Exploration visuelle des données

Nicoleta ROGOVSCHI

nicoleta.rogovschi@parisdescartes.fr

M2-INFO

Analyse Discriminante Linéaire (ADL)

Plan du cours

- Introduction et définitions
- ADL pour 2 classes
- ADL pour classes
- Exemple
- ACP vs ADL
- Conclusions

Réduction des dimensions par extraction de caractéristiques

Deux grandes familles de méthodes :

- **Méthodes linéaires**
  - Analyse en Composantes Principales (ACP)
  - Analyse Discriminante Linéaire (ADL)
  - Multi-Dimensional Scaling (MDS)
  - …
- **Méthodes non-linéaires**
  - Isometric feature mapping (Isomap)
  - Locally Linear Embedding (LLE)
  - Kernel PCA
  - Segmentation spectrale (spectral clustering)
  - Methodes supervisées (S-Isomap)
  - …

Introduction

- Analyse Discriminante Linéaire
  - Une méthode pour l'analyse de données de grande dimension dans le cas de l’apprentissage supervisé (les classes (les labels) sont disponibles dans l’ensemble de données)
  - Elle trouve un espace à faible dimension optimal telle que, lorsque les points sont projetés, les données de différentes classes sont bien séparées
  - Utile pour l’extraction des caractéristiques pour faciliter la classification supervisée

Introduction

- ADL tente de déterminer la contribution des variables qui expliquent l'appartenance des individus à des groupes.
- L'analyse discriminante linéaire permet aussi d'affecter de nouveaux individus aux groupes.

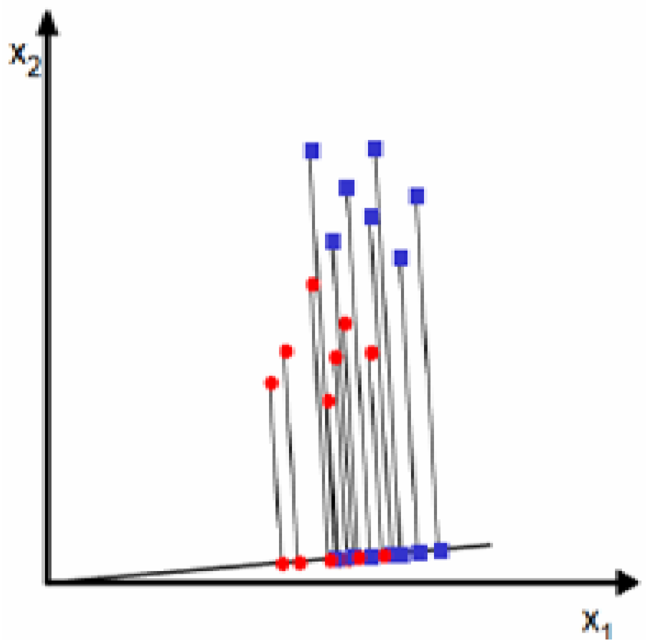
ADL pour 2 classes

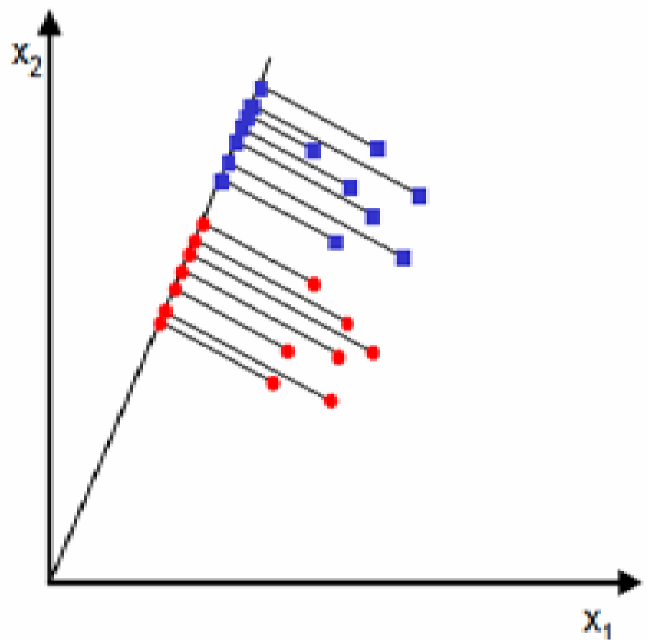
- L'objectif principal d'ADL est de réaliser une réduction de dimensions tout en préservant le plus d'information discriminatoire possible de chaque classe
  - On suppose qu'on a un ensemble d'échantillons de D-dimensions $\left\{x^{(1}, x^{(2}, \ldots, x^{(N}\right\}$, $\boldsymbol{N_1}$ desquelles appartiennent à la classe $\boldsymbol{r_1}$ et $\boldsymbol{N_2}$ à la classe $\boldsymbol{r_2}$.
  - Nous chercheons à obtenir un scalaire **y** en projetant les échantillons $\boldsymbol{x}$ sur une ligne $\boldsymbol{w}$

$$y = w^T x$$

ADL pour 2 classes

- De toutes les lignes possibles on veut trouver celle qui maximise la séparabilité des classes





ADL pour 2 classes

Afin de trouver un bon vecteur de projection, on doit définir une mesure de séparation entre les projections.

– Le vecteur de la moyenne de chaque classe de ***x*** et ***y*** espaces des caractéristiques est

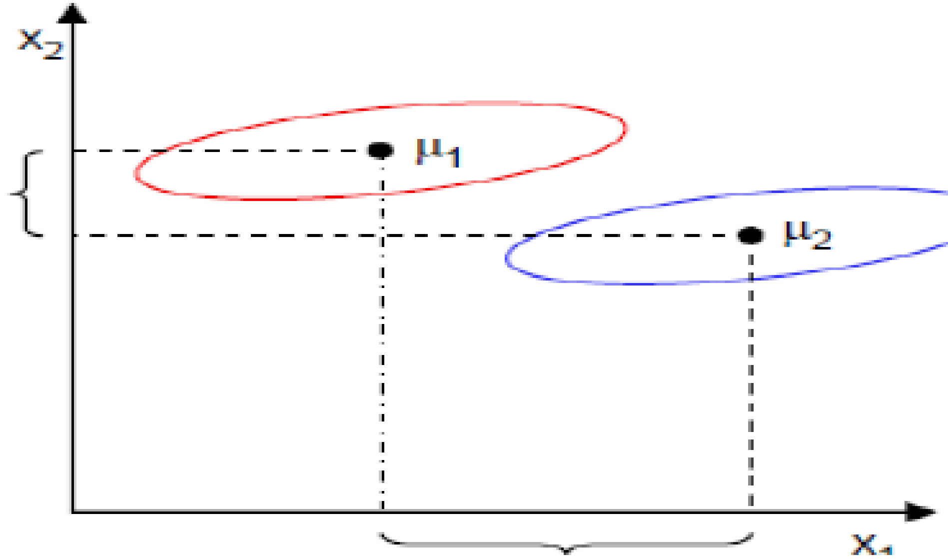
$$\mu_i = \frac{1}{N_i} \sum_{x \in r_i} x \quad \text{et} \qquad \tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in r_i} y = \frac{1}{N_i} \sum_{x \in r_i} w^T x = w^T \mu_i$$

ADL pour 2 classes

- On peut choisir comme fonction objective la distance entre les moyennes projetées

$$J(w) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|w^T(\tilde{\mu}_1 - \tilde{\mu}_2)\right|$$

- Néanmoins, la distance entre les moyennes projetée n'est pas une très bonne mesure, car elle ne tient pas compte de l'écart-type à l'intérieur des classes



ADL pour 2 classes

- La solution proposée par Fisher est de maximiser une fonction qui représente la différence entre les moyennes, normalisée par une mesure de dispersion intra-classe
  - Pour chaque classe on définit la dispersion, qui est équivalente à la variance, comme suit:

$$\tilde{s}_i^{\,2} = \sum_{y \in r_i} (y - \tilde{\mu}_i)^2$$

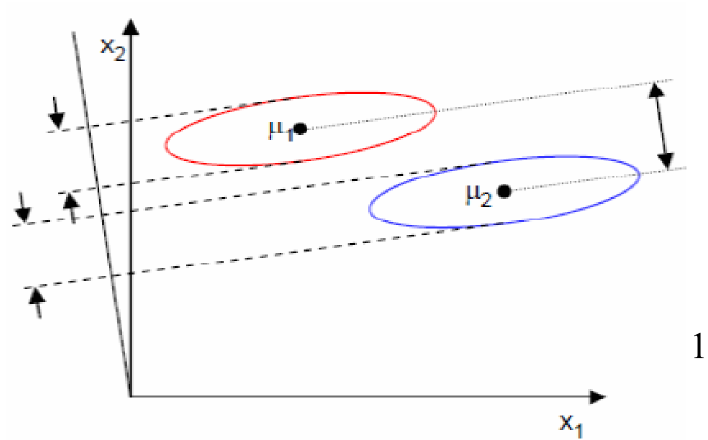
ou la quantité $(\tilde{s}_1^2 + \tilde{s}_2^2)$ est appelée la <u>variance intra-classes</u> des observations projetées

ADL pour 2 classes

Le discriminant linéaire de Fisher est définie comme une fonction linéaire $\boldsymbol{w^T x}$ qui maximise le critère suivant :

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

Toutefois, nous allons chercher une projection ou les exemples de la même classe sont proche les uns des autres et en même temps les moyennes projetées seront le plus éloignées possible.



ADL pour 2 classes

- Afin de trouver la projection optimale w*, on doit exprimer J(w) comme une fonction explicite de **w**
- On définit une mesure de dispersion dans l'espace de dimensions multivariées **x**, qu'on appelle matrices de covariance.

$$S_i = \sum_{x \in r_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_1 + S_2 = S_w$$

ou S_W est appelé <u>matrice de covariance intra-classes</u>

ADL pour 2 classes

- La projection de ***y*** peut être exprimé comme une fonction de matrice de covariance dans l'espace de caractéristiques ***x***

$$\tilde{s}_i^{\,2} = \sum_{y \in r_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in r_i} (w^T x - w^T \mu_i)^2 = \sum_{x \in r_i} w^T (x - \mu_i)(x - \mu_i)^T w = w^T S_i w$$

$$\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2} = w^T S_w w$$

- D'une manière similaire, la différence entre les moyennes projetées peut être exprimé en terme de moyennes dans l'espace d'origine

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2) w = w^T S_B w$$

s_B

ADL pour 2 classes

- La matrice S_B est appelé la <u>matrice de covariance inter-classes.</u>

- On peut exprimer le critère de Fisher en terme de S_W et S_B de la manière suivante :

$$J(w) = \frac{w^T S^B w}{w^T S_w w}$$

ADL pour 2 classes

- Pour trouver le maximum de J(W) on va calculer les dérivées de J(W) et égaler à zéro.

$$\frac{d}{dw}\left[J(w)\right]=\frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right]=0 \Rightarrow$$

$$\Rightarrow \left[w^T S_W w\right]\frac{d\left[w^T S_B w\right]}{dw}-\left[w^T S_B w\right]\frac{d\left[w^T S_W w\right]}{dw}=0 \Rightarrow$$

$$\Rightarrow \left[w^T S_W w\right]2S_B w-\left[w^T S_B w\right]2S_W w=0$$

- On divise par $w^T S_W w$

$$\left[\frac{w^T S_W w}{w^T S_W w}\right]S_B w-\left[\frac{w^T S_B w}{w^T S_W w}\right]S_W w=0 \Rightarrow$$

$$\Rightarrow S_B w - J S_W w = 0 \Rightarrow$$

$$\Rightarrow S_W^{-1} S_B w - J w = 0$$

ADL pour 2 classes

- Résoudre le problème $S_W^{-1} S_B w = Jw$ nous mène à

$$w^* = \arg\max_w \left\{ \frac{w^T S_B w}{w^T S_w w} \right\} = S_w^{-1}(\mu_1 - \mu_2)$$

- Ce qui représente le critère linéaire discriminant de Fisher (1936).

LDA for Multiple Classes

- On peut facilement généraliser l'ADL de Fisher pour un problème à C-classes.
  - A la place d'une projection y, on va maintenant chercher (C-1) projections $[y_1, y_2, \ldots, y_{C-1}]$. On va avoir (C-1) vecteurs de projection w_i qu'on peut arranger par colonnes dans une matrice de projection
    $W=[w_1|w_2|\ldots| w_{C-1}]$:

$$y_i = w_i^T x \Rightarrow y = W^T x$$

ADL multi-classes

- Adaptation des formules
  - La projection intra-classes est égale à

$$S_W = \sum_{i=1}^{c} S_i$$

ou $S_i = \sum_{x \in r_i} (x - \mu_i)(x - \mu_i)^T$ et $\mu_i = \frac{1}{N_i} \sum_{x \in r_i} x$

  - La projection inter-classes est égale à

$$S_B = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T$$

ou $\mu = \frac{1}{N} \sum_{\forall x} x = \frac{1}{N} \sum_{x \in r_i} N_i \mu_i$

Et $S_T = S_B + S_W$ est appelé la ***matrice totale des projections.***

ADL multi-classes

- D'une manière similaire on peut définir le vecteur des moyennes et les matrices des covariances pour les exemples projetés :

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in r_i} y \qquad \tilde{S}_W = \sum_{i=1}^{C} \sum_{y \in r_i} (y - \tilde{\mu}_i)(y - \tilde{\mu}_i)^T$$

$$\tilde{\mu} = \frac{1}{N} \sum_{\forall y} y \qquad \tilde{S}_B = \sum_{i=1}^{C} N_i (\tilde{\mu}_i - \mu)(\tilde{\mu}_i - \mu)^T$$

- A partir des dérivées calculées pour le problème à 2 classes, on peut écrire:

$$\tilde{S}_W = W^T S_W W$$

$$\tilde{S}_B = W^T S_B W$$

ADL multi-classes

- On rappelle qu'on cherche une projection qui maximise le taux entre la projection inter-class et la projection intra-class. Comme la projection n'est qu'un scalaire (elle a C-1 dimensions), on peut utiliser le déterminant des matrices des projections pour obtenir une fonction objective scalaire:

$$J(W) = \frac{\left|\tilde{S}_B\right|}{\left|\tilde{S}_W\right|} = \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|}$$

- On va chercher la matrice de projection W* qui maximise ce taux

ADL multi-classes

- On peut monter que la matrice de projection optimale W* est celle dont les colonnes sont les vecteurs propres qui correspondent au plus grandes valeurs propres du problème suivant :

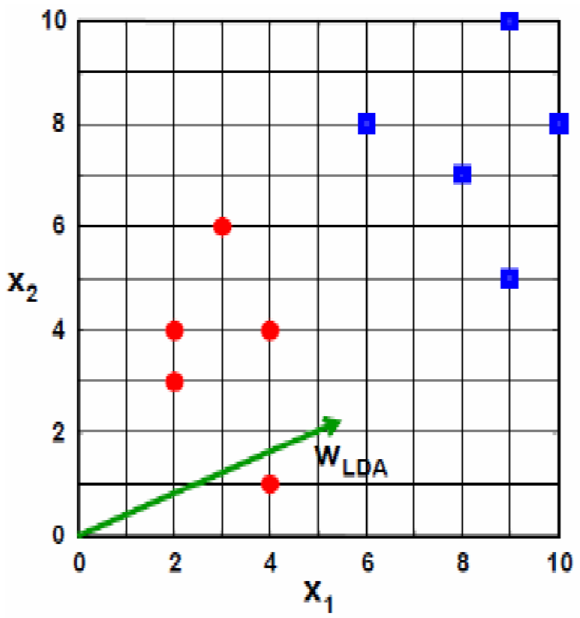
$$W^* = \left[w_1^* \middle| w_2^* \middle| \ldots \middle| w_{C-1}^*\right] = \arg\max_{W} \left\{ \frac{\left|W^T S_B W\right|}{\left|W^T S_w W\right|} \right\} = (S_B - \lambda_i S_W) w_i^* = 0$$

Exemple

- On a un jeu de données bidimensionnel

  $X1=(x_1, x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$

  $X2=(x_1, x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$



- Solution:
  - Statistiques de base:

$$S_1=\begin{bmatrix}0.80 & -0.40\\ -0.40 & 2.60\end{bmatrix}; \qquad S_2=\begin{bmatrix}1.84 & -0.04\\ -0.04 & 2.64\end{bmatrix};$$

$$\mu_1=[3.00 \quad 3.60]; \qquad \mu_2=[8.40 \quad 7.60];$$

  - Les projections inter- et intra- classes sont :

$$S_B=\begin{bmatrix}29.16 & 21.60\\ 21.60 & 16.00\end{bmatrix}; \qquad S_W=\begin{bmatrix}2.64 & -0.44\\ -0.44 & 5.28\end{bmatrix};$$

Exemple

– La projection ADL est obtenu comme solution au problème suivant :

$$S_W^{-1}S_B v = \lambda v \Rightarrow \left|S_W^{-1}S_B - \lambda I\right| = 0 \Rightarrow \begin{vmatrix} 11.89-\lambda & 8.81 \\ 5.08 & 3.76-\lambda \end{vmatrix} = 0 \Rightarrow \lambda = 15.65$$

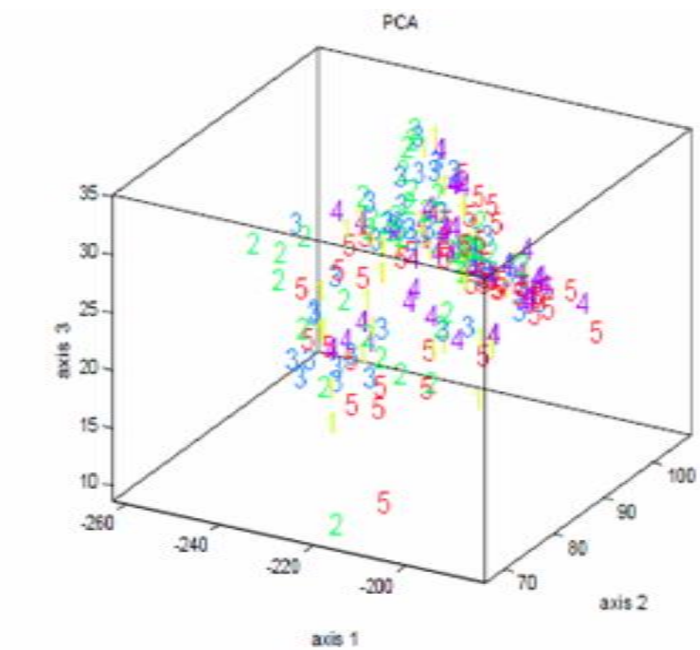
$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 15.65\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \Rightarrow \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$$

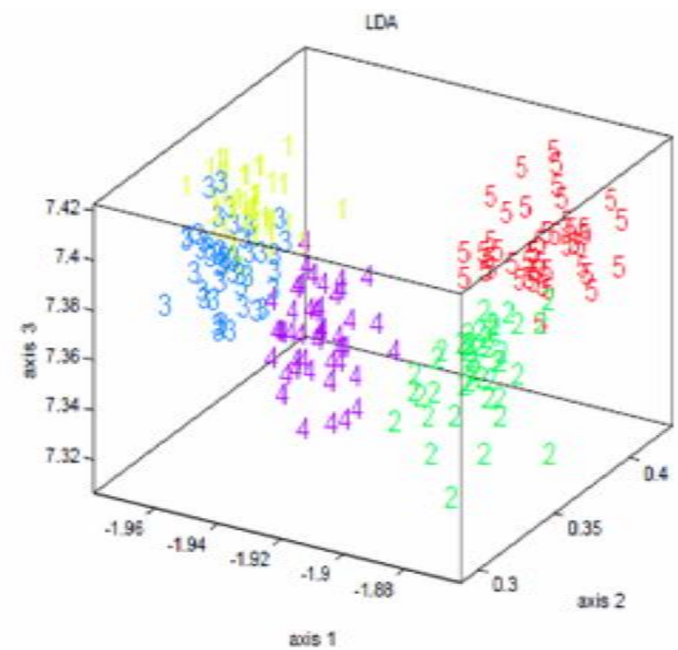
– Ou directement par :

$$W^* = S_W^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} -0.91 & -0.39 \end{bmatrix}^T$$

ACP vs ADL

- On a un jeu de données qui caractérise 5 types de café, de taille 45 x 60





ACP vs ADL

- Pour des jeux de données de petite taille, ACP donne des meilleurs résultats que l'ADL.
- Quand le nombre d'exemples est assez grand et représentatif pour chaque classe, l'ADL montre des meilleurs performances que l'ACP.

Conclusions

- ADL est une méthodes simple assez connue pour le traitement de données de grande tailles, quand les étiquettes des classes sont disponibles.
- C'est une méthode linéaire pour la réductions des dimensions en projetant les données d'origine dans un espace de dimensions C-1

Conclusions

- Il a une série de limitations dans l'ADL classique
- Il existe plusieurs extensions de l'ADL standard (ADL généralisée, ADL non-paramétrique, ADL orthonormé), qui essayent de dépasser ces défauts.