Licence d'Informatique 2ème année
AD Utilisateur (C5-160412-INFO)
1ère chance - 26 mars 2021 (1h45)

Carl FRÉLICOT – Dpt Info / Lab MIA

- Autorisés : calculatrice (pas une application sur smartphone), un formulaire **manuscrit** sans **aucun exemple numérique**. 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 (en **capitales**), Prénom : Note : /20

On s'intéresse aux données `Wine.tab` de Orange étudiées en TP consistant à l'observation de 13 *features* chimiques sur 178 vins italiens de qualités différentes donnée par une *categorical* `Wine` de modalités : 1, 2, 3 et d'effectifs : 59, 71, 48.

Exercice 1

Décrivez ce que pourrait être selon vous l'objectif du *canvas* donné.

L'objectif du *canvas* est la réduction de la dimension : $(178 \times 13) \rightarrow (9 \times 5)$ à partir d'un échantillon aléatoire ($n = 178 \times 5\% \approx 9$) stratifié : [59, 71, 48] $\times 5\% \approx$ [3,4,2].

Les $p = 5$ variables les plus discriminantes au sens de la statistique de Fisher sont sélectionnées.

Exercice 2

Pour des questions de temps d'épreuve, nous avons extrait le sous-tableau ci-contre composé de $n = 9$ vins et $p = 5$ variables (que vous pouvez écrire H, P, C, A, OD), en dessous duquel des statistiques ont été calculées.

	Wine	Hue	Proline	Color	Alcohol	OD
1	2	0.93	465	6	11.56	3.69
2	2	1	680	2.8	11.64	2.75
3	1	1.02	1290	5.25	14.39	3.58
4	2	0.96	495	3.05	12.69	2.06
5	2	1.23	428	3.8	11.66	2.14
6	1	1.18	1020	3.7	13.07	2.69
7	1	1.03	770	4.5	12.93	3.52
8	3	0.61	425	7.1	13.11	1.33
9	3	0.61	560	9.2	14.13	1.6
$\bar{x}$		0.952	681.4	5.044	12.8	2.6
s		0.205	282.7	1.975	0.982	0.826

1) Donnez des éléments de calcul permettant d'obtenir le *boxplot* de la variable `Proline` que vous dessinerez verticalement.
   ATTENTION : c'est le *boxplot* tel qu'il doit être calculé qui est demandé, pas celui retourné par Orange.

deux solutions acceptées

1) $Q_2 = 560$, $Q_1 = \frac{428+465}{2} = 446.5$, $Q_3 = \frac{770+1020}{2} = 895$ **et** $EIQ = 448.5$ **;**
moustaches et 425 et 1290, pas d'outlier

2) $Q_2 = 560$, $Q_1 = 465$, $Q_3 = 770$ **et** $EIQ = 305$ **;**
moustaches 425 et 1020 et outlier en 1290

2) Utilisez les valeurs que la sortie Orange affiche afin de calculer la valeur manquante de sortie du *widget* `Rank` pour la `Proline` du *canvas* de l'Exercice 1.

$s_B^2 = \frac{3}{9}(1026.67 - 681.44)^2 + \frac{4}{9}(517 - 681.44)^2 + \frac{2}{9}(492.5 - 681.44)^2 = 59\,678.88$ **et**
$s_W^2 = \frac{3}{9}212.34^2 + \frac{4}{9}97.05^2 + \frac{2}{9}67.5^2 = 20\,228.01$ **;**

alors $F = \frac{59\,678.88/2}{20\,228.01/6} = 8.85$

3) Afin d'expliquer les valeurs de `Color` (C) par celles de `Hue` (H), on donne : $\sum_{k=1}^{9} H_k \times C_k = 40.28$. Calculez le coefficient de corrélation entre les deux variables. Commentez le résultat en termes de possibilité d'ajustement par un premier modèle (1), et lequel ?

$s_{CH}^2 = \frac{40.28}{9} - 5.044 \times 0.952 = -0.369$ **et** $r_{CH} = \frac{-0.369}{0.205 \times 1.975} = -0.811$ **;**

C et H sont fortement anti-corrélées et on doit pouvoir ajuster par un modèle affine décroissant

4) Deux autres modèles (2) et (3) sont proposés. Lequel retiendriez-vous ? Vous écrirez l'expression mathématique du modèle avec des coefficients fictifs : a, b, c, etc.

au sens de l'écart-type résiduel le modèle (3) est très légèrement meilleur que le (2) :
$1.0243 < 1.0254$ **;**

mais le principe de parcimonie impose de retenir le modèle (2) plus simple dont l'expression est : $C = a \times H^2 + b \times H + c$

Exercice 3

1) Une ACP est réalisée sur le sous-tableau (Exercice 2). Auriez-vous fait les mêmes choix que l'analyste ?

L'analyste a eu raison de choisir une ACP normée : $s_P^2 = 282.7^2 >> s_H^2 = 0.205^2$.

Le % cumulé de variance expliquée suggère de retenir $q = 3$ ou 4 composantes (coude) et non 5.

Donnez des éléments d'interprétation (individus, variables) du 1er plan factoriel.

PC1 : axe colorimétrique, oppose les vins (1) et (2) à forte H et OD et faible C à gauche aux vins (3) à forte C et faibles H et OD à droite

PC2 : axe (de force?) opposant les vins (1) à forts A et P à droite aux vins (2) à gauche à faibles A et P

2) Une AFD est réalisée sur le sous-tableau (Exercice 2). Donnez des éléments d'interprétation (individus, variables) du 1er plan discriminant.

On retrouve des éléments similaires à ceux de l'ACP.

LDx : ordonne les vins par qualité : (3) à gauche ($C+$, OD- et H-), (2) au centre et (1) à droite (C-, $OD+$ et $H+$)

LDy : oppose les vins (1) en haut ($A+$) aux vins (2) (A-) en bas

3) Soit un nouveau vin $x = (1,600, 5, 11, 3)$. Projetez-le dans le 1er plan discriminant, puis prédisez visuellement sa qualité. Vous préciserez quelle règle vous avez utilisé et noterez $\widehat{x}$ le point résultant.

$\widehat{x}_x = {}^t x\ LD_x = [1, 600, 5, 11, 3][0.94, -0, -0.1, 0.15, 0.3] = 2.99$ **et**

$\widehat{x}_y = {}^t x\ LD_y = [1, 600, 5, 11, 3][0.82, 0, 0.18, 0.54, -0.05] = 7.51$

4) Calculez, par produit scalaire, la distance euclidienne entre $\widehat{x}$ et $\overline{x}_3$.

$d^2(\widehat{x}, \overline{x}_3) = {}^t(\widehat{x} - \overline{x}_3)(\widehat{x} - \overline{x}_3) = [0.89, -2.67][0.89, -2.67] = 7.921$

Exercice 4

1) On a obtenu `Cluster` à partir d'une exécution de l'algorithme des *k-means* sur les données (Exercice 2) et on donne la table de contingence qui croise ce résultat avec la vérité-terrain `Wine`. Calculez la contribution au $\chi^2 = 9$ de la 1ère qualité de vins et du 3ème cluster.

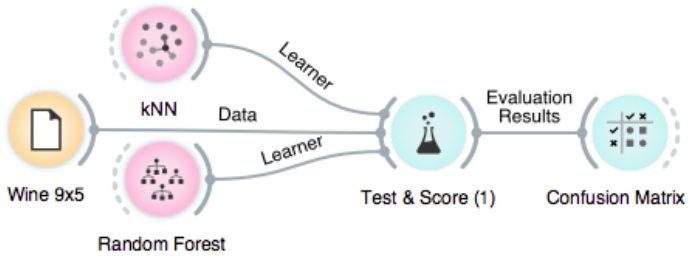
$t_{13} = \frac{3\times1}{9} = \frac{1}{3}$ **et** $e_{13} = \frac{(1-/3)^2}{1/3} = \frac{4}{3}$,
soit $\frac{4/3}{9} = 14.81\%$ **de la valeur du** χ^2 **d'indépendance**

2) Peut-on dire que `Cluster` et `Wine` sont indépendantes ? Cela veut-il dire que le résultat du *clustering* est conforme à la vérité-terrain ?

la p-value vaut 6.01%, un risque assez faible de rejeter à tort l'hypothèse d'indépendance entre *Cluster* et *Wine* ;
elles sont donc plutôt dépendantes et *Cluster* est plutôt conforme à la vérité-terrain

Exercice 5

1) Dessinez le *canvas* permettant de produire les fenêtres données.



2) Calculez la valeur de Rappel de la 1ère catégorie de vins, puis la valeur de Précision de la 2ème.

confusion vins (1) :

2	**1**
1	**5**

alors $R_1 = \frac{2}{2+1} = 2/3 = 0.67$

confusion vins (2) :

2	**2**
2	**3**

alors $P_2 = \frac{2}{2+3} = 2/5 = 0.4$

3) Étant donnée que la valeur de Précision de la 1ère catégorie de vins vaut 2/3, que vaut le score global manquant ?

comme $P_3 = 0$ **(voir confusion), alors** $P = \frac{3\times\frac{2}{3}+3\times\frac{2}{5}+2\times0}{9} = 0.40$