TP1 : Statistiques descriptives

Nicolas Cisternas
cisternas@et.esiea.fr

François Lelievre
flelievre@et.esiea.fr

Romaric Prodhomme
rprodhomme@et.esiea.fr

Hugo Revault d'Allonnes
revaultdallonn@et.esiea.fr

I. Données & analyse monodimensionnelle

Lors de ce premier TP d'ingénierie des données, il nous a été demandé d'étudier le dataset des Iris. Il est composé de 4 descripteurs, 150 exemples et 3 classes. Nous pouvons observer les moyennes (notées 'M') et variances (notées 'V') globales sur la Fig. 1. Dans une optique de classification, nous choisirons le descripteur ayant la variance la plus élevée, c'est-à-dire « petal length ». On remarque ici que le pire descripteur pour faire de la classification est « sepal width ». En effet, plus la variance est élevée, plus il est facile de distinguer les éléments les uns des autres.

```
sepal length (cm): M: 5.843 | V: 0.686
sepal width (cm): M: 3.057 | V: 0.19
petal length (cm): M: 3.758 | V: 3.116
petal width (cm): M: 1.199 | V: 0.581
```

Fig. 1. Moyennes et Variances globales des 4 descripteurs.

II. Analyse conditionnelle

Nous avons dû étudier chaque descripteur à l'intérieur de chaque classe. Nous en avons extrait les moyennes et variances intra-classe que nous pouvons visualiser sur la Fig. 3. Ensuite, grâce au théorème de la variance totale, nous avons pu déterminer les variances interclasses de chacun des descripteurs, visibles sur la Fig. 2. Enfin, nous avons représenté chaque descripteur dans un graphe afin de rendre le tout plus explicite. Les classes 0, 1 et 2 correspondent respectivement à 'setosa', 'versicolor' et 'virginica', voir Fig. 4, Fig. 5, Fig. 6 et Fig. 7. D'après les graphes, on remarque que la faible variance intra-classe et la grande variance interclasse de « petal length » nous permet d'identifier aisément chaque classe. Les classes ne se chevauchent pas et on voit bien que la classe 0 est comprise entre 1 et 2cm, la classe 1 entre 3 et 5cm, la classe 2 entre 5 et 7cm. On le remarque d'autant plus sur la Fig. 6.2.

```
sepal length (cm) : V interclasse 0.421
sepal width (cm) : V interclasse 0.076
petal length (cm) : V interclasse 2.914
petal width (cm) : V interclasse 0.536
```

Fig. 2. Variances interclasses des 4 descripteurs.

```
classe  setosa
sepal length (cm): M: 5.006| V: 0.124
sepal width (cm): M: 3.428| V: 0.144
petal length (cm): M: 1.462| V: 0.03
petal width (cm): M: 0.246| V: 0.011

classe  versicolor
sepal length (cm): M: 5.936| V: 0.266
sepal width (cm): M: 2.77| V: 0.098
petal length (cm): M: 4.26| V: 0.221
petal width (cm): M: 1.326| V: 0.039

classe  virginica
sepal length (cm): M: 6.588| V: 0.404
sepal width (cm): M: 2.974| V: 0.104
petal length (cm): M: 5.552| V: 0.305
petal width (cm): M: 2.026| V: 0.075
```

Fig. 3. Moyennes et Variances globales des 4 descripteurs.

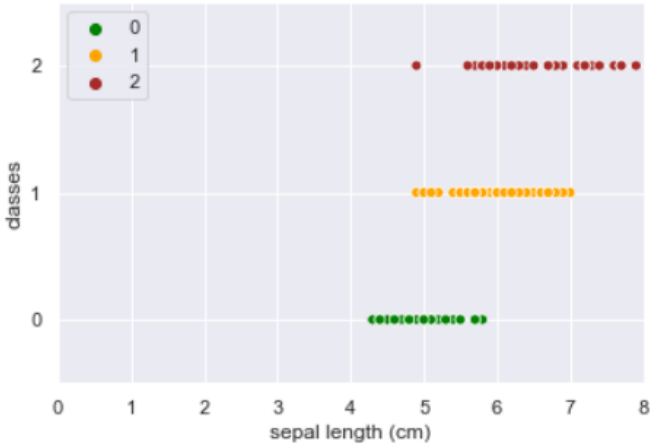

Fig. 4. Répartition de Sepal length.

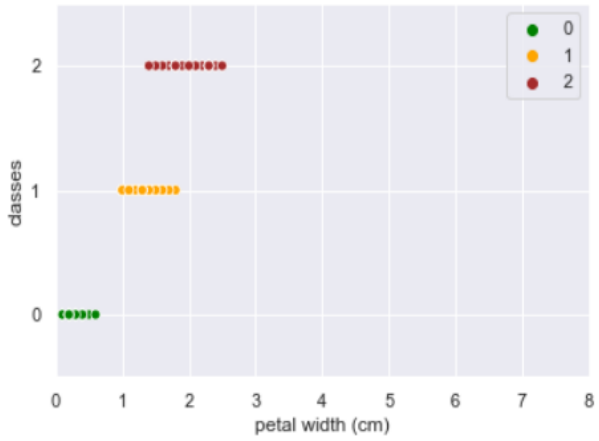

Fig. 5. Répartition de Petal width.

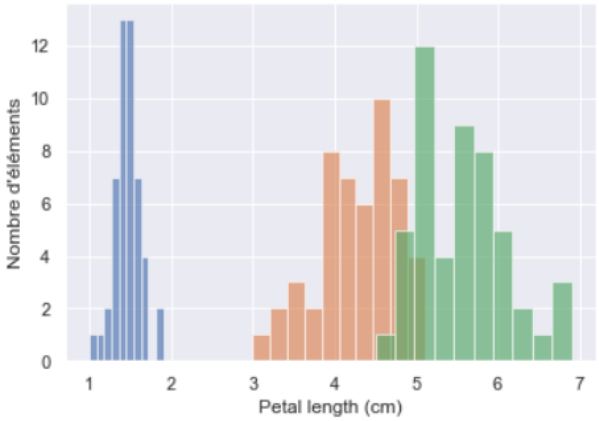

Fig. 6.2. Histogramme de Petal length.

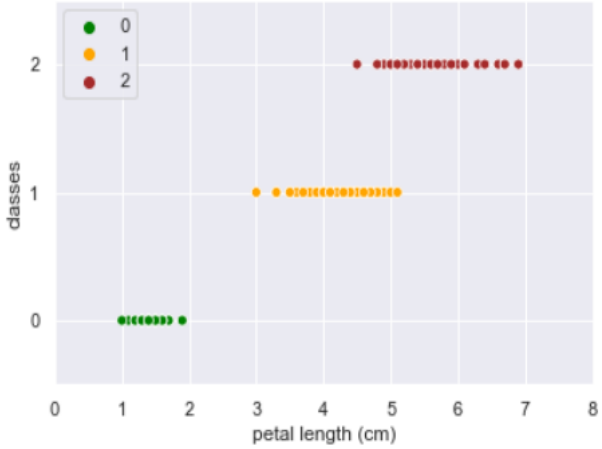

Fig. 6. Répartition de Petal length.

III. Analyse Bidimensionnelle

Dans cette partie nous avons représenté dans des graphes la corrélation entre les descripteurs selon 2 configurations, on a aussi représenté leurs moyennes. Configuration 1: 'sepal length' et 'sepal width', Fig. 8. Configuration 2 : 'sepal width' et 'petal width' Fig. 9. Dans la 1ère configuration, on remarque graphiquement et grâce aux matrices de covariances que les descripteurs sont très corrélés dans la classe 'setosa', c'est-à-dire que pour une iris de classe setosa, il y a un lien apparent entre la longueur des sépales et leur largeur, tandis que dans la 2ème configuration, ils sont très corrélés dans la classe versicolor, c'est-à-dire que pour une iris de classe versicolor, il y a un lien apparent entre la largeur des pétales et celle des sépales.

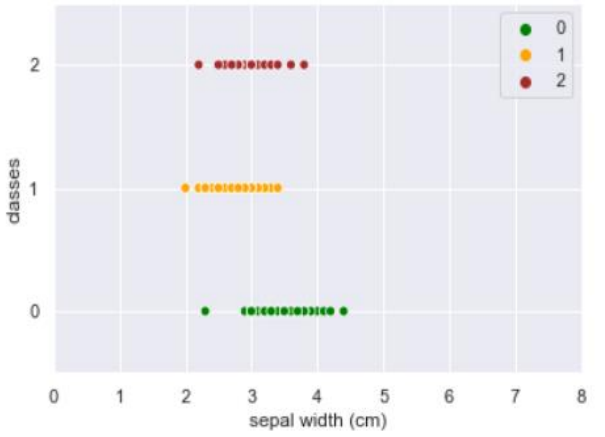

Fig. 7. Répartition de Sepal width.

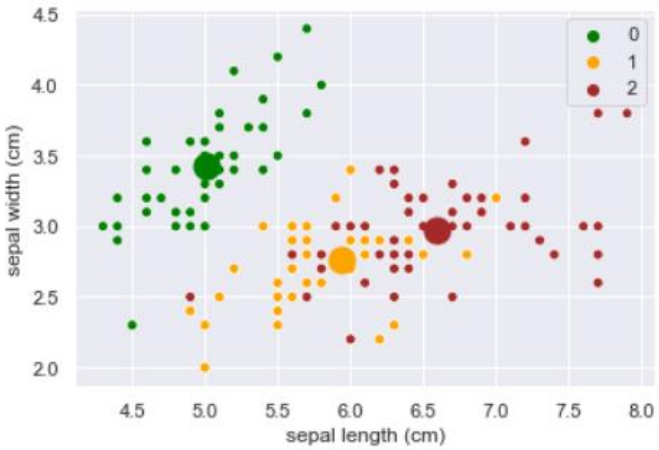

Fig. 8. Sepal width selon sepal length.

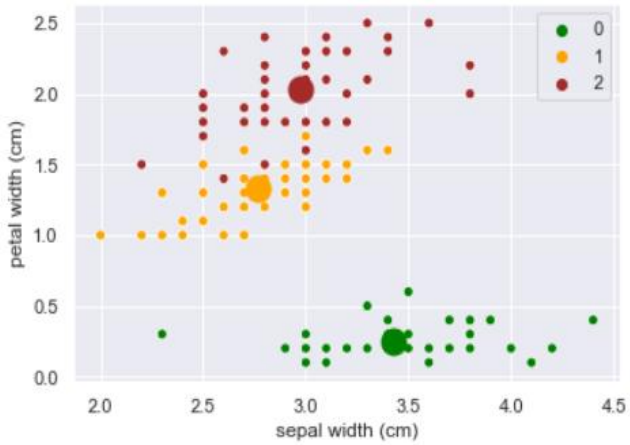

Fig. 9. Petal width selon sepal width.

Enfin, nous avons représenté les frontières de décisions si l'on utilise l'algorithme *nearest*-mean sur la configuration 2, Fig. 10.

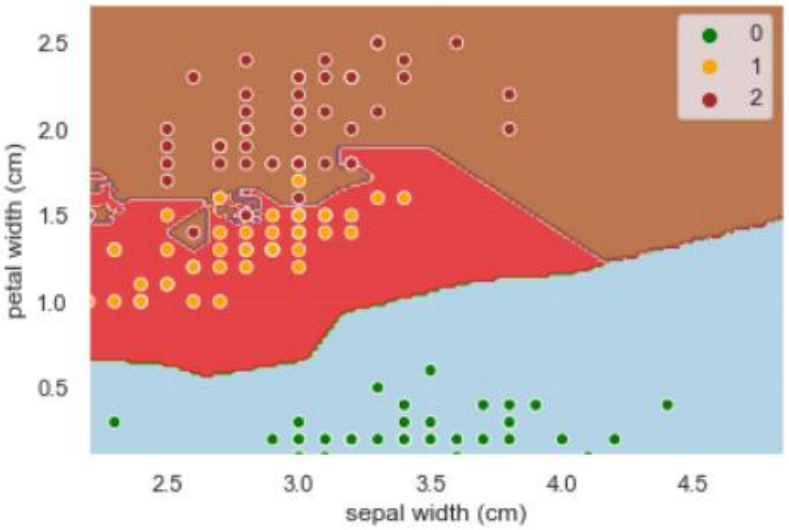

Fig. 10. Représentation des frontières sur la configuration 2.