Devoir Surveillé **Correction**

Semestre : 1

Module: **Intelligence Artificielle et Systèmes Experts – Application Data Mining**
Enseignants: **Mohamed Heny SELMI, Walid AYEDI**
Classes: **4 INFO A1, 2, 3 et 4**
Documents autorisés : **NON**
Nombre de pages : **4**
Date : **Mercredi 07/11/2012**
Heure : **14h15mn**
Durée : **1h30min**

Exercice 1 :

On propose d'analyser un échantillon de 37 véhicules caractérisés par 11 variables quantitatives via une Analyse en Composantes Principales.

Liste des variables actives :

PUIS puissance, CYLI cylindrée, VITE vitesse, LONG longueur, LARG largeur, HAUT hauteur, POID poids, COFF coffre, RESE réservoir, CONS consommation, CO2 emission_CO2 C

Liste des variables illustratives : PRIX prix

Tab1. Valeurs propres

HISTOGRAMME DES 11 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	7.4633	67.85	67.85	**********
2	1.4790	13.45	81.29	**********
3	0.9396	8.54	89.84	**********
4	0.3534	3.21	93.05	****
5	0.2764	2.51	95.56	***
6	0.1995	1.81	97.38	***
7	0.1255	1.14	98.52	**
8	0.0892	0.81	99.33	*
9	0.0554	0.50	99.83	*
10	0.0148	0.13	99.97	*
11	0.0038	0.03	100.00	*

Tab2. Coordonnées des variables sur les axes 1 à 5

VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR				
IDEN - LIBELLE COURT	1	2	3	4	5	1	2	3	4	5
PUIS - puissance	0.93	-0.26	0.16	0.08	0.07	0.93	-0.26	0.16	0.08	0.07
CYLI - cylindrée	0.86	-0.04	0.13	0.38	0.28	0.86	-0.04	0.13	0.38	0.28
VITE - vitesse	0.85	-0.37	-0.19	0.04	0.02	0.85	-0.37	-0.19	0.04	0.02
LONG - longueur	0.93	0.07	-0.27	-0.04	-0.05	0.93	0.07	-0.27	-0.04	-0.05
LARG - large	0.88	0.20	-0.24	-0.01	-0.15	0.88	0.20	-0.24	-0.01	-0.15
HAUT - hauteur	0.09	0.81	0.56	0.00	0.00	0.09	0.81	0.56	0.00	0.00
POID - poids	0.91	0.30	0.08	0.09	-0.13	0.91	0.30	0.08	0.09	-0.13
COFF - coffre	0.65	0.51	-0.39	-0.26	0.30	0.65	0.51	-0.39	-0.26	0.30
RESE - réservoir	0.90	0.21	-0.13	0.11	-0.25	0.90	0.21	-0.13	0.11	-0.25
CONS - consommation	0.83	-0.34	0.36	-0.25	0.01	0.83	-0.34	0.36	-0.25	0.01
CO2 - emission_CO2	0.88	-0.26	0.33	-0.22	-0.01	0.88	-0.26	0.33	-0.22	-0.01

Tab3. Coordonnées, contributions et cosinus carrés des individus tab3

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
ALFA 156	2.70	14.27	2.57	-2.38	0.84	-0.46	0.83	2.4	10.4	2.1	1.6	6.8	0.46	0.40	0.05	0.01	0.05
AUDIA3	2.70	2.83	-1.36	-0.30	-0.68	-0.32	-0.15	0.7	0.2	1.3	0.8	0.2	0.66	0.03	0.16	0.04	0.01
AUDIA8	2.70	33.41	5.56	-0.57	-0.53	0.10	-0.51	11.2	0.6	0.8	0.1	2.5	0.92	0.01	0.01	0.00	0.01
AVENSIS	2.70	3.05	-0.39	0.81	-1.33	-0.08	0.39	0.1	1.2	5.1	0.0	1.5	0.05	0.22	0.58	0.00	0.05
BMW X5	2.70	20.56	3.49	2.30	0.97	0.74	-1.04	4.4	9.7	2.7	4.2	10.6	0.59	0.26	0.05	0.03	0.05
BMW530	2.70	9.89	2.91	-0.55	-0.67	-0.16	0.56	3.1	0.6	1.3	0.2	3.1	0.86	0.03	0.04	0.00	0.03
CHRYS300	2.70	40.15	5.66	-1.36	1.03	1.46	1.59	11.6	3.4	3.0	16.2	24.9	0.80	0.05	0.03	0.05	0.06
CITRONC2	2.70	16.26	-3.97	-0.42	0.25	-0.23	0.11	5.7	0.3	0.2	0.4	0.1	0.97	0.01	0.00	0.00	0.00
CITRONC4	2.70	2.45	-0.85	-0.03	-0.88	0.82	-0.31	0.3	0.0	2.2	5.2	0.9	0.30	0.00	0.32	0.28	0.04
CITRONC5	2.70	5.55	2.12	-0.56	-0.01	-0.56	0.34	1.6	0.6	0.0	2.4	1.1	0.81	0.06	0.00	0.06	0.02
CLIO	2.70	13.47	-3.42	-0.61	-0.72	0.67	0.22	4.2	0.7	1.5	3.5	0.5	0.87	0.03	0.04	0.03	0.00
CORSA	2.70	14.13	-3.70	-0.27	-0.41	0.18	0.12	5.0	0.1	0.5	0.3	0.2	0.97	0.01	0.01	0.00	0.00
FIESTA	2.70	12.96	-3.48	-0.03	-0.61	0.44	0.03	4.4	0.0	1.1	1.5	0.0	0.94	0.00	0.03	0.02	0.00
GOLF	2.70	5.52	-1.96	0.68	-0.60	0.32	0.02	1.4	0.8	1.0	0.8	0.0	0.70	0.08	0.06	0.02	0.00
LAGUNA	2.70	2.05	0.70	-0.52	-0.89	-0.31	-0.15	0.2	0.5	2.3	0.7	0.2	0.24	0.13	0.38	0.05	0.01
MAZDARX8	2.70	12.53	1.37	-2.75	0.29	-1.46	-0.71	0.7	13.8	0.2	16.3	4.9	0.15	0.60	0.01	0.17	0.04
MEGANECC	2.70	3.32	0.00	-1.48	-0.16	0.49	-0.80	0.0	4.0	0.1	1.8	6.3	0.00	0.66	0.01	0.07	0.19
MERC_A	2.70	4.59	-1.21	1.14	-0.05	0.38	0.55	0.5	2.4	0.0	1.1	3.0	0.32	0.28	0.00	0.03	0.07
MERC_E	2.70	10.28	2.56	0.33	-1.41	1.05	0.18	2.4	0.2	5.7	8.4	0.3	0.64	0.01	0.19	0.11	0.00
MODUS	2.70	6.94	-2.28	-0.01	1.12	0.23	-0.19	1.9	0.0	3.6	0.4	0.4	0.75	0.00	0.18	0.01	0.01
MONDEO	2.70	6.00	1.10	0.07	-1.62	-0.66	-0.09	0.4	0.0	7.5	3.3	0.1	0.20	0.00	0.44	0.07	0.00
MURANO	2.70	23.49	4.06	1.14	2.20	-0.25	-0.30	6.0	2.4	14.0	0.5	0.9	0.70	0.06	0.21	0.00	0.00
MUSA	2.70	8.65	-2.16	1.45	1.02	0.37	0.41	1.7	3.8	3.0	1.0	1.6	0.54	0.24	0.12	0.02	0.02
OUTLAND	2.70	5.99	1.48	0.54	1.42	-0.78	-0.05	0.8	0.5	5.8	4.6	0.0	0.36	0.05	0.34	0.10	0.00
P1007	2.70	11.05	-2.98	0.51	1.23	0.17	-0.52	3.2	0.5	4.4	0.2	2.6	0.80	0.02	0.14	0.00	0.02
P307CC	2.70	3.89	0.09	-1.64	0.39	0.09	-0.37	0.0	4.9	0.4	0.1	1.3	0.00	0.69	0.04	0.00	0.04
P407	2.70	1.68	0.68	-0.19	-0.85	-0.32	-0.39	0.2	0.1	2.1	0.8	1.5	0.28	0.02	0.43	0.06	0.09
P607	2.70	8.57	2.70	-0.07	-0.89	0.24	-0.40	2.6	0.0	2.3	0.5	1.5	0.85	0.00	0.09	0.01	0.02
PANDA	2.70	23.76	-4.74	-0.14	0.97	-0.23	0.51	8.1	0.0	2.7	0.4	2.6	0.94	0.00	0.04	0.00	0.01
PASSAT	2.70	2.09	0.52	-0.20	-0.78	-0.74	0.19	0.1	0.1	1.7	4.2	0.4	0.13	0.02	0.29	0.26	0.02
PTCRUISER	2.70	5.23	0.62	-0.96	1.81	0.20	-0.18	0.1	1.7	9.5	0.3	0.3	0.07	0.17	0.63	0.01	0.01
SANTA_FE	2.70	19.48	1.29	3.76	-0.19	-1.51	1.01	0.6	25.8	0.1	17.5	9.9	0.08	0.72	0.00	0.12	0.05
TWINGO	2.70	22.31	-4.53	-0.97	0.46	-0.10	0.13	7.4	1.7	0.6	0.1	0.2	0.92	0.04	0.01	0.00	0.00
VECTRA	2.70	2.85	0.17	0.40	-1.53	-0.02	0.22	0.0	0.3	6.7	0.0	0.5	0.01	0.06	0.82	0.00	0.02
VELSATIS	2.70	6.59	1.62	1.59	-0.60	0.21	-0.94	0.9	4.6	1.0	0.3	8.6	0.40	0.38	0.06	0.01	0.13
X-TRAIL	2.70	3.74	-0.13	1.50	1.04	0.14	-0.21	0.0	4.1	3.1	0.1	0.4	0.00	0.60	0.29	0.01	0.01
YARIS	2.70	17.44	-4.11	-0.19	0.35	-0.13	-0.13	6.1	0.1	0.4	0.1	0.2	0.97	0.00	0.01	0.00	0.00

Fig1. Cercle de corrélation

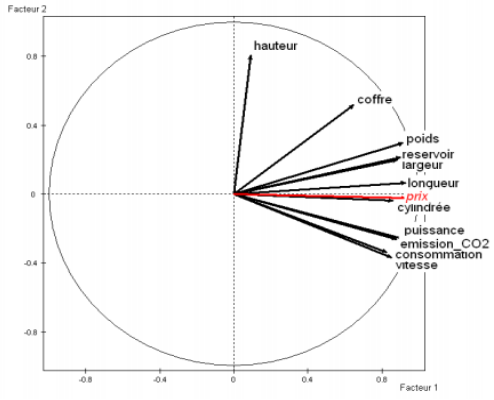
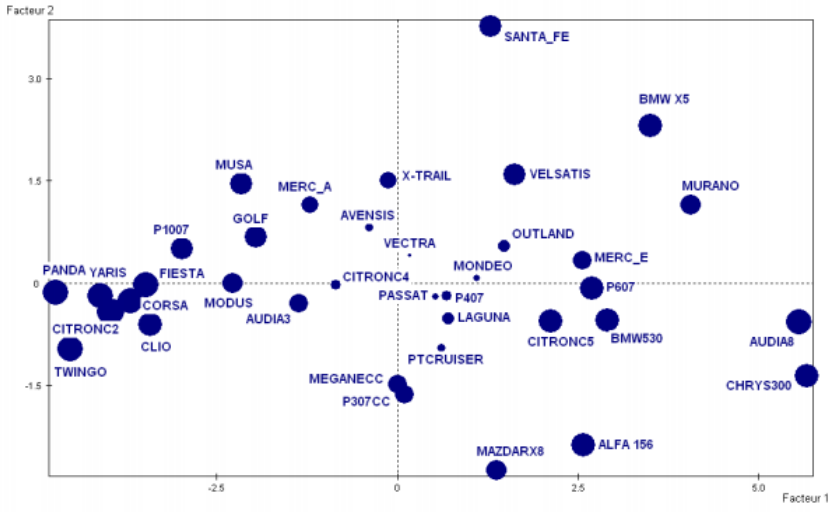


Fig2. Plan factoriel (CP1, CP2)



1. Expliquer les objectifs de l'Analyse en Composantes Principales (ACP) en Informatique décisionnelle.

L'ACP permet une analyse multi variée d'un échantillon décrit par des variables quantitatives. Elle permet d'étudier, d'une part, les corrélations entre les variables et d'autre part, d'analyser la distribution du nuage de l'échantillon sur les plans factoriels. [2]

2. Etude du tableau des valeurs propres

I. A quoi correspond la somme des valeurs propres ?

La somme des valeurs propres quantifie l'information ou l'inertie totale contenue dans l'échantillon étudié. [1]

II. On choisit de n'étudier que les deux premières composantes principales. Justifier ce choix en analysant le tableau des valeurs propres (tab1).

Les deux premières composantes correspondent aux deux valeurs propres supérieures à 1 et quantifient plus que 80% de l'inertie totale de l'échantillon étudié. [1]

3. Interpréter les deux premières composantes principales en analysant les contributions et les corrélations des variables par rapport à chaque composante en se basant sur la table2

La première composante principale est corrélée avec les variables qui caractérisent la puissance, la vitesse, la longueur, la largeur, le réservoir, la consommation et l'émission de CO2 des voitures. La deuxième composante est corrélée significativement avec seulement la variable caractérisant la hauteur des voitures. [2]

4. La représentation graphique des individus montre que la majorité des véhicules sont bien représentées dans le plan (CP1, CP2). Justifier cette affirmation en se basant sur la table3

La qualité des représentations des individus est quantifiée via le cosinus de l'angle que fait véhicules avec chaque composante du plan factoriel dans l'espace des véhicules. Plus le cosinus carré est proche de 1, plus la véhicule est bien représentée ou bien projetée sur le plan factoriel. En examinant les cosinus carrés des véhicules par rapport aux deux composantes de la table 3, on remarque que la majorité des véhicules ont des cosinus carrés supérieurs à 0.5. Ceci confirme la bonne qualité des représentations des véhicules dans le plan (CP1, CP2). [2]

5. Commenter les positions des véhicules suivantes par rapport aux variables sur le plan factoriel :

i. AUDIA8, CHRYS300 : **ces véhicules sont corrélés positivement avec CP1. Ainsi, ces derniers sont puissants, rapides et par conséquent, consomment beaucoup de carburant et émettent beaucoup de CO2. Ils sont caractérisés aussi par des dimensions importantes en longueurs et en largeurs. [1]**

ii. PANDA, YARIS : **ces véhicules sont corrélés négativement avec CP1. Ainsi, ces derniers ne sont ni puissants ni rapides et par conséquent, consomment peu de carburant et émettent peu de CO2. Ils sont caractérisés aussi par des dimensions faibles en longueurs et en largeurs. [1]**

iii. SANTA_FE, BMW X5 : **ces véhicules sont corrélés positivement avec CP2. Ces derniers sont caractérisés par des dimensions en hauteur importantes relativement aux autres véhicules. [1]**

iv. MAZDARX8, ALFA 156 : **ces véhicules sont corrélés négativement avec CP2. Ces derniers sont caractérisés par des dimensions en hauteur faibles relativement aux autres véhicules. [1]**

Exercice 2 :

On souhaite découper la clientèle d'une banque en groupes homogènes. On propose d'appliquer le découpage via la méthode K-means.

1. Expliquer l'utilité d'un tel découpage dans le domaine de la banque, en se basant sur un exemple.

Avec un tel découpage, le banquier peut caractériser chaque groupe par les produits ou les services de la banque préférés par les clients de ce cluster. Ainsi, si le banquier souhaite faire une action marketing pour un produit ou un service précis, il applique cette action aux clients appartenant au groupe adéquat au lieu de l'appliquer à tous les clients de la banque. [2]

2. Parmi les critères d'arrêt de l'algorithme K-means, on cite la stabilisation de l'inertie totale des groupes résultants. Expliquer comment cette stabilisation de l'inertie permet de générer des groupes homogènes.

On a V=A+E avec V la matrice de variance-covariance ou l'inertie totale des individus, A la matrice intra-classes et E la matrice inter-classes. A chaque itération de l'algorithme, A diminue et E augmente. Dans le cas où A atteint son seuil minimal et E atteint son seuil maximal, la somme de A et de E se stabilisent et on converge vers des groupes homogènes et stables. [2]

3. Citer deux inconvénients de la méthode K-means.

Deux inconvénients de K-means :

1. **Le choix subjectif du nombre de groupes à construire au début de l'algorithme (le choix de K). [1]**
2. **Le choix aléatoire des K centres au début de l'algorithme. [1]**

4. Un expert de la banque propose de construire deux groupes de clients : clients à tendance épargne et clients à tendance non épargne. Afin de quantifier la qualité du découpage, on propose de croiser le résultat de l'algorithme K-means sur un échantillon de clients caractérisé par une variable qui identifie les clients réellement à tendance épargne par le caractère E et les autres clients par le caractère NE.

Le résultat du croisement génère la table de confusion suivante :

	1	2
E	4	54
NE	72	7

En se basant sur les résultats de la table de confusion, quantifier la qualité du découpage obtenu.

D'après la table de confusion, les clients à tendance épargne sont représentés par le groupe 2 et les autres clients par le groupe 1. Ainsi, le taux de bonne classification totale est égale à ((72+54)/(72+54+4+7))* 100 =91,97%, le taux de bonne classification des clients à tendance épargne est égale à (54/(54+4))* 100 =93,1% et le taux de bonne classification des clients à tendance non épargne est égale à (72/(72+7))* 100 =91.13%. [2]