Licence d'Informatique 2
Analyse de Données Utilisateur (C5–160412)
TD 5 – Apprentissage Supervisé (Prédiction)

Carl FRÉLICOT – Dpt Info / Lab MIA

Les méthodes d'analyse d'un tableau de données, originellement descriptives et explicatives, peuvent être vues comme les fondations de la science des données (*Data Science*) et de l'apprentissage automatique (*Machine Learning*) pour l'intelligence artificielle en ce sens qu'elles permettent de :

- caractériser, une à une, deux par deux, les variables observées — [TD1, TD2]
- expliquer, réduire leur dimensionnalité — ACP, AFC, ACM [TD3]
- classifier (segmenter) les observations — *apprentissage non supervisé* [TD4]
- classer (prédire) de nouvelles observations — *apprentissage supervisé* [TD5]

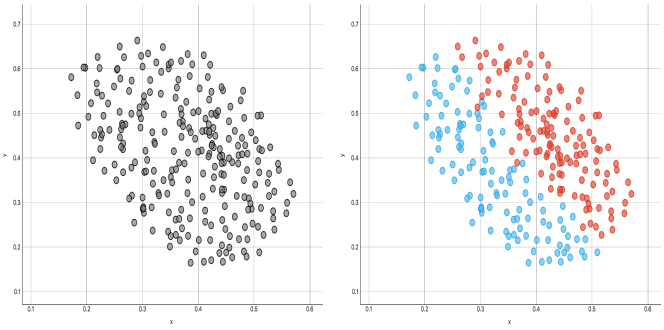
C'est ce dernier aspect qui est abordé dans cette séance. La classification supervisée consiste à *apprendre* d'un tableau de données pour lesquelles on possède (par expertise ou par *clustering*) une variable indicatrice d'appartenance à un groupe ce qui est nécessaire à la *prédiction* du classement dans l'un des groupes d'autres observations.
Il existe beaucoup d'approches de ce problème et un grand nombre de méthodes plus ou moins sophistiquées. Nous nous limitons aux données quantitatives et n'abordons en détail que les deux premières, fondées sur des notions déjà vues.

1. Approche Factorielle Discriminante — AFD

(a) Dans les figures ci-contre, quelle direction définit la variable (artificielle) décrivant le mieux le nuage ?
Le problème s'étend bien évidement à plus qu'une direction, et à $p > 2$ dimensions.

(b) À droite, il s'agit donc de trouver le sous-espace le plus *discriminant*. L'approche est ici *factorielle*[a], comme au TD3, le critère est différent. Sur quelles notions vues précédemment peut-il reposer ?

(c) Que suffirait-il alors de faire pour prédire le groupe d'une nouvelle observation (x, y) ?



[a] s'il y a m groupes, on montre qu'il y a $q = min(p, m-1)$ facteurs

Exercice 1

Le tableau ci-contre montre les notes de $n = 9$ élèves de terminale dans $p = 5$ matières et l'avis donné par le conseil de classe pour le baccalauréat portée sur le livret scolaire. L'avis est l'indicatrice de groupe.

	Prénom	math.	info.	fran.	angl.	arts	livret
1	Thomas	6.0	6.0	5.0	5.5	8.0	D
2	Margaux	8.0	8.0	8.0	8.0	9.0	D
3	Florian	6.0	7.0	11.0	10.0	11.0	P
4	Lucie	15.0	14.0	16.0	15.0	8.0	F
5	Victor	14.0	14.0	12.0	12.5	10.0	F
6	Elena	11.0	10.0	5.5	7.0	13.0	P
7	Hugo	5.5	7.0	14.0	11.5	10.0	P
8	Nabil	13.0	12.5	8.5	9.5	12.0	F
9	Juliette	9.0	9.5	12.5	12.0	18.0	F

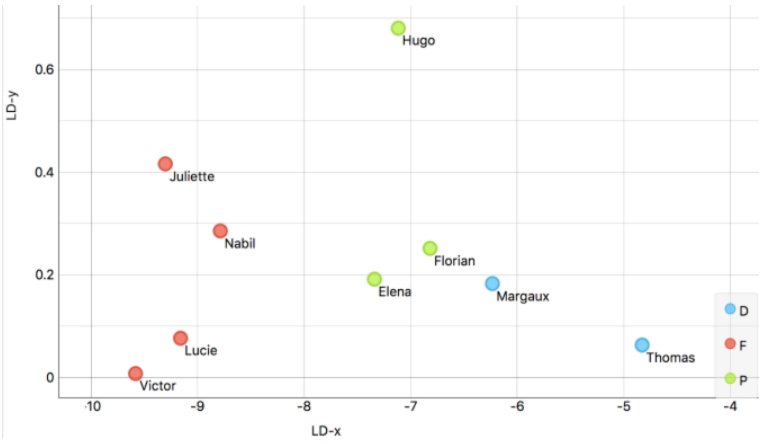
1-1) Combien y a-t-il de facteurs discriminants ?

1-2) Les vecteurs directeurs des 2 premiers axes sont donnés ci-dessous :

	component	math.	info.	fran.	angl.	arts
1	LD-x	0.4	-0.9	-0.1	0.1	-0.2
2	LD-y	-0.0	0.2	0.6	-0.8	0.1

0.3961	-0.8825	-0.1373	0.0710	-0.2009
-0.0356	0.2202	0.5580	-0.7964	0.0682

La projection des élèves dans le plan qu'ils définissent est visualisée ci-contre. Comment ont été calculées les coordonnées de ce nuage de points ?



1-3) Calculez les coordonnées d'un des élèves.

1-4) Soit un élève ayant obtenu 10 dans les cinq matières dont on souhaite prédire l'avis à partir de ses notes. Calculez ses coordonnées dans le plan discriminant.

1-5) Sans calcul, où se situeraient un élève ayant obtenu 0 partout ?

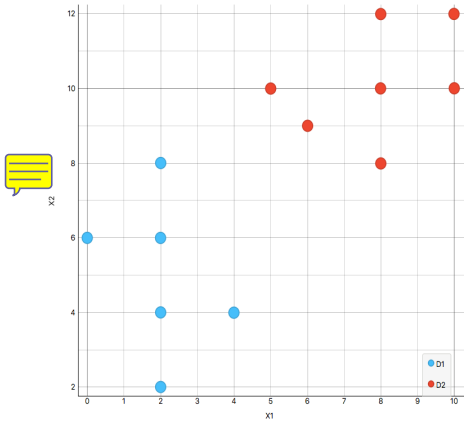
1-6) Et 20 partout ?

2. Approche par Voisins

La méthode (ou règle) dite du *Plus Proche Voisin* (*Nearest Neighbor*) est très populaire car très simple à mette en œuvre. Elle consiste, pour prédire le groupe d'un individu inconnu x, à chercher le point du tableau de données dont il est le plus *proche* et choisir son groupe. L'utilisateur doit bien sûr choisir une *distance*.
La méthode des K-*Plus Proches Voisins* (K-*NN*) consiste à prédire pour (ou classer) x (dans) le groupe majoritairement représenté parmi ses K plus proches voisins dans le tableau de données. Le nombre K est un paramètre choisi par l'utilisateur.

Exercice 2

Considérons le nuage de points ci-contre, et la distance de Manhattan.



2-1) Classez les points $x = (4, 8)$, $y = (6, 4)$ et $z = (7, 6)$ par la règle du 1-PPV.

2-2) Trouvez un point pour lequel on ne saura pas prédire le groupe. Que faire en pareil cas ?

2-3) Pouvez-vous extrapoler la frontière de décision entre les deux groupes ?

2-4) Reclassez les trois points par la règle des 3-PPV.

(HW) Chez vous, vous reprendrez les questions précédentes avec la distance euclidienne usuelle.

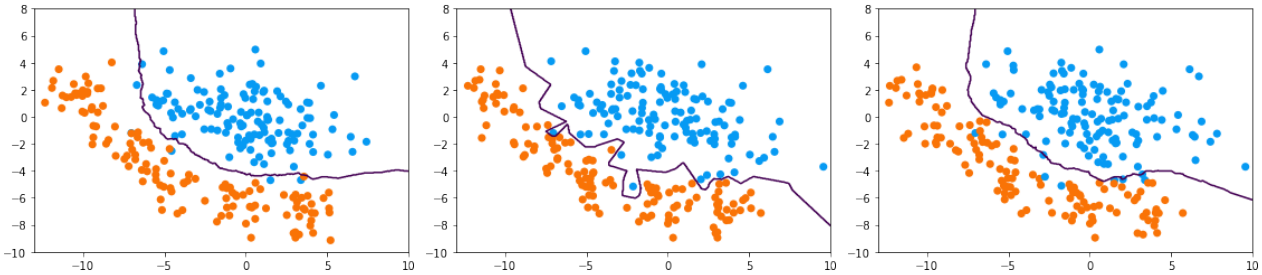
2-5) Pourquoi ne serait-il pas judicieux d'utiliser ici la règle des 2-PPV ou celle des 4-PPV ?

(a) Dans certains ouvrages/cours, on lit que pour éviter toute ambiguïté de classement, il faut choisir K impair. Qu'en pensez-vous ?

(b) En pratique, l'avantage principal de la méthode des K-PPV est sa simplicité. Quel est son énorme inconvénient ? Pensez algorithmique...

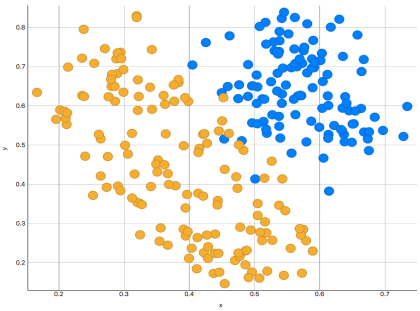
(c) Les trois figures ci-dessous représentent des tableaux de données générés aléatoirement (donc presque similaires) et la frontière donnée par les K-PPV pour différentes valeurs de K : 1,10 et 50.

- Remettez-les dans l'ordre.
- Laquelle des trois règles vous semble la meilleure ?



3. Approche Fonctionnelle

Une autre approche consiste à apprendre les paramètres d'une fonction f qui sépare le mieux possible les groupes. Considérons le nuage de points ci-contre.



(a) Comment situeriez-vous la moins complexe possible des fonctions $f(x, y)$ sépare les deux groupes ?

(b) Comment l'utiliser pour prédire ?

(c) La fonction $f(x, y)$ d'ordre immédiatement supérieur ?

(d) Une qui sépare parfaitement les deux groupes ?

(e) Selon vous, laquelle des trois présente le meilleur *pouvoir prédictif* pour des inconnus ?

4. Approche Probabiliste

Une autre approche consiste à utiliser une modélisation probabiliste des groupes. Les paramètres des lois sont appris sur le tableau de données, et il s'agit, pour un inconnu x de calculer les probabilités $P(i|x)$ $(i = 1, m)$ et sélectionner le groupe le plus probable. Deux exemples de frontière de prédiction sont données ci-contre pour un modèle normal. Commentez-les.