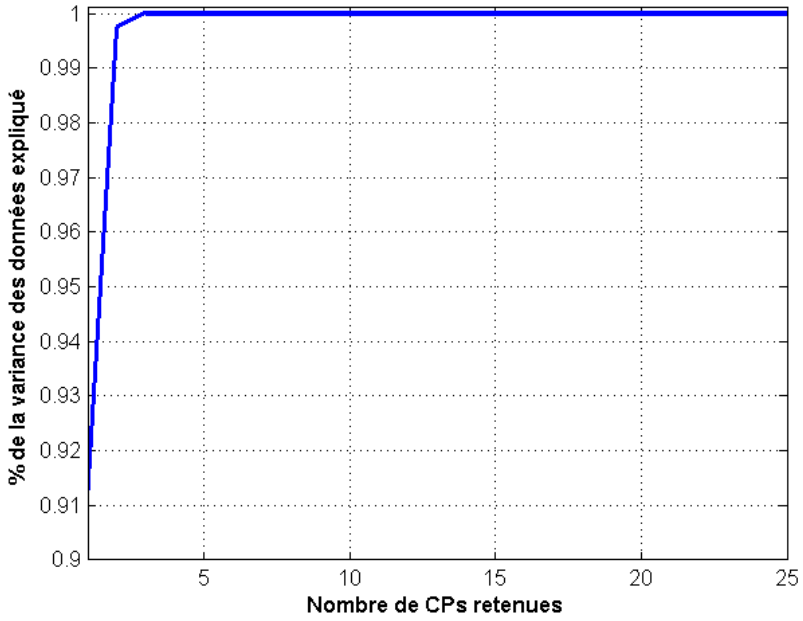
$\lambda_2 = 0.48$



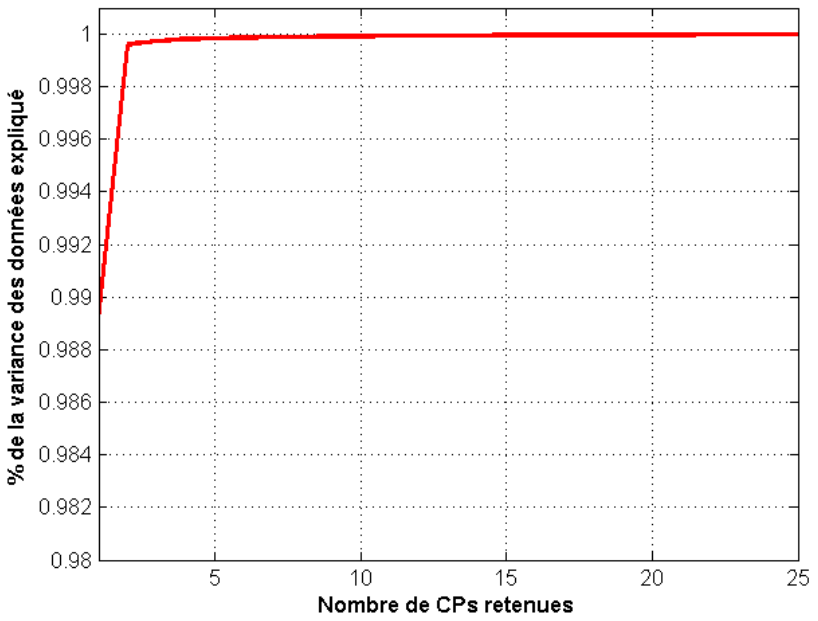
Analyse à composantes principales (ACP)

Avec les données de spams de HP Labs, on obtient les graphes de Scree suivants avec les courriels spams et non-spams:

Spams



Non-spams



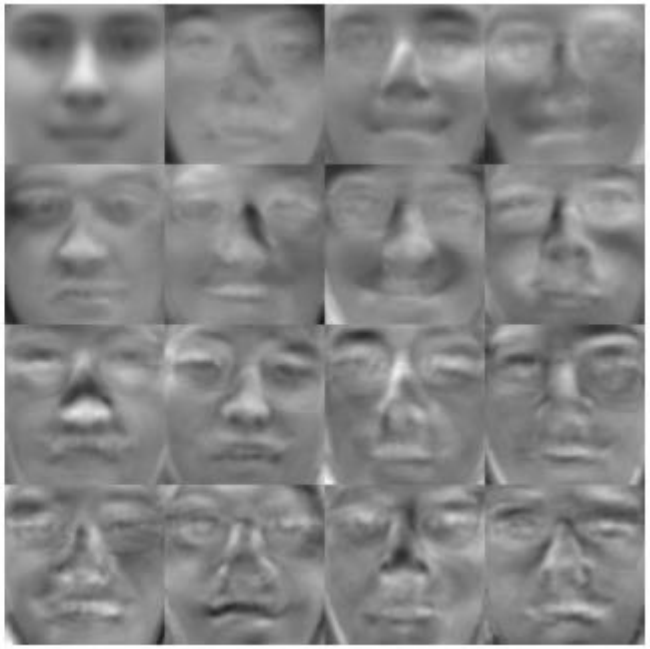
https://archive.ics.uci.edu/ml/datasets/Spambase

Application de l'ACP

L'un des premiers algorithmes de la reconnaissance faciale est basé sur l'ACP (Eigenface)

Données d'apprentissage

M=15 premières CPs



Extraction d’attributs:
Analyse discriminante linéaire (ADL)

Analyse discriminante linéaire (ADL)

- L'ACP est une méthode qui réduit la dimensionnalité des données en préservant l'information de variance.

- Lorsqu'on a besoin de l'information de classification, l'ACP ne possède pas de mécanisme pour l'encoder.

- L'analyse discriminante linaire (ADL) est une méthode qui permet de réduire la dimensions des données en préservant la plus forte discrimination entre les classes.

- Contrairement à l'ACP, l'ADL est une méthode supervisée.

Analyse discriminante linéaire (ADL)

- Soit une classification à $K = 2$ classes (C_1 et C_2).
- Soit $\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(N)}, y^{(N)})\}$ un ensemble de N données où $y^{(i)} = 1$ si $x^{(i)} \in C_1$ et $y^{(i)} = 0$ si $x^{(i)} \in C_2$.
- On voudrait trouver une direction de projection $\boldsymbol{w}$ telle que $\|\boldsymbol{w}\| = 1$ et $z_1 = \boldsymbol{w}^T x$ a le maximum de séparation entre les classes C_1 et C_2.
- Soit $\boldsymbol{\mu}^{(1)}$ et $\boldsymbol{\mu}^{(2)}$ les moyennes des données dans les classes C_1 et C_2, respectivement. Soit $m_1 \in \mathbb{R}$ et $m_2 \in \mathbb{R}$ les moyennes des classes après la projection des données sur $\boldsymbol{w}$.

Analyse discriminante linéaire (ADL)

Exemple: ($D = 2$)

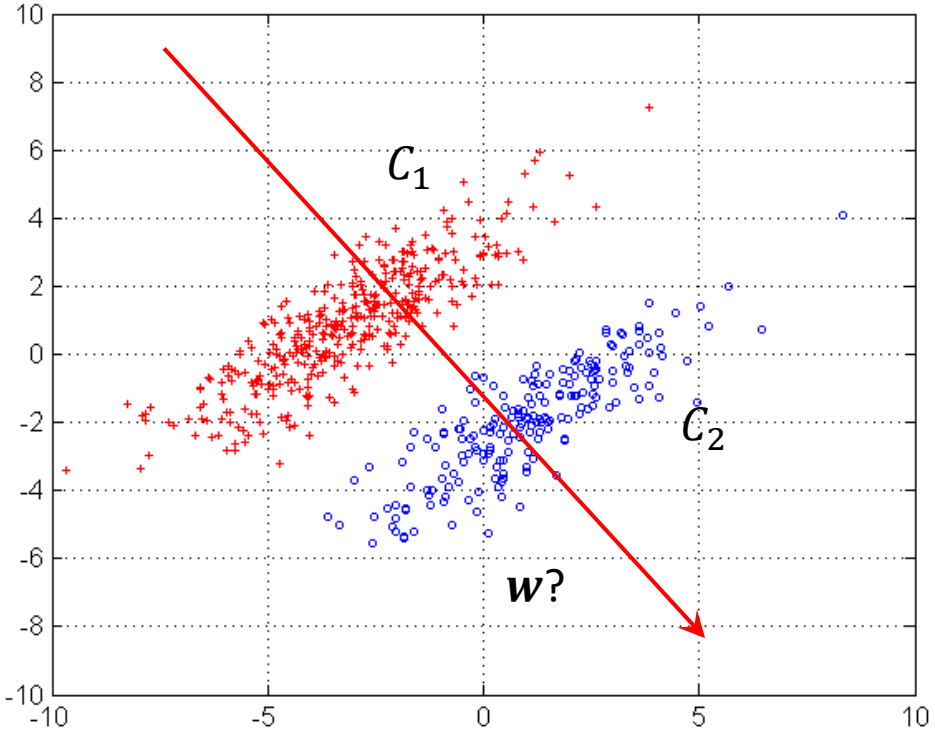
Quelle est la direction de projection $\boldsymbol{w}$ qui maximise la discrimination entre les deux classes C_1 et C_2?

$$\boldsymbol{\mu}^{(1)} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

$$\boldsymbol{\mu}^{(2)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

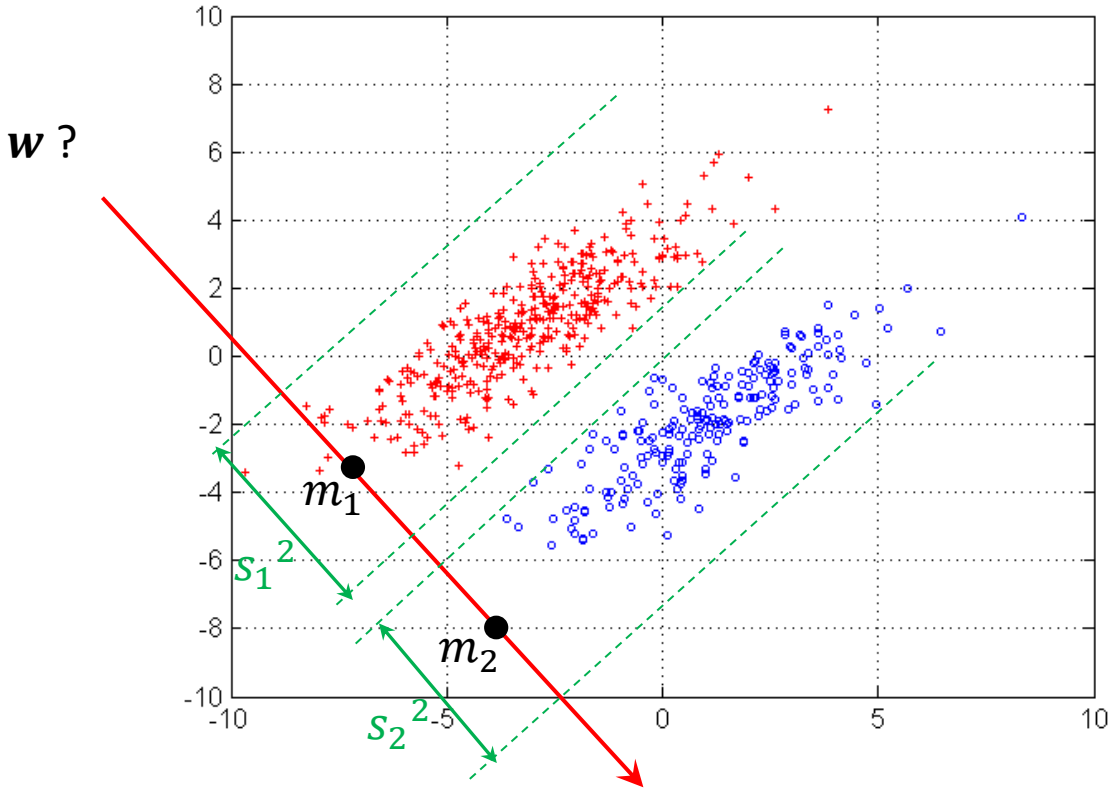
$$\boldsymbol{\Sigma}^{(1)} = \begin{bmatrix} 4 & 3 \\ 3 & 3 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{(2)} = \begin{bmatrix} 4 & 3 \\ 3 & 3 \end{bmatrix}$$



Analyse discriminante linéaire (ADL)

Soit m_1 et m_2 les moyennes des données de C_1 et C_2 après leurs projections sur $\boldsymbol{w}$ et s_1 et s_2 leurs variances.



Analyse discriminante linéaire (ADL)

- On remarque que:

$$m_1 = \frac{\sum_{i=1}^{N} \boldsymbol{w}^T x^{(i)} y^{(i)}}{\sum_{i=1}^{N} y^{(i)}} = \boldsymbol{w}^T \boldsymbol{\mu}^{(\mathbf{1})}$$

$$m_2 = \frac{\sum_{i=1}^{N} \boldsymbol{w}^T x^{(i)} (1 - y^{(i)})}{\sum_{i=1}^{N} (1 - y^{(i)})} = \boldsymbol{w}^T \boldsymbol{\mu}^{(\mathbf{2})}$$

$${s_1}^2 = \sum_{i=1}^{N} (\boldsymbol{w}^T x^{(i)} - m_1)^2 y^{(i)}$$

$${s_2}^2 = \sum_{i=1}^{N} \left(\boldsymbol{w}^T x^{(i)} - m_2\right)^2 (1 - y^{(i)})$$

Analyse discriminante linéaire (ADL)

- La direction $\boldsymbol{w}$ qui maximise la discrimination entre C_1 et C_2 est celle qui maximise la quantité:

$$J(\boldsymbol{w}) = \frac{(m_1 - m_2)^2}{{s_1}^2 + {s_2}^2}$$

- On remarque que :

$$(m_1 - m_2)^2 = (\boldsymbol{w}^T \boldsymbol{\mu}^{(1)} - \boldsymbol{w}^T \boldsymbol{\mu}^{(2)})^2$$

$$= \boldsymbol{w}^T (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})^T \boldsymbol{w}$$

$$= \boldsymbol{w}^T \boldsymbol{S}_{inter} \boldsymbol{w}$$

- On appelle $\boldsymbol{S}_{inter}$ la matrice de variance interclasses.

Analyse discriminante linéaire (ADL)

- Par ailleurs, on a:

$$s_1{}^2 = \sum_{i=1}^{N} (\boldsymbol{w}^T x^{(i)} - m_1)^2 y^{(i)}$$

$$= \sum_{i=1}^{N} (\boldsymbol{w}^T x^{(i)} - m_1)(\boldsymbol{w}^T x^{(i)} - m_1)^T y^{(i)}$$

$$= \sum_{i=1}^{N} \boldsymbol{w}^T (x^{(i)} - \boldsymbol{\mu}^{(1)})(x^{(i)} - \boldsymbol{\mu}^{(1)})^T \boldsymbol{w}\, y^{(i)}$$

$$= \boldsymbol{w}^T \boldsymbol{S_1} \boldsymbol{w}$$

- On appelle $\boldsymbol{S_1}$ la matrice de variance intra-classe de C_1 .

Analyse discriminante linéaire (ADL)

- De même, on définira la matrice intra-classe de C_2 : $\boldsymbol{S_2}$.
- La variance totale intra classe après projection est:

$$s_1{}^2 + s_2{}^2 = \boldsymbol{w}^T \boldsymbol{S_1} \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{S_1} \boldsymbol{w}$$

$$= \boldsymbol{w}^T (\boldsymbol{S_1} + \boldsymbol{S_2}) \boldsymbol{w}$$

$$= \boldsymbol{w}^T \boldsymbol{S}_{intra}\, \boldsymbol{w}$$

- La fonction à maximiser sur $\boldsymbol{w}$ est alors donnée par:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_{inter}\, \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_{intra}\, \boldsymbol{w}}$$

Analyse discriminante linéaire (ADL)

- Après la dérivation de $J(\boldsymbol{w})$ par rapport à $\boldsymbol{w}$, on obtient:

$$\boldsymbol{w} \approx \boldsymbol{S}_{intra}{}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$$

Exemple:

- Pour notre exemple

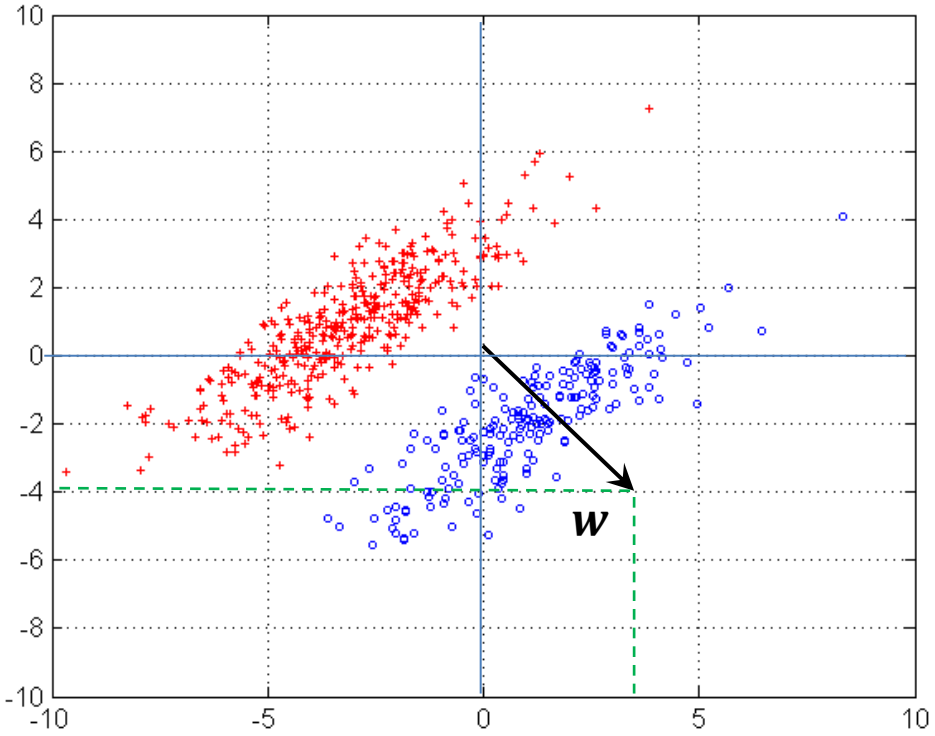
$$\boldsymbol{S}_{intra} = \boldsymbol{S}_1 + \boldsymbol{S}_2 = 2\begin{bmatrix} 4 & 3 \\ 3 & 3 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{S}_{intra}{}^{-1} \approx \begin{bmatrix} 0.5 & -0.50 \\ -0.5 & 0.67 \end{bmatrix}$$

$$\boldsymbol{w} = \boldsymbol{S}_{intra}{}^{-1}\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\right) = \begin{bmatrix} 0.5 & -0.50 \\ -0.5 & 0.67 \end{bmatrix}\begin{bmatrix} 4 \\ -3 \end{bmatrix}$$

$$= \begin{bmatrix} 3.5 \\ -4 \end{bmatrix}$$

Analyse discriminante linéaire (ADL)

- Le vecteur $\boldsymbol{w}$ obtenu est en effet celui désiré.



Références

1. M. S. Allili. Techniques d'apprentissage automatique (Cours de 2e cycle). Université du Québec en Outaouais (UQO), Québec, Canada. Hivers 2015.
2. S. Rogers et M Girolami. A first Course in machine learning, CRC press, 2012.
3. C. Bishop. Pattern Recognition and Machine learning. Springer 2006.
4. R. Duda, P. Storck et D. Hart. Pattern Classification. Prentice Hall, 2002.