

Prof. Mohamed Aly LOULY

Data Science et Machine Learning

1

Analyse Discriminante (AD)

Classification supervisée

Classement ou Scoring

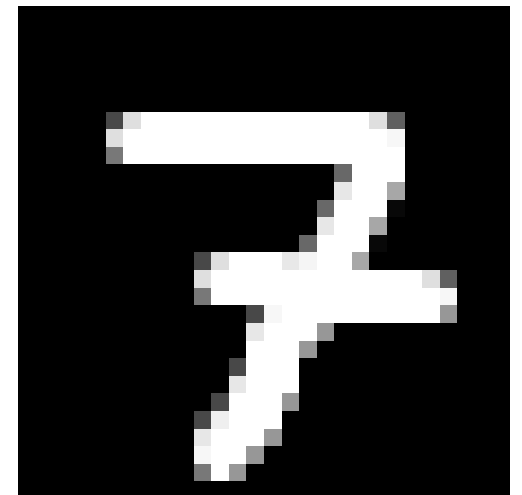
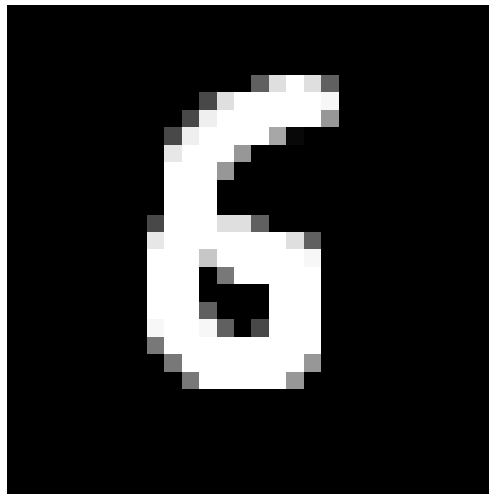
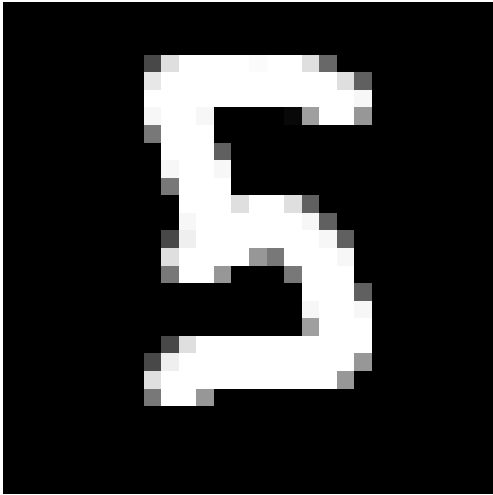
LES DEUX MÉTHODES DE L'AD

L'Analyse Discriminante (AD) contient deux méthodes :

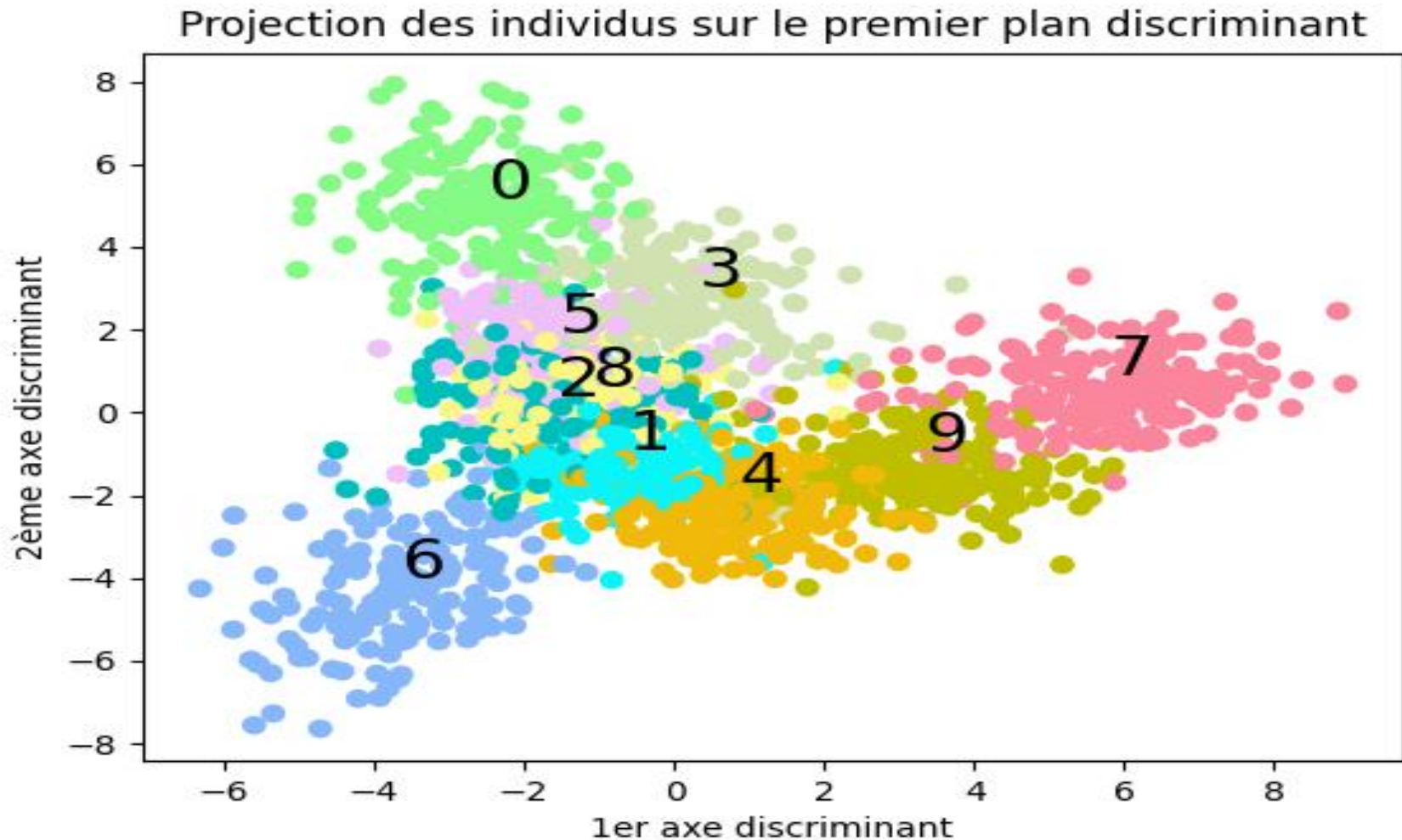
L'Analyse Factorielle Discriminante (AFD) : une méthode géométrique descriptive qui permet de visualiser des individus sur un plan en maximisant la discrimination entre leurs classes.

L'Analyse Discriminante Décisionnelle (ADD) : une méthode géométrique et probabiliste qui explique une variable qualitative par plusieurs variables quantitatives. Elle permet de faire le classement de chaque nouvel individu à la classe qui lui correspond le mieux.

L'ADD POUR LA RECONNAISSANCE DES FORMES



L'AFD POUR LA RECONNAISSANCE DES FORMES



LES DONNÉES DE L'AD

L'Analyse Discriminante (AD) fonctionne avec n individus appartenant à m classes différentes pour lesquels nous avons observé p variables.

Les observations sont données par une matrice X de taille (n, p) . Les lignes de cette matrice représentent les individus et les colonnes représentent les variables. Le coefficient x_{ij} désigne l'observation de la variable j sur l'individu i .

Les classes des individus sont données par un vecteur Y de taille n , prenant des valeurs entières entre 1 et m . La valeur y_i précise la classe de l'individu i .

CENTRAGE ET NORMALISATION DES DONNÉES

Il est souvent nécessaire de réduire toutes les variables pour uniformiser leur contribution à la distance entre les individus. Il est également plus pratique de les centrer. On note donc par **Z** le tableau des données ainsi transformées. :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

VARIANCE TOTALE ET VARIANCES DES CLASSES

La matrice de variance covariance est la suivante pour les données centrées réduites :

$$T = \frac{1}{n} Z^T Z = \frac{1}{n} \sum_{i=1}^n Z_i^T Z_i$$

La matrice de variance covariance W_k de chaque classe est donnée par l'expression suivante, où g_k désigne la moyenne des individus de la classe C_k et n_k leur nombre :

$$W_k = \frac{1}{n_k} \sum_{i \in C_k} Z_i^T Z_i - g_k^T g_k$$

VARIANCES ENTRE ET INTRA CLASSES

La matrice de variance entre classes (**Between**) est la matrice de variance covariance des moyennes :

$$B = \sum_{k=1}^m \frac{n_k}{n} g_k^T g_k$$

La matrice de variance intra classes (**Within**) est la moyenne des matrices de variance covariance des classes :

$$W = \sum_{k=1}^m \frac{n_k}{n} W_k$$

DÉCOMPOSITION DE LA MATRICE DE VARIANCE COVARIANCE

La matrice de variance covariance totale est la somme des variances inter classes et intra classes :

$$T = B + W$$

Variance Totale = Variance Expliquée (B) + Variances Résiduelles (W)

LES AXES DISCRIMINANTS

Les axes discriminants sont donnés par les vecteurs propres $u_1, u_2, \dots, u_{m-1}$ de $T^{-1}B$, associés à ses valeurs propres non nulles $\lambda_1, \lambda_2, \dots, \lambda_{m-1}$, énumérées dans l'ordre décroissant.

LA PROJECTION DES INDIVIDUS

La projection des individus sur l'axe discriminant k est alors :

$$F_k = Z u_k$$

En désignant par U la matrice dont les colonnes sont les vecteurs $u_1, u_2, \dots, u_q$, on obtient les différentes projections dans les colonnes de la matrice F suivante : $F = Z U$.

Les centres de gravité g_k peuvent être projetés de la même manière sur les axes discriminants.

AFD POUR EXPLIQUER LA PAUVRETÉ PAR L'ACCÈS À L'EAU

Les 53 moughataas de la Mauritanie ont été classées dans 3 groupes selon le taux de la pauvreté. La 1^{ère} classe regroupe les moughataas pour lesquelles le taux de pauvreté n'atteint pas le 1/3 de la population. La 2^{ème} classe regroupe celles dont le taux est entre 1/3 et 2/3. La 3^{ème} est celle des moughataas dont le taux de pauvreté dépasse les 2/3.

Nous souhaitons expliquer ce classement par 7 variables qui indiquent le taux de la population utilisant chaque moyen d'accès à l'eau : citerne, revendeur, fontaine, fleuve, pluie, robinet, puits.

MATRICE DE VARIANCE COVARIANCE TOTALE

<i>T</i>	citerne	revendeur	fontaine	fleuve	pluie	robinet	puits
citerne	1,00	-0,06	-0,11	-0,21	-0,08	-0,07	-0,36
revendeur	-0,06	1,00	0,30	-0,32	-0,13	-0,26	-0,60
fontaine	-0,11	0,30	1,00	-0,11	-0,09	-0,12	-0,17
fleuve	-0,21	-0,32	-0,11	1,00	0,21	-0,30	0,46
pluie	-0,08	-0,13	-0,09	0,21	1,00	-0,03	0,09
robinet	-0,07	-0,26	-0,12	-0,30	-0,03	1,00	-0,38
puits	-0,36	-0,60	-0,17	0,46	0,09	-0,38	1,00

MATRICE DE VARIANCE COVARIANCE ENTRE CLASSES

<i>B</i>	citerne	revendeur	fontaine	fleuve	pluie	robinet	puits
citerne	0,03	0,08	0,04	-0,06	-0,01	0,04	-0,08
revendeur	0,08	0,34	0,15	-0,23	-0,10	0,16	-0,42
fontaine	0,04	0,15	0,07	-0,10	-0,04	0,07	-0,17
fleuve	-0,06	-0,23	-0,10	0,16	0,05	-0,11	0,26
pluie	-0,01	-0,10	-0,04	0,05	0,04	-0,04	0,13
robinet	0,04	0,16	0,07	-0,11	-0,04	0,07	-0,19
puits	-0,08	-0,42	-0,17	0,26	0,13	-0,19	0,53

MATRICE DE VARIANCE COVARIANCE INTRA CLASSES

W	citerne	revendeur	fontaine	fleuve	pluie	robinet	puits
citerne	0,97	-0,14	-0,15	-0,15	-0,07	-0,11	-0,27
revendeur	-0,14	0,66	0,15	-0,09	-0,03	-0,42	-0,18
fontaine	-0,15	0,15	0,93	-0,01	-0,05	-0,19	0,00
fleuve	-0,15	-0,09	-0,01	0,84	0,16	-0,19	0,20
pluie	-0,07	-0,03	-0,05	0,16	0,96	0,01	-0,04
robinet	-0,11	-0,42	-0,19	-0,19	0,01	0,93	-0,19
puits	-0,27	-0,18	0,00	0,20	-0,04	-0,19	0,47

INVERSE DE LA MATRICE DE VARIANCE COVARIANCE TOTALE

T^{-1}							
	3,09	3,49	0,45	0,77	0,34	2,92	4,01
	3,49	6,89	0,26	1,24	0,53	4,96	6,70
	0,45	0,26	1,17	0,14	0,11	0,53	0,65
	0,77	1,24	0,14	1,59	-0,12	1,14	0,77
	0,34	0,53	0,11	-0,12	1,10	0,39	0,56
	2,92	4,96	0,53	1,14	0,39	4,95	5,44
	4,01	6,70	0,65	0,77	0,56	5,44	8,23

LA MATRICE PRINCIPALE DE L'AFD

$T^{-1}B$							
	0,09	0,09	0,07	-0,15	0,05	0,05	-0,01
	0,18	0,33	0,19	-0,34	0,02	0,16	-0,26
	0,03	0,07	0,04	-0,06	0,00	0,03	-0,06
	0,01	0,03	0,01	-0,02	-0,01	0,01	-0,03
	0,02	-0,03	0,00	-0,01	0,03	-0,01	0,06
	0,14	0,23	0,14	-0,25	0,03	0,11	-0,16
	0,12	-0,10	0,02	-0,10	0,17	-0,03	0,30

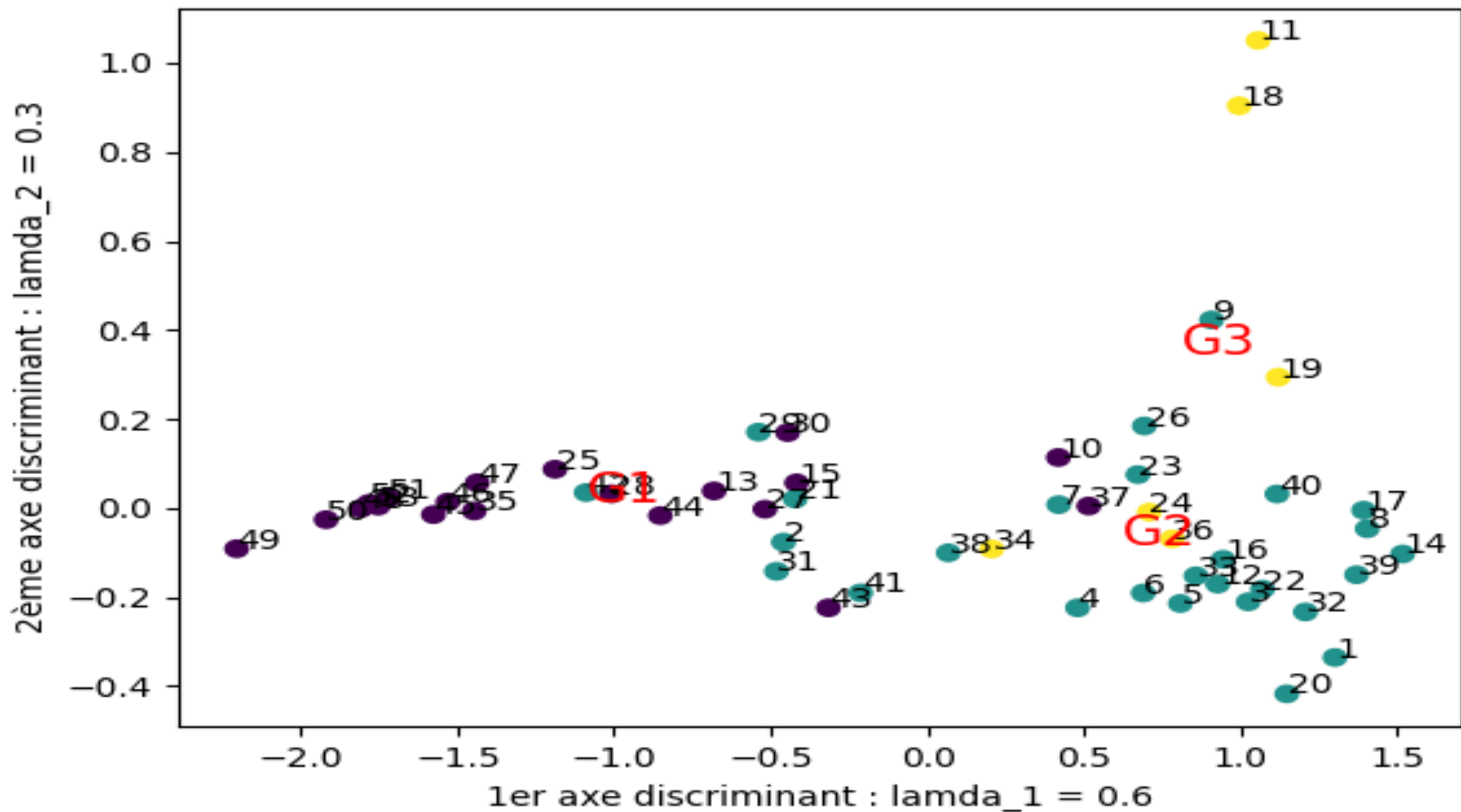
LES AXES DISCRIMINANTS

U1	U2
-0,19	0,35
-0,76	0,51
-0,16	0,08
-0,06	0,00
0,08	0,10
-0,52	0,42
0,29	0,65

Valeurs propres non nulles : $\lambda_1 = 0,6$ et $\lambda_2 = 0,3$.

PROJECTION PAR L'AFD

Projection des individus sur le premier plan discriminant



L'ANALYSE DISCRIMINANTE DÉCISIONNELLE ET PRÉDICTIVE

L'Analyse Discriminante Décisionnelle (ADD) est utilisée comme une méthode **décisionnelle** et **prédictive**. Elle permet de **classer** l'individu dans la classe qui lui correspond le mieux.

Il y a deux approches de classement : une approche **géométrique** et une approche **probabiliste**.

CLASSEMENT AVEC UNE RÈGLE GÉOMÉTRIQUE

La règle du classement géométrique consiste à affecter l'individu x à la classe C_k dont le centre de gravité g_k est le plus proche de cet individu. La distance généralement utilisée pour ce classement est celle qui est associée à la **métrique** de la matrice T^{-1} :

$$dist^2(x, g_k) = (x - g_k) T^{-1} (x - g_k)^T$$

La métrique associée à la matrice W^{-1} est parfois utilisée à la place de celle de la matrice T^{-1} .

CLASSEMENT AVEC LE MODÈLE PROBABILISTE BAYÉSIEN

La règle du modèle du classement probabiliste bayésien consiste à affecter l'individu x à la classe C_k la plus probable. C'est la classe qui maximise la probabilité conditionnelle $P(C_k/x)$:

$$P(C_k/x) = \frac{P(x/C_k)P(C_k)}{P(x)}$$

Il suffira de chercher la classe C_k qui maximise le numérateur car $P(x)$ ne dépend pas de k .

CLASSEMENT AVEC LE MODÈLE BAYÉSIEN NORMAL DE MÊME VARIANCE

Sous l'hypothèse de matrices de variance covariance identiques W , le score à minimiser devient :

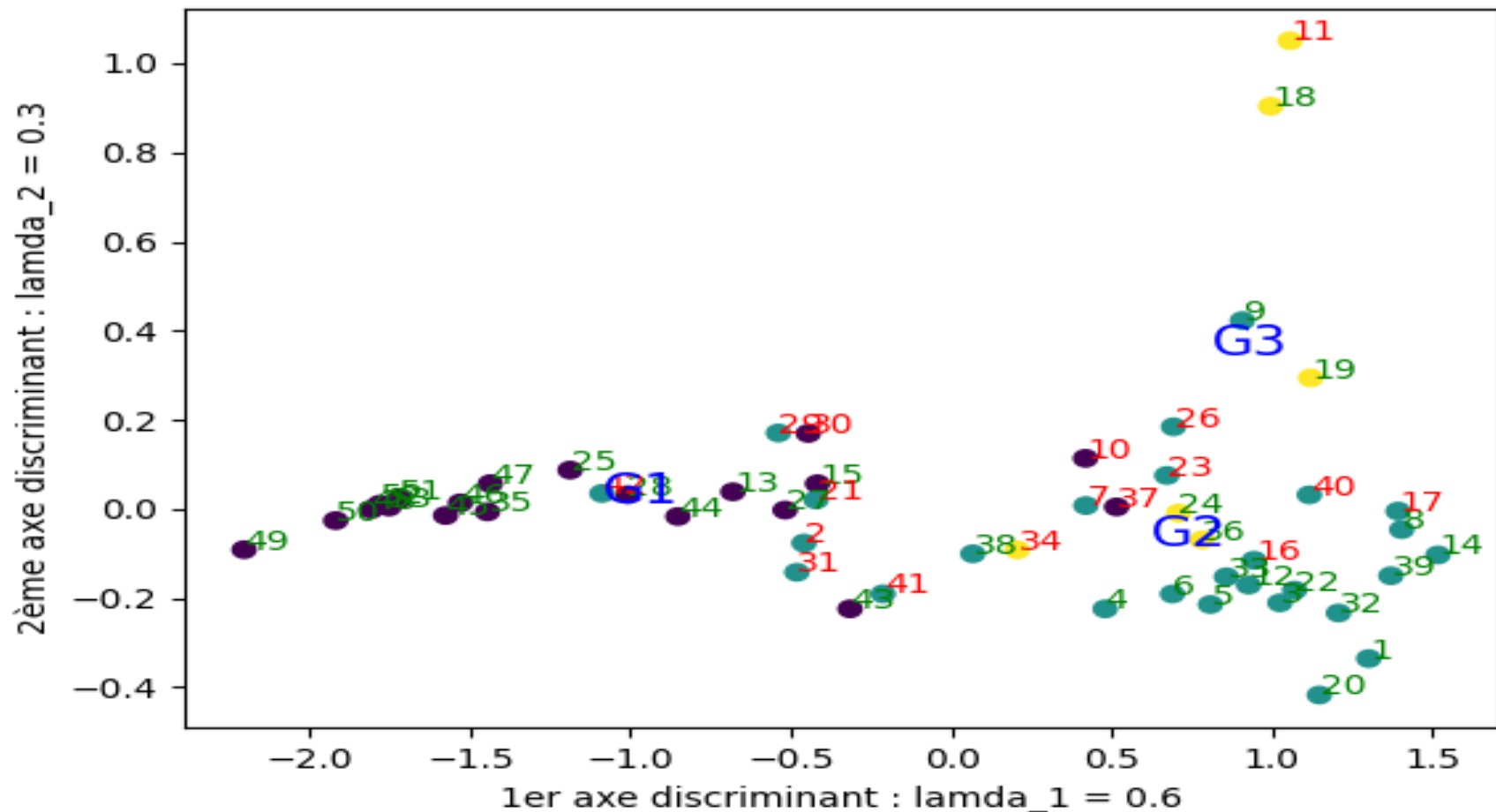
$$sc_k = (x - g_k) W^{-1} (x - g_k)^T - 2 \log(n_k)$$

On remarque que le premier terme est celui de la règle du classement géométrique avec la distance associée à la matrice W^{-1} .

Les deux classements sont donc identiques en cas de classes équiprobables, de mêmes nombres n_k .

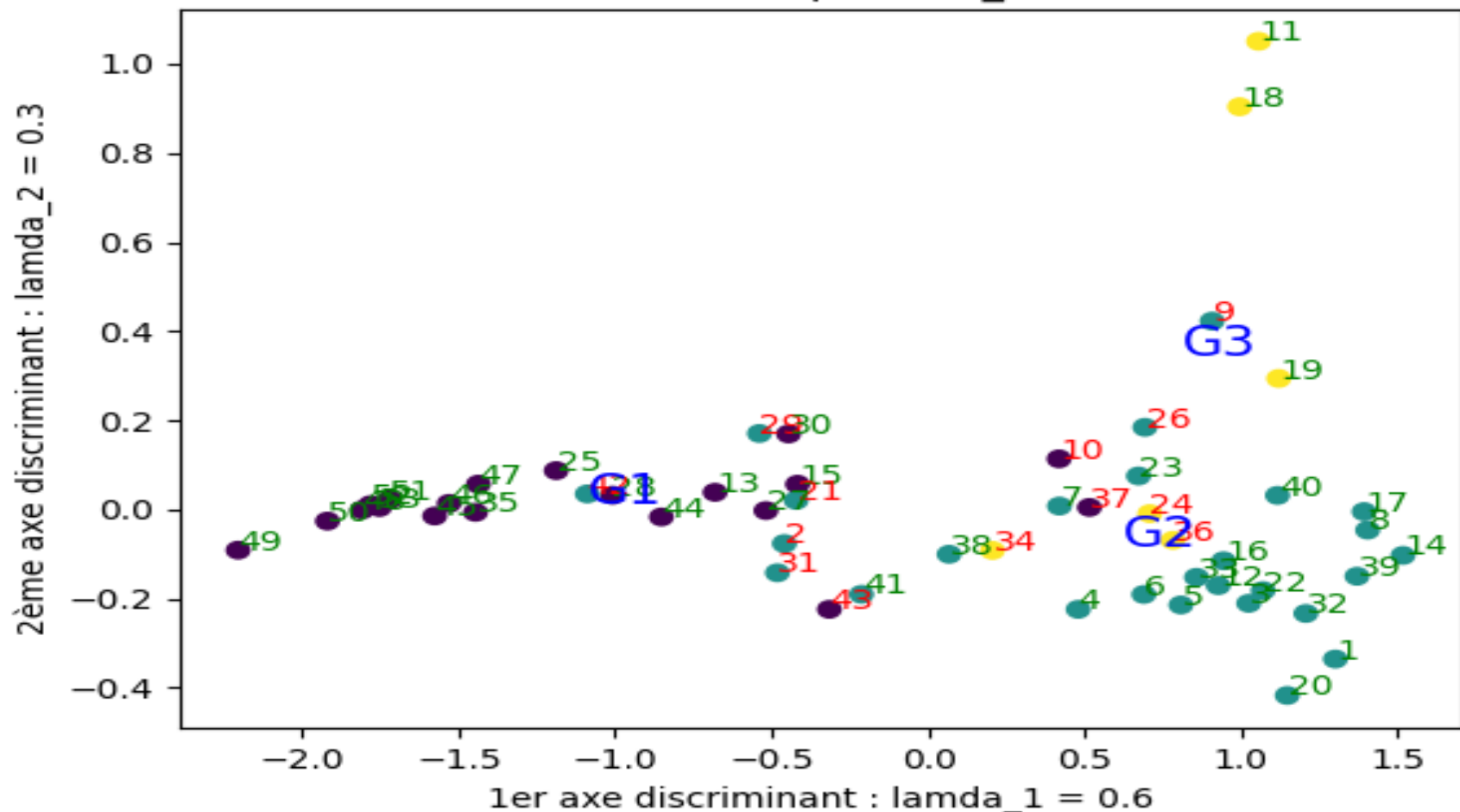
CLASSEMENT AVEC LA DISTANCE EUCLIDIENNE

Classement avec la distance euclidienne : $36/53 = 68\%$



CLASSEMENT AVEC LA DISTANCE ASSOCIÉE À LA MÉTRIQUE DE L'INVERSE DE T

Classement avec la métrique de T_{inv} : $40/53 = 75\%$



CLASSEMENT PROBABILISTE BAYÉSIEN NORMAL

Classement bayésien normal : $41/53 = 77\%$

