Licence d'Informatique 2ème année
Analyse de Données Utilisateur (C5–160412)
Contrôle du 14 mai 2019 (2h30)

Carl FRÉLICOT – Dpt Info / Lab MIA

- Autorisés : calculatrice (pas une application sur smartphone), l'aide-mémoire déposé sur l'ENT sans **aucune annotation**. Tout **autre document est interdit**
- Utilisez les cadres réservés pour inscrire vos réponses.
  Vous devez **écrire la formule** utilisée avant de donner le détail du moindre calcul.
- Une réponse **non justifiée** sera considérée comme **fausse**.

Nom, prénom : Note : /20

Exercice 1

On a testé deux logiciels (`Blue` et `Yellow`) de *Machine Learning* et leur association `Green` sur 10 apprenants à qui on a demandé de graduer sur une échelle de 0 à 20 leur ressenti en termes d'effet potentiel sur leur compréhension du *Machine Learning* : sans effet (`sans`), amélioration (`amel.`) et amélioration significative (`amel.+`). Les résultats sont donnés dans le tableau ci-contre, ainsi que leurs projections dans le 1er plan discriminant sauf une qui a été effacée.

	logiciel	sans	amel.	amel.+	LD-x	LD-y
1	Blue	2.0	4.0	10.0	10.425	2.134
2	Blue	2.0	16.0	2.0	10.156	-4.933
3	Yellow	11.0	7.0	5.0	7.246	7.539
4	Green	16.0	11.0	19.0	20.836	13.149
5	Yellow	19.0	4.0	4.0	4.267	15.608
6	Green	3.0	13.0	17.0	21.084	0.527
7	Yellow	12.0	6.0	11.0	11.685	10.155
8	Green	10.0	14.0	20.0	23.684	6.819
9	Blue	1.0	14.0	2.0	9.146	-4.914
10	Blue	7.0	15.0	9.0		1.378
	$\overline{x}$	8.3	10.4	9.9	13.37	4.75
	s	6.0	4.45	6.49	6.20	6.73

Des statistiques ont par ailleurs été calculées, et on donne :

	component	sans	amel.	amel.+
1	LD-x	-0.0656	0.5379	0.8405
2	LD-y	0.8693	-0.4441	0.2172

1. Combien y a-t-il au plus d'axes discriminants pour ces données ?

2. Calculez la coordonnée effacée.

3. Pour la note `amel.+`, on a relevé les statistiques conditionnelles ci-contre. Calculez la statistique de Fisher permettant de savoir si elle sépare bien les logiciels.

logiciel	`Blue`	`Yellow`	`Green`
$\overline{x}_j$	5.75	6.67	18.67
s_j	3.77	3.09	1.25

4. Faites les calculs permettant de tracer le boxplot de la note `amel.+`, puis tracez-le (verticalement, à droite).

5. Quelle erreur de saisie minimale aurait-il fallu faire sur la note 20 (note entière) pour que celle-ci soit en dehors de l'épure ?

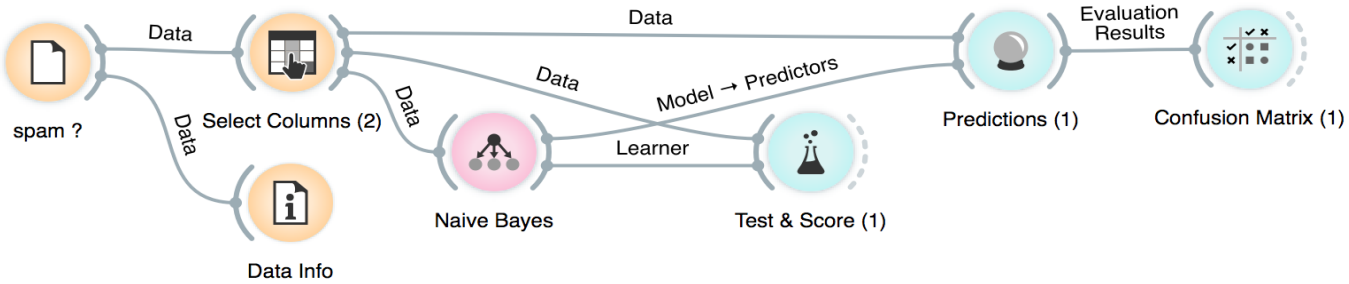
6. Afin d'expliquer (ou non) `LD-y` par `amel` avec un modèle affine, on a calculé en plus de ce qui figure en page 1 : $\sum x \times y = 304.6$, $\sum y^2 = 678.2$ et $\sum x^2 = 1\,280$. Faites les calculs nécessaires et donnez le modèle de prévision.

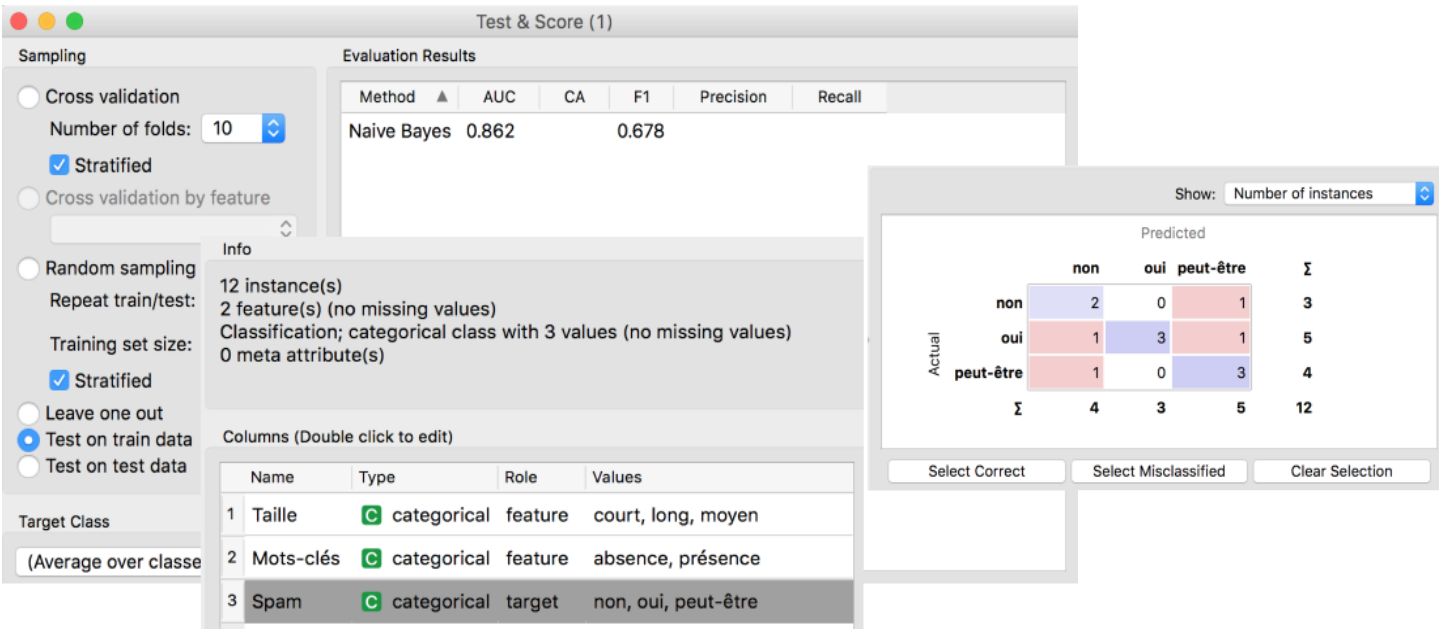
7. Calculez la distance[1] entre `amel` et `amel+`.

[1] faites un choix adapté

Exercice 2

Ci-dessous un canvas Orange et quelques sorties où trois valeurs d'évaluation ont été effacées.





1. Les données sont des courriels à filtrer. Expliquez très brièvement la raison d'être du canvas.

2. Calculez les mesures de rappel et précision de la modalité `peut-être`.

3. On donne : $R_{non} = 66.7\%$, $R_{oui} = 60\%$, $P_{non} = 50\%$ et $P_{oui} = 100\%$. Calculez les trois valeurs manquantes.

Exercice 3

Considérons le tableau de données en dimension 3 de l'Exercice 1, mais limité aux 5 derniers individus (time is money) renommés x, y, z, t et u. Appelons-le T et $\overline{T}$ celui composé des 5 premiers individus.

1. Calculez les distances de Manhattan, puis de Chebychev entre les deux premiers individus de T.

2. Comme il y a 3 modalités d'utilisation des logiciels, on a exécuté trois itérations des `K-means` sur T, avec $K = 3$ et la distance euclidienne usuelle. Faites les calculs permettant de compléter les tableaux ci-dessous.

$Y^{(0)} = [1, 2, 3, 1, 2] \rightarrow V^{(1)} = \begin{bmatrix} \overline{x}_1 = (\quad , \quad , \quad) \\ \overline{x}_2 = (9.5, 10.5, 10), \\ \overline{x}_3 = (10, 14, 20) \end{bmatrix} \rightarrow$

$d^2(\overline{x}_j, x_i)$	x	y	z	t	u
$\overline{x}_1$	57.5	158.5	174.5	57.5	27.5
$\overline{x}_2$	97.5	27.5	112.5	148.5	
$\overline{x}_3$	59	149		405	131
$Y^{(1)}$	1	2		1	1

$\rightarrow V^{(2)} = \begin{bmatrix} \overline{x}_1 = (\quad , \quad , \quad), \\ \overline{x}_2 = (12, 6, 11), \\ \overline{x}_3 = (\quad , \quad , \quad) \end{bmatrix} \rightarrow$

$d^2(\overline{x}_j, x_i)$	x	y	z	t	u
$\overline{x}_1$	60.22	136.22	153.89	60.89	12.22
$\overline{x}_2$	166		149	266	110
$\overline{x}_3$	59	149		405	131
$Y^{(2)}$	3			1	1

$\rightarrow V^{(3)} = \begin{bmatrix} \overline{x}_1 = (4, 14.5, 5.5), \\ \overline{x}_2 = (\quad , \quad , \quad), \\ \overline{x}_3 = (6.5, 13.5, 18.5) \end{bmatrix} \rightarrow$

$d^2(\overline{x}_j, x_i)$	x	y	z	t	u
$\overline{x}_1$	135.5	166.5	246.5	21.5	
$\overline{x}_2$	166	0	149	266	110
$\overline{x}_3$	14.75	142.75	14.75	302.75	92.75
$Y^{(3)}$					

itération 1 :

itération 2 :

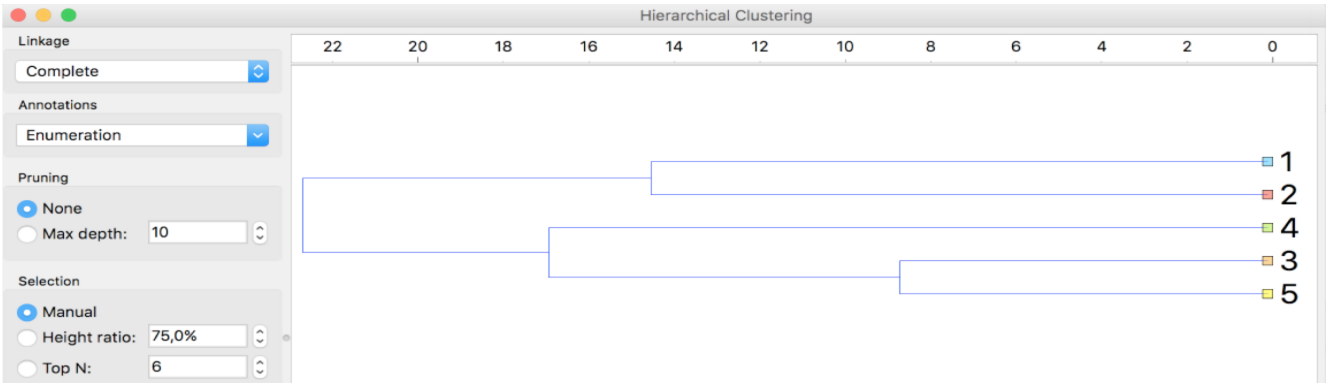
itération 3 :

(...)

Nom, prénom :

3. Était-il judicieux d'itérer davantage ?

4. En supposant $\overline{T}$ contenue dans un `Data Table`, "dessinez" le canevas Orange qui a permis d'obtenir la figure ci-dessous, ainsi que la visualisation du tableau de distances.



5. Quelle partition retiendriez-vous de cette hiérarchie?

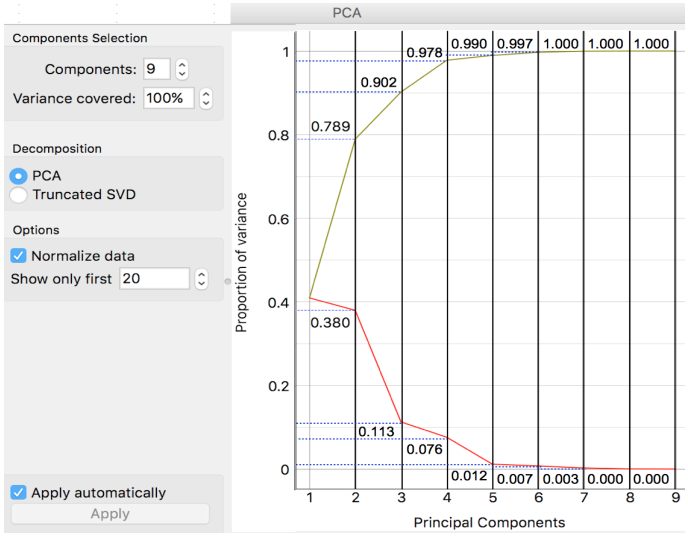
6. Analysez, aux fins d'interprétation, ce qu'on peut finalement retenir de ces deux études (sur T et $\overline{T}$) ?

Exercice 4

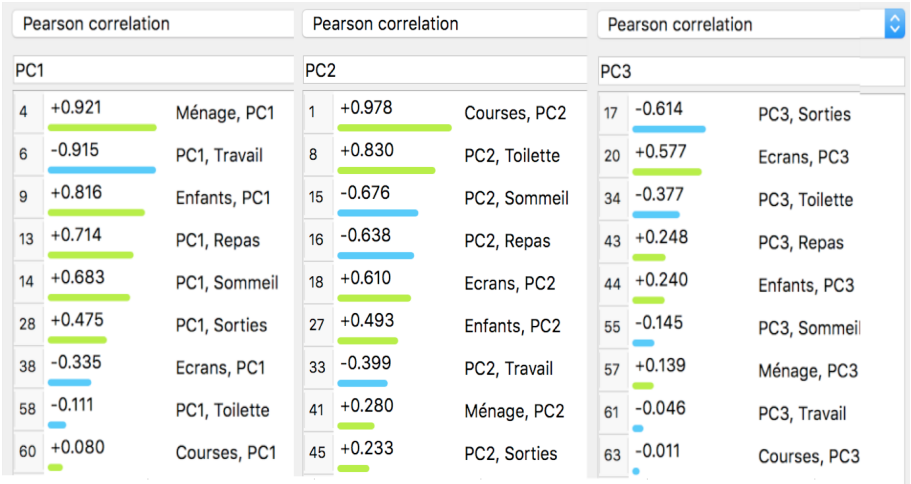
Dans le tableau de données qui suit sont reportés les temps quotidien moyen (en heures) passé par des individus typiques à diverses activités. Une Analyse en Composantes Principales aété réalisée (sorties ci-après).

Les méta-données sont illustratives et servent notamment à coder les individus : FNMu est une Femme Non active Mariée américaine alors que HACe est un Homme Actif Célibataire européen.

	Gen.	Actif	Civil.	Géo.	Code	Travail	Ménage	Enfants	Courses	Toilette	Repas	Sommeil	Ecrans	Sorties
1	F	oui	célibat.	US	FACu	5.80	2.50	0.30	1.40	1.20	1.00	7.60	1.15	3.05
2	F	oui	célibat.	US	FACu	5.80	1.96	0.18	1.41	1.26	0.96	7.75	1.32	3.36
3	F	oui	célibat.	EUR	FACe	5.81	3.07	0.30	0.80	0.95	1.42	8.16	0.87	2.62
4	F	non	célibat.	EUR	FNCe	4.63	2.62	0.14	0.92	0.97	1.47	8.49	0.84	3.92
5	F	non	marié/e	US	FNMu	0.10	4.95	1.10	1.70	1.10	1.30	7.85	1.60	4.30
6	F	non	marié/e	US	FNMu	2.00	4.26	0.90	1.61	1.12	1.19	7.76	1.43	3.73
7	F	non	marié/e	EUR	FNMe	0.27	5.68	0.87	1.12	0.90	1.80	8.43	1.25	3.68
8	F	non	marié/e	EUR	FNMe	1.90	5.28	0.69	1.02	0.83	1.74	8.24	1.19	3.11
9	H	oui	célibat.	US	HACu	7.00	0.50	0.00	1.50	1.05	1.00	7.60	1.50	3.85
10	H	oui	célibat.	EUR	HACe	7.53	0.95	0.07	0.57	0.85	1.50	8.08	1.15	3.30
11	H	oui	célibat.	EUR	HACe	7.48	0.72	0.00	0.62	0.77	1.40	8.13	1.00	3.88
12	H	oui	marié/e	US	FAMu	7.50	0.60	0.10	1.20	0.95	1.15	7.60	1.75	3.15
13	H	oui	marié/e	US	HAMu	8.50	0.65	0.10	1.15	0.90	1.15	7.65	1.80	2.10
14	H	oui	marié/e	EUR	HAMe	7.53	0.97	0.10	0.52	0.85	1.52	8.08	1.22	3.21



Code	PC1	PC2	PC3
FACu	-1.182	1.671	-0.700
FACu	-1.333	1.879	-1.090
FACe	0.253	-1.592	0.069
FNCe	1.216	-1.630	-1.707
FNMu	2.466	3.108	0.054
FNMu	1.290	2.624	0.013
FNMe	3.725	-0.392	0.684
FNMe	2.617	-0.934	1.237
HACu	-2.050	1.441	-0.854
HACe	-0.702	-2.246	-0.073
HACe	-0.533	-2.463	-0.983
FAMu	-2.232	0.670	0.897
HAMu	-2.815	0.124	2.248
HAMe	-0.721	-2.260	0.207



1. Quelle autre(s) méthode(s) que l'ACP aurai(en)t certainement été intéressante(s) à réaliser sur de telles données ?

2. Vous devez :
   - commenter (critiquer ?) les choix de l'analyste
   - donner, quels que soient ces choix, des éléments d'interprétation des trois premières composantes principales à partir de l'analyse des individus et des variables

(...)