HEIG-VD

Rapport

MLG Labo1

Auteur :
Basile Châtillon
Valentin Finini

7 mars 2018

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD

www.heig-vd.ch

Question 1

Regarding the wine database, by looking at the boxplots generated during the Exploratory analysis of data (section 6) , which features seems the most discriminative ? why ?

Nous avons décidé de prendre la feature des flavonoïdes. En effet, pour différencier les vins, il faut prendre une caractéristique telle que le recouvrement entre les boîtes à moustache est minimale. En effet, la boîte de la boîte à moustache contient 50% de la population. Dès lors, on peut déjà être sur qu'une majorité de vins seront bien classés. En se basant sur ce critère, on peut donc constater que le meilleur est donc celui des flavonoïdes. Un autre bon critère aurait été de prendre le "total phenols". Les boîtes ne se chevauchent pas non plus se qui garantie déjà une bonne précision. Par contre on peut cette fois constater que les moustaches sont plus étendues. Du coup cela signifie que plus de vins seront possiblement mal classés.

Question 2

Can you estimate the performance of a single-rule classification method like the one presented in section 7 ?

Pour estimer la performance de la "single-rule" nous avons divisé le nombre de prédictions correctes par le nombre total de vin à classer. On a donc trouvé que pour la règle du taux d'alcool contenu dans le vin, que la performance est d'environ 33,7%. Cela signifie qu'un peu moins d'un tiers des fois ou on cherche à prédire la classe en utilisant cette règle, notre prédiction s'avérera incorrecte. On est presque entrain de tirer la classe au hasard. On en conclut que la performance de cette règle est mauvaise.

Question 3

Define a rule that uses the most discriminative feature to classify the wine observations.

Nous avons donc repris la méthode de classification proposée pour le taux d'alcool contenu dans le vin que nous avons adapté au taux de flavonoïdes. Pour définir les valeurs limites des différentes classes, nous avons pris pour la première, le quartile de 25% de la classe n°1 des vins. Et pour la deuxième, le quartile de 75% de la classe n°3.

L'implémentation en Python de la règle se trouve ci-dessous.

```
pred = []

# The values below are taken from the lower and upper quartile
for row in df['flavanoids']:
    if row >= 2.68: # quartile_25% class 1
        pred.append(1);
    elif row > 0.92: # quartile_75% class 3
        pred.append(2);
    else:
        pred.append(3)

# A new column is added to the dataframe
df['prediction_flavanoids'] = pred
```

Question 4

Compute the confusion matrix of the resulting rule-based system defined in Q3.

Voici la matrice de confusion associée à notre règle :

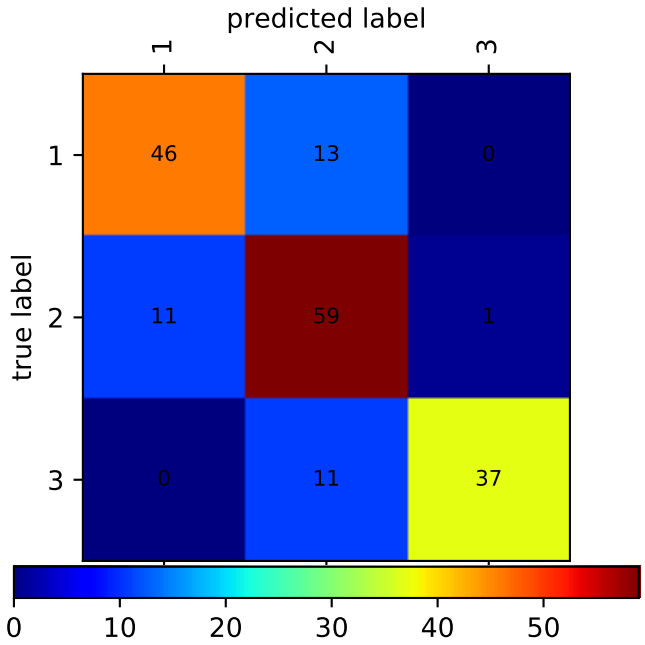

FIGURE 1 – Confusion Matrix

Question 5

Compute the precision, the recall and the F1-score of the classification system defined in Q3 for only one class using the values of the confusion matrix ?

Nous avons décidé de calculer les différentes valeurs pour la classe n°1. Nous avons donc commencé par créer la table de confusion pour la classe n°1 en se basant sur la matrice de confusion calculée précédemment.

		Actual class	
		class 1	non-class 1
Predicted class	class 1	46	11
	non-class 1	13	108

TABLE 1 – Table de confusion de la classe n°1

Nous obtenons donc les différentes valeurs : $tp = 46$, $fp = 1$, $tn = 13$ et $fn = 107$. À partir de ces variables nous pouvons calculer les valeurs demandées.

$$\text{precision} = \frac{tp}{tp + fp} = \frac{46}{46 + 11} \simeq 0.81$$

$$\text{recall} = \frac{tp}{tp + fn} = \frac{46}{46 + 13} \simeq 0.78$$

$$F_1\text{- score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.81 \times 0.79}{0.81 + 0.78} \simeq 0.79$$

Nous avons également utilisé la méthode de la librairie skmetrics.classification_report :

	precision	recall	f1-score	support
1	0.81	0.78	0.79	59
2	0.71	0.83	0.77	71
3	0.97	0.77	0.86	48
avg / total	0.81	0.80	0.80	178

On peut constater que nous avons trouvé les mêmes valeurs que celles retournées par la fonction de la bibliothèque. C'est chouette.