Examen
Introduction à la Science des Données

RICM5 (2015-2016)

Auteurs: Massih-Reza Amini, Ahlame Douzal
Durée : 2 heures, documents de cours autorisés
Les deux parties sont à rédiger et à rendre sur feuilles séparées

20 janvier 2016

1 Partie I (1h)

Question 1 (4 pt)
Soit la base d'apprentissage de taille 4 suivante :

$$S = \left\{ \left(\begin{pmatrix} 1 \\ 3 \end{pmatrix}, +1 \right); \left(\begin{pmatrix} -3 \\ -1 \end{pmatrix}, +1 \right); \left(\begin{pmatrix} 1 \\ -3 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 3 \\ -1 \end{pmatrix}, -1 \right) \right\}$$

1.1 (0.5 pt) Dessiner les points dans un repère orthonormé du plan.

1.2 (1.5 pt) On considère le modèle de perceptron pour séparer les points de classe $+1$ à ceux de la classe -1. On suppose que le vecteur des poids initial est le vecteur null, que le pas d'apprentissage est fixé à 1, et que le biais $w_0 = 0$. Quel est dans ce cas, le vecteur poids trouvé par l'algorithme du perceptron après 2 mises à jour si l'on considère que les mises à jour se font dans le sens trigonométrique en commençant par le point de coordonnées $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$, c.à.d. $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ puis $\begin{pmatrix} -3 \\ -1 \end{pmatrix}$ puis $\begin{pmatrix} 1 \\ -3 \end{pmatrix}$ puis ... ?

1.3 (0.5 pt) Quelle est l'équation de la droite séparatrice trouvée après ces deux mises à jour ?

1.4 (1 pt) À quelle distance se trouve les points de la base d'apprentissage à cette droite ? En déduire la marge.

1.5 (0.5 pt) Le résultat du théorème de Novikoff se vérifie-t-il sur cet exemple ?

Question 2 (3 pt)

On applique l'algorithme d'Adaboost sur une base d'apprentissage de taille 10 ;

$$S = \{(\mathbf{x}_i, y_i); i \in \{1, \ldots, 10\}\} \in (\mathcal{X} \times \{-1, +1\})^{10}$$

2.1 (1 pt) À l'étape 1 les exemples sont assignés des poids uniformes : $\forall i, D_1(i) = \frac{1}{10}$. On suppose qu'après la phase d'entraînement, le premier classifieur $h_1 : \mathcal{X} \to \{-1, +1\}$ classifie mal 3 exemples de S. Estimer l'erreur $\epsilon_1 = \sum_{i:h_1(\mathbf{x}_i) \neq y_i} D_1(i)$ et en-déduire le poids α_1 associé à h_1 trouvé par l'algorithme.

2.2 (2 pt) Calculer les nouveaux poids D_2 des exemples mal classés et des exemples bien classés par h_1.

Question 3 (3 pt)

On considère que l'espace d'entrée est de dimension d, $\mathcal{X} \subseteq \mathbb{R}^d$. Calculer les gradients des erreurs instantanées convexes bornant l'erreur $0/1$ suivant en $\mathbf{w} \in \mathbb{R}^d$:

$$\begin{aligned} \ell_q(\mathbf{x}, y, \mathbf{w}) &= (y - \langle \mathbf{w}, \mathbf{x} \rangle)^2 \\ \ell_l(\mathbf{x}, y, \mathbf{w}) &= \ln(1 + e^{-y\langle \mathbf{w}, \mathbf{x} \rangle}) \\ \ell_e(\mathbf{x}, y, \mathbf{w}) &= e^{-y\langle \mathbf{w}, \mathbf{x} \rangle} \end{aligned}$$

2 Partie II (1h)

2.1 Questions de cours

1. Quelles interprétations correspondent aux valeurs -1, 0 et 1 prises par :
   a) le taux de Kendall, b) le coefficient des rangs de Spearman ?
2. Enumérez les stratégies adoptées pour mesurer la similarité entre des variables binaires non symétriques ?
3. Que représente la validation croisée en apprentissage machine ?
4. Appliquez à la main l'algorithme des k-means aux données représentées dans la Figure 1 ; les centres initiaux étant A_1, A_4 and A_7. Illustrez les partitions obtenues à chaque itération.

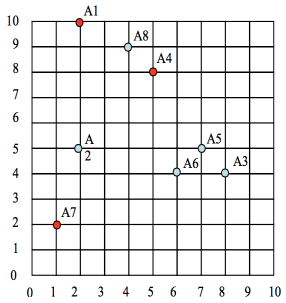

FIGURE 1 –

5. On applique l'algorithme PAM (Partitioning Around Medoids) aux données représentées dans la Figure 2. Indiquez les régions (de A à F) où seront sélectionnés les 3 premiers medoids durant la phase "BUILD". Justifiez vos réponses.

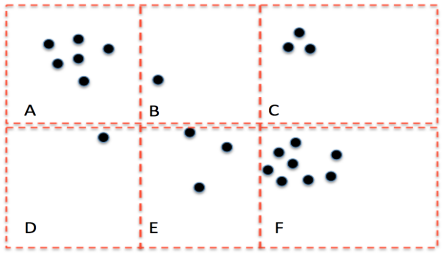

FIGURE 2 –

2.2 Analyse des données de cépages

On s'intéresse à l'analyse des données Wines [1] donnant la description de 177 vins issus de trois cultivars (i.e. plants de vigne) et provenant de la région du Piémont en Italie. La composition des vins est mesurée par 13 analyses chimiques et spectroscopiques ("alcohol", "malic acid", "ash", "ash alkalinity", "magnesium", "tot. phenols", "flavonoids" , "non-flav. phenols", "proanth", "col. int.", "col. hue", "OD ratio" et "proline"). Un 14ème descripteur "classeW" indique le cultivar de chaque vin. Les 13 descripteurs étant de type numérique et "classeW" nominal. Le tableau suivant indique,

1. http ://kdd.ics.uci.edu

à titre d'exemple, la composition de 3 échantillons de vins par 4 descripteurs chimiques :

```
     alcohol     malic acid    ash      ash alkalinity ...
[1,]   13.20       1.78        2.14           11.2  ...
[2,]   13.16       2.36        2.67           18.6   ...
[3,]   14.37       1.95        2.50           16.8  ...
....
```

2.2.1 Analyse des dépendances

La figure 1 représente 3 distributions des 177 vins décrites respectivement par les couples ("tot. phenols", "flavonoids"), ("magnesium", "ash alkalinity") et ("alcohol", "proline") :

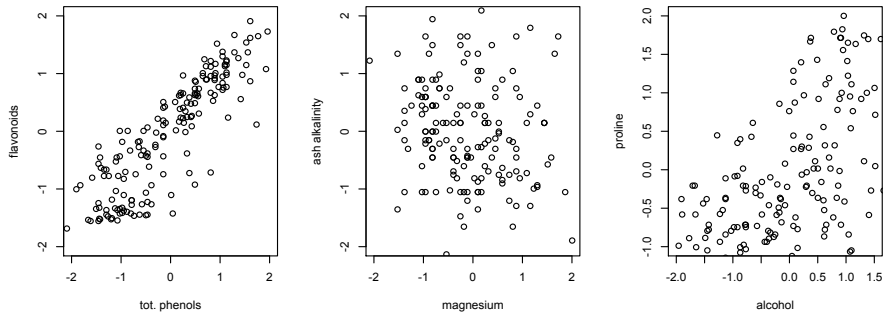

FIGURE 3 – Distributions des 177 vins

A partir de ces 3 distributions, analysez les dépendances entre les 3 couples de descripteurs.

2.2.2 Prédiction de cultivars

Notre objectif, dans cette section, est la prédiction des cultivars à partir de la composition chimique et spectroscopique des vins. Pour cela, une classification par arbre est utilisée, les résultas obtenus sont indiqués ci-dessous :

Caractéristiques de l'arbre :

```
node), split, n, loss, yval, (yprob)
```

```
    * denotes terminal node
1) root 177 106 2 (0.32768362 0.40112994 0.27118644)
  2) proline>=0.0314527 66  10 1 (0.84848485 0.06060606 0.09090909)
    4) flavonoids>=0.1417439 58   2 1 (0.96551724 0.03448276 0.00000000) *
    5) flavonoids< 0.1417439 8   2 3 (0.00000000 0.25000000 0.75000000) *
  3) proline< 0.0314527 111  44 2 (0.01801802 0.60360360 0.37837838)
    6) OD ratio>=-0.6939325 65   4 2 (0.03076923 0.93846154 0.03076923) *
    7) OD ratio< -0.6939325 46   6 3 (0.00000000 0.13043478 0.86956522)
     14) col. hue>=-0.2486876 7   2 2 (0.00000000 0.71428571 0.28571429) *
     15) col. hue< -0.2486876 39   1 3 (0.00000000 0.02564103 0.97435897) *
```

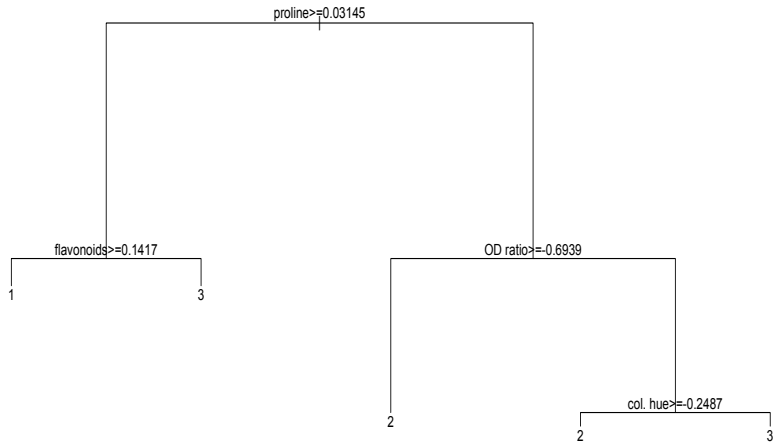

FIGURE 4 – Arbre de classification pour la prédiction de cultivars

1. Interprétez les informations relatives aux noeuds étiquetés 2) et 4).
2. Evaluez l'EAC (l'Erreur Apparente de Classification) de l'arbre induit.
3. Quel est le cultivar le mieux prédit ?
4. Indiquez les règles discriminantes retenues par l'arbre pour la prédiction des cultivars.