M1 ISIFAR

Mémoire de T.P.E

L'Analyse Discriminante Linéaire

Zakarya Hakkou & Christian Ramires

Rapport rédigé sous R Markdown

Table des matières

Introduction

L'objet de ce mémoire est de présenter les aspects théoriques et généraux de ce qu'on appelle communément l'analyse discriminante et en particulier l'analyse discriminante linéaire.

En quoi consiste l'analyse discriminante?

En termes simples, l'analyse discriminante est une méthode permettant de décrire un ensemble d'individus et de prédire une variable qualitative à k classes, à l'aide de variables prédictives, généralement numériques. Donc, l'analyse discriminante a pour but de déterminer des différences entre des groupes de données afin d'éclairer des décisions et ainsi développer des solutions efficaces.

Les objectifs sont doubles:

- d'une part, il y a un objectif *descriptif*, c'est-à-dire qu'on cherche les combinaisons linéaires de variables qui permettent de séparer le mieux possible les k catégories de variables qualitatives et ainsi donner une représentation graphique, qui rende compte au mieux de cette séparation.
- d'autre part, l'objectif est *décisionnel*, c'est-à-dire qu'on cherche à décider dans quelle classe affecter un nouvel individu pour lequel on connaît les valeurs des variables prédictives. Nous étudierons cet aspect dans ce rapport, puisque l'analyse discriminante linéaire se trouve dans l'approche décisionnelle.

Dans le cadre professionnel, les champs d'application de l'analyse discriminante sont multiples. Nous pouvons citer quelques domaines:

- **Médecine**: à partir de mesures de laboratoire, on cherche une fonction discriminante permettant de prédire au mieux le type de maladie d'un patient, ou de son évolution afin d'orienter le traitement.
- **Finance**: prévision du comportement et du risque des demandeurs de crédit par exemple - préoccupation primordiale des banques -. En effet, l'usage de l'analyse discriminante au sein des banques et assurances s'avère très utile et efficace.

Ce rapport rassemble une explication détaillée de deux méthodes décisionnelles: l'analyse discriminante linéaire puis l'analyse discriminante quadratique -appuyée de preuves et d'exemples concrets-. Enfin, nous commenterons les avantages et les limites de ces outils de décision.

Notre problématique est donc la suivante: quels sont les liens et les différences existants entre l'analyse discriminante linéaire et l'analyse discriminante quadratique? Quelle est l'analyse la plus optimale?

Dans ce rapport, nous tâcherons donc de répondre in fine à cette problématique.

Préliminaires

On dispose d'un échantillon de n individus sur lequel on a observé des variables explicatives $X_i = (X_i^1, ..., X_i^p)$ et une variable qualitative $Y_i \in 1, ..., K$ qui correspond à chaque classe de l'individu i.

A l'aide de ces variables, on va prédire un nouvel individu $n+1$ en tenant compte de ses propriétés que l'on note $(X_{n+1}^1, ..., X_{n+1}^p)$.

Notations:

- $C_1, ..., C_K$ sont les classes.
- $(\pi_1, ..., \pi_K)$ est la distribution de Y où $\pi_k = \mathbb{P}(Y = k)$ est la probabilité d'appartenir à la classe k et $\sum_{k=1}^{K} \pi_k = 1$.
- f_k est la densité conditionnelle du vecteur $X \in \mathbb{R}^p$ sachant que l'on est dans la classe C_k.
- $\mu_k \in \mathbb{R}^p$ est le vecteur des moyennes théoriques dans un cas gaussien.
- Σ_k est la matrice de variance-covariance théorique dans un cas gaussien.

I. Analyse Discriminante Linéaire

a) Présentation de l'Analyse Discriminante Linéaire

L'analyse discriminante linéaire a pour objectif d'expliquer et de prédire les valeurs d'une variable qualitative, notée Y, à partir de variables explicatives quantitatives et/ou qualitatives, notées $X = (X_1, ...X_p)$.

Cette méthode s'appuie sur deux hypothèses:

- les densités conditionnelles sont gaussiennes, c'est-à-dire que:

$$f_k(x) = \frac{1}{(\sqrt{2\pi})^p \sqrt{det\Sigma_k}} exp(-\frac{1}{2} *^t (x - \mu_k)\Sigma_k^{-1}(x - \mu_k))$$

- l'homoscédasticité, c'est-à-dire qu'il y a égalité des matrices de variances-covariances conditionnelles:

$$\Sigma_1 = ... = \Sigma_K = \Sigma$$

L'ADL peut être présentée à travers l'approche bayesienne, qui consiste à modéliser la probabilité d'appartenance à une certaine classe et donc à lui affecter une nouvelle observation.

Formule de Bayes:

$$\mathbb{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

avec:

- f_k: la densité conditionnelle du vecteur $X \in \mathbb{R}^p$ sachant que l'on est dans la classe C_k
- $\pi_k = \mathbb{P}(Y = k)$: la probabilité d'appartenir à la classe C_k et $\sum_{k=1}^{K} \pi_k = 1$

Afin de trouver la probabilité conditionnelle - $\mathbb{P}(Y = k|X = x)$ - la plus grande, une méthode consiste à chercher le maximum de vraisemblance:

$$\hat{Y} = \underset{k\in(1,\dots,K)}{\operatorname{argmax}}\ \mathbb{P}(Y = k|X) \iff \hat{Y} = \underset{k\in(1,\dots,K)}{\operatorname{argmax}}\ log\mathbb{P}(Y = k|X) \quad (1)$$

Analyse de la vraisemblance dans le cas où il y a deux classes (K = 2)

Soit un échantillon dans lequel il existe deux classes, notées C_k et C_l. Il suffit d'analyser le log-ratio suivant, que l'on note R:

$$R = log\frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = l|X = x)} = log\frac{\pi_k}{\pi_l} + log\frac{f_k(x)}{f_l(x)}$$

$$R = log\frac{\pi_k}{\pi_l} + {}^t x\Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2} * {}^t\mu_k\Sigma^{-1}\mu_k + \frac{1}{2} * {}^t\mu_l\Sigma^{-1}\mu_l$$

- Si $R > 0$, alors $P(Y = k|X = x) > \mathbb{P}(Y = l|X = x)$. Donc Y appartient à la classe C_k.
- Si $R < 0$, alors $P(Y = k|X = x) < \mathbb{P}(Y = l|X = x)$. Donc Y appartient à la classe C_l.

Analyse de la vraisemblance dans le cas où il y a plus de deux classes (K>2)

Une autre manière de résoudre (1) serait d'utiliser la fonction linéaire discriminante. Cela revient à choisir:

$$\hat{Y} = \underset{k\in(1,\dots,K)}{\operatorname{argmax}}\ \delta_k(x) \ (2)$$

Preuve

On sait que:

$$\mathbb{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Comme $\sum_{l=1}^{K} \pi_l f_l(x)$ ne dépend pas de k, alors:

$$\hat{Y} = \underset{k\in(1,\dots,K)}{\operatorname{argmax}}\ \pi_k f_k(x)$$

Ici:

$$\pi_k f_k(x) = \pi_k \frac{1}{(\sqrt{2\pi})^p\sqrt{det\Sigma}} exp(-\frac{1}{2} * {}^t(x - \mu_k)\Sigma^{-1}(x - \mu_k))$$

En passant au logarithme, on obtient:

$$\log \pi_k f_k(x) = \log \pi_k - \frac{p}{2} log(2\pi) - \frac{1}{2}\log(det\Sigma) - \frac{1}{2} * {}^t(x - \mu_k)\Sigma^{-1}(x - \mu_k)$$

Comme $\frac{p}{2}log(2\pi)$ et $\frac{1}{2}log(det\Sigma)$ sont constantes alors le problème (1) devient:

$$\hat{Y} = \underset{k\in(1,...,K)}{\text{argmax}} \left(log(\pi_k) - \frac{1}{2} *^t (x-\mu_k)\Sigma^{-1}(x-\mu_k) \right)$$

En développant, on retrouve:

$$\hat{Y} = \underset{k\in(1,...,K)}{\text{argmax}} \left(log(\pi_k) +^t x.\Sigma^{-1}\mu_k - \frac{1}{2} *^t \mu_k\Sigma^{-1}\mu_k \right)$$

Ainsi, on trouve finalement:

$$\hat{Y} = \underset{k\in(1,...,K)}{\text{argmax}} \ \delta_k(x) \ (2) \quad avec: \quad \delta_k(x) = log(\pi_k) + x.\Sigma^{-1}\mu_k - \frac{1}{2} *^t \mu_k\Sigma^{-1}\mu_k$$

Les estimations

En pratique, on ne connaît ni les $(\pi_k)_{k=1}^K$, ni les $(\mu_k)_{i=1}^K$, ni Σ. Il faut donc les estimer par:

- $\hat{\pi}_k = \frac{n_k}{n}$ où $n_k = \sum_{i=1}^k 1_{Y_i=k}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i\in C_k} X_i$
- $\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^K \sum_{i\in C_k} (X_i - \hat{\mu}_k) \ ^t(X_i - \hat{\mu}_k)$ (Estimateur sans biais)

b) Algorithme de l'ADL appliqué à la base de données Iris

Nous nous intéressons dans cette étude à la base de données **Iris de Fisher**, qui constitue un dataset de référence pour l'analyse discriminante. Ce jeu de données contient 50 fleurs de chacune des 3 espèces de fleurs (Setosa, Virginica et Versicolor) - soit 150 fleurs - et donne les mesures en centimètres des longueurs et largeurs des sépales et pétales.

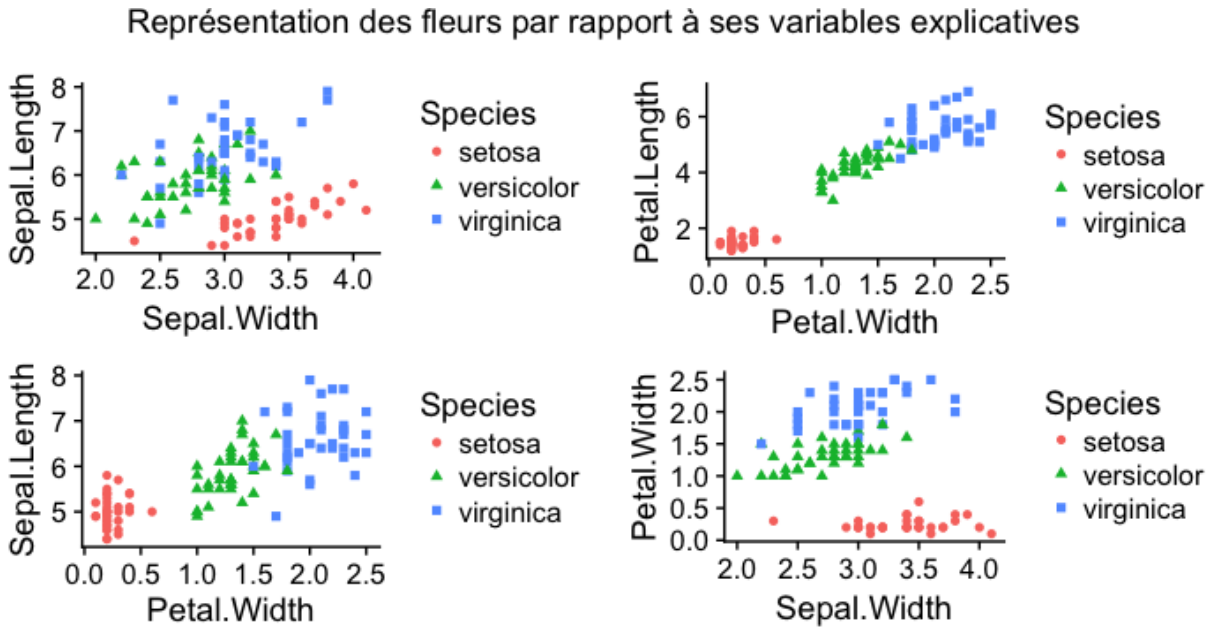
Les variables quantitatives sont:

- *Sepal.Length*: Longueur des sépales
- *Sepal.Width*: Largeur des sépales
- *Petal.Length*: Longueur des pétales
- *Petal.Width*: Largeur des pétales

La variable qualitative est:

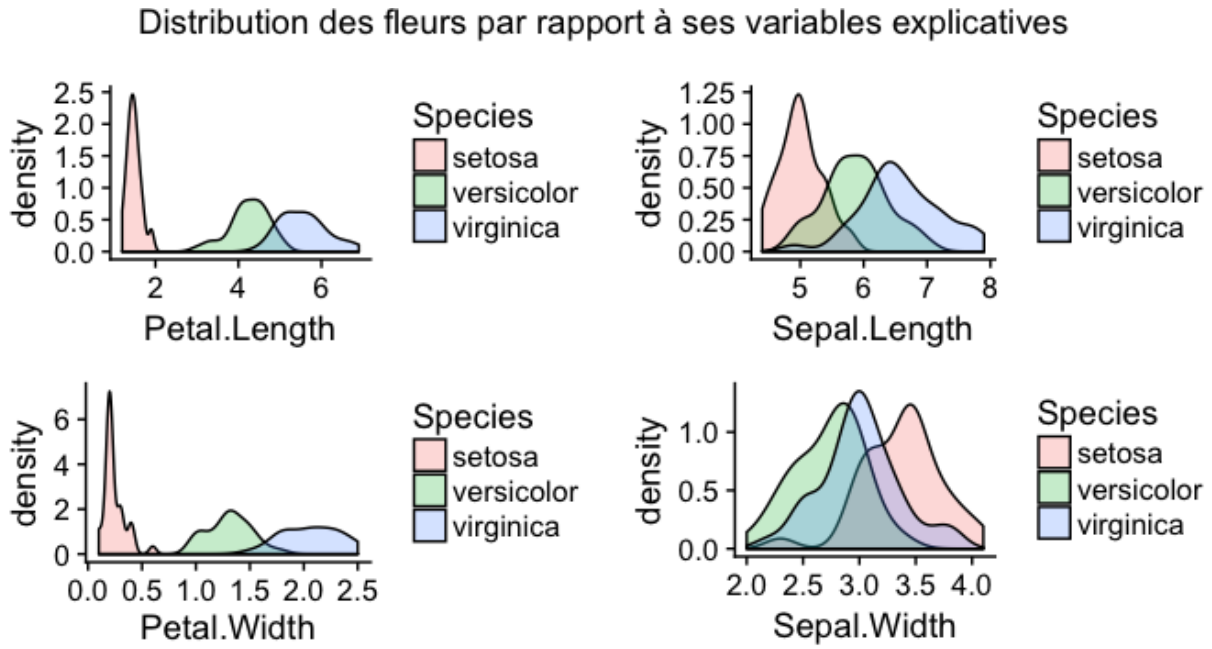
- *Species*: Espèces de fleurs

Nous prenons 110 fleurs parmi 150 afin de constituer la base de données d'apprentissage.
L'objectif est d'essayer de prédire le type d'espèce (setosa, virginica, versicolor) des 40 fleurs restantes, grâce à l'analyse discriminante linéaire.



Cette analyse montre qu'il existe une séparation flagrante entre les espèces(setosa, virginica, versicolor), en fonction de leur caractéristique(longueur sépales, largeur sépales, longueur pétales, largeur pétales). Cependant, nous pouvons voir que dans le premier graphique, comparant la longueur et largeur des sépales, cette séparation n'est pas constatée entre les virginica et les versicolor.

Maintenant, nous allons analyser la distribution des 110 fleurs par rapport à leurs variables explicatives afin de pouvoir se placer dans les hypothèses de l'ADL (densités gaussiennes et homoscédasticité).

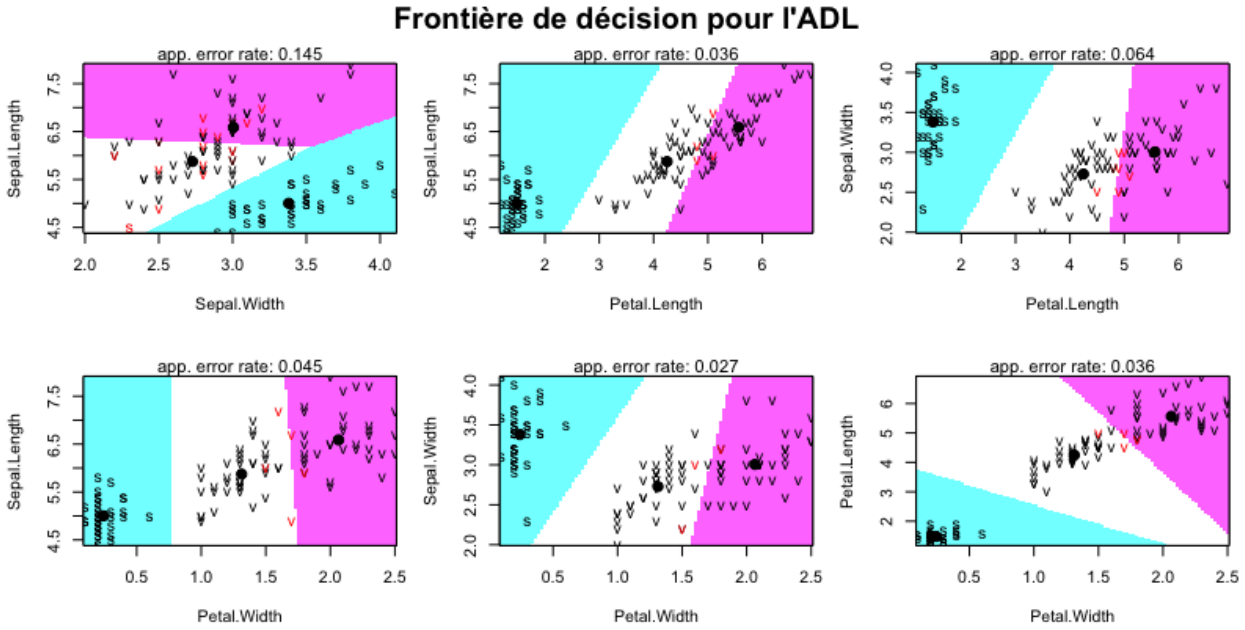


On constate que les densités sont semblables à une densité gaussienne, on peut donc supposer l'hypothèse de gaussianité.

De plus, on remarque que les variances des fleurs "versicolor" et "virginica" sont quasiment les mêmes, contrairement à la variance des fleurs "setosa".

Même si l'hypothèse d'homoscédasticité n'est pas vérifiée, on la supposera afin de pouvoir utiliser l'ADL et ainsi connaître son efficacité.

Afin de représenter les frontières de décision, on utilise la commande “partimat” du package “KlarR”.



On constate que chaque graphique est divisé en trois zones par des fonctions linéaires:

- la zone bleu clair représente les fleurs setosa
- la zone blanche représente les fleurs versicolor
- la zone magenta représente les fleurs virginica

Il ne faut pas oublier que chaque représentation a un certain taux d’erreur, affiché sur chaque graphique.

Afin de connaître le type de fleur d’un nouvel individu, on peut voir ses caractéristiques. Grâce à ses spécificités, nous pourrons placer la fleur sur une des trois zones. Par exemple, si un nouvel individu est placé sur la zone bleu clair, nous pouvons dire qu’il s’agit d’une “setosa”.

Commentaires

A l’aide de l’algorithme de l’ADL qu’on a programmé, on obtient les résultats suivants:

- On trouve qu’il y a ... setosa qui ont été prédites parmi ... de notre échantillon de 40 fleurs.

```
## [1] 12 13
```

- On trouve donc ... versicolor prédites parmi ... de notre échantillon de 40 fleurs.

```
## [1] 13 16
```

- Enfin, on trouve ... virginica prédites parmi ... de notre échantillon de 40 fleurs.

```
## [1]  8 11
```

Notre algorithme nous donne une erreur globale de:

```
## [1] 0.175
```

II. Analyse Discriminante Quadratique

a) Présentation de l'Analyse Discriminante Quadratique

L'analyse discriminante quadratique a le même objectif que l'ADL, c'est-à-dire qu'elle permet d'expliquer et prédire les valeurs d'une variable qualitative, notée Y, à partir de variables explicatives quantitatives et/ou qualitatives, notées $X = (X_1, ...X_p)$.

Cette méthode s'appuie également sur deux hypothèses:

- les densités conditionnelles sont gaussiennes.
- l'hétéroscédasticité, c'est-à-dire que les matrices de variances-covariances des variables explicatives sont différentes/hétérogènes -les Σ_k différents - pour chaque groupe.

Si ces 2 hypothèses sont vérifiées, alors on aboutit à des fonctions discriminantes quadratiques.

En ADQ, les probabilités sont toujours calculées à l'aide du théorème de Bayes, en faisant pour chaque groupe l'hypothèse de normalité $Y|X = k \sim \mathcal{N}(\mu_k, \Sigma_k)$ pour $k = 1, ..., K$.

Afin de trouver la probabilité conditionnelle - $\mathbb{P}(Y = k|X = x)$ - la plus grande, une méthode consiste à chercher le maximum de vraisemblance comme pour l'ADL:

$$\hat{Y} = \underset{k\in(1,...,K)}{\operatorname{argmax}} \ \mathbb{P}(Y = k|X) \iff \hat{Y} = \underset{k\in(1,...,K)}{\operatorname{argmax}} \ log\mathbb{P}(Y = k|X) \quad (3)$$

Analyse de la vraisemblance dans le cas où il y a deux classes (K = 2)

Soit un échantillon dans lequel il existe deux classes, notées C_k et C_l. Il suffit d'analyser le log-ratio suivant, que l'on note R:

$$R = log\frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = l|X = x)} = log\frac{\pi_k}{\pi_l} + log\frac{f_k(x)}{f_l(x)}$$
$$R = log\frac{\pi_k}{\pi_l} - \frac{1}{2}\ {}^t(x - \mu_k)\Sigma_k^{-1}(x - \mu_k) + \frac{1}{2}\ {}^t(x - \mu_l)\Sigma_l^{-1}(x - \mu_l)$$

- Si $R > 0$, alors $P(Y = k|X = x) > \mathbb{P}(Y = l|X = x)$. Donc Y appartient à la classe C_k.
- Si $R < 0$, alors $P(Y = k|X = x) < \mathbb{P}(Y = l|X = x)$. Donc Y appartient à la classe C_l.

Analyse de la vraisemblance dans le cas où il y a plus de deux classes (K > 2)

Une autre manière de résoudre (3) serait d'utiliser la fonction quadratique discriminante. Cela revient à choisir:

$$\hat{Y} = \underset{k\in(1,...,K)}{\operatorname{argmax}} \ \delta_k(x) \ (4) \quad avec: \quad \delta_k(x) = log(\pi_k) - \frac{1}{2}log(\det\Sigma_k) - \frac{1}{2}\ {}^t(x - \mu_k)\Sigma_k^{-1}(x - \mu_k)$$

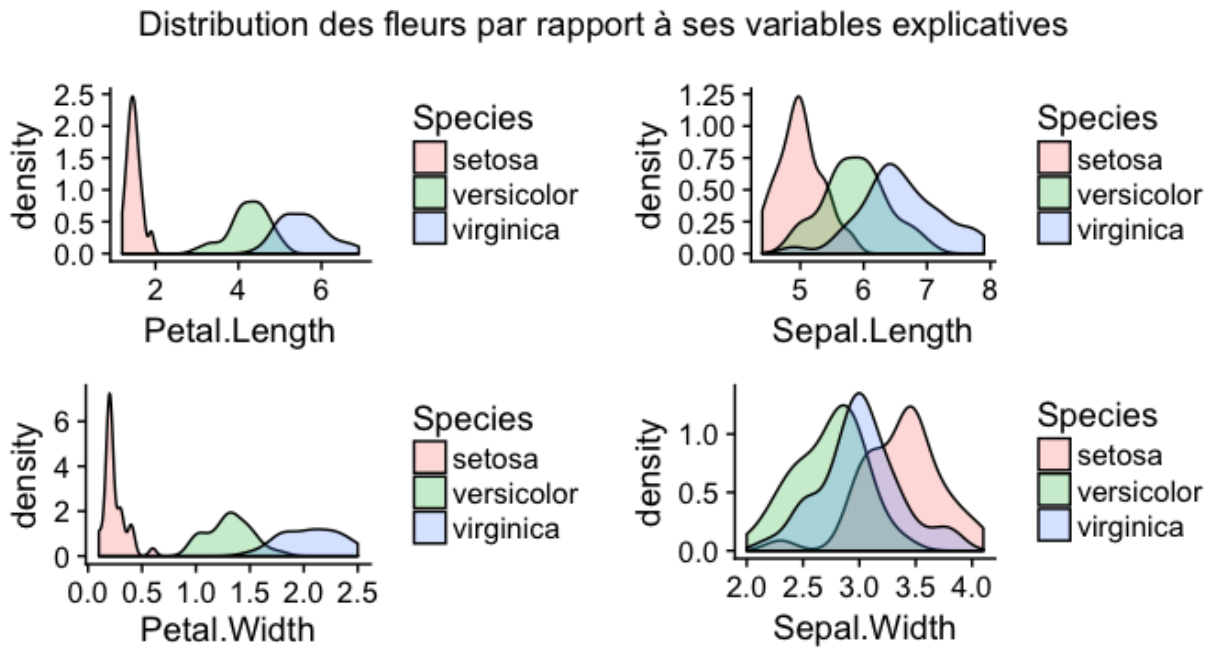
Les estimateurs

- $\hat{\pi}_k = \frac{n_k}{n}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i \in C_k} X_i$
- $\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i \in C_k} (X_i - \hat{\mu}_k)^t (X_i - \hat{\mu}_k)$ (Estimateur sans biais)

b) Algorithme de l'ADQ appliqué à la base de données Iris

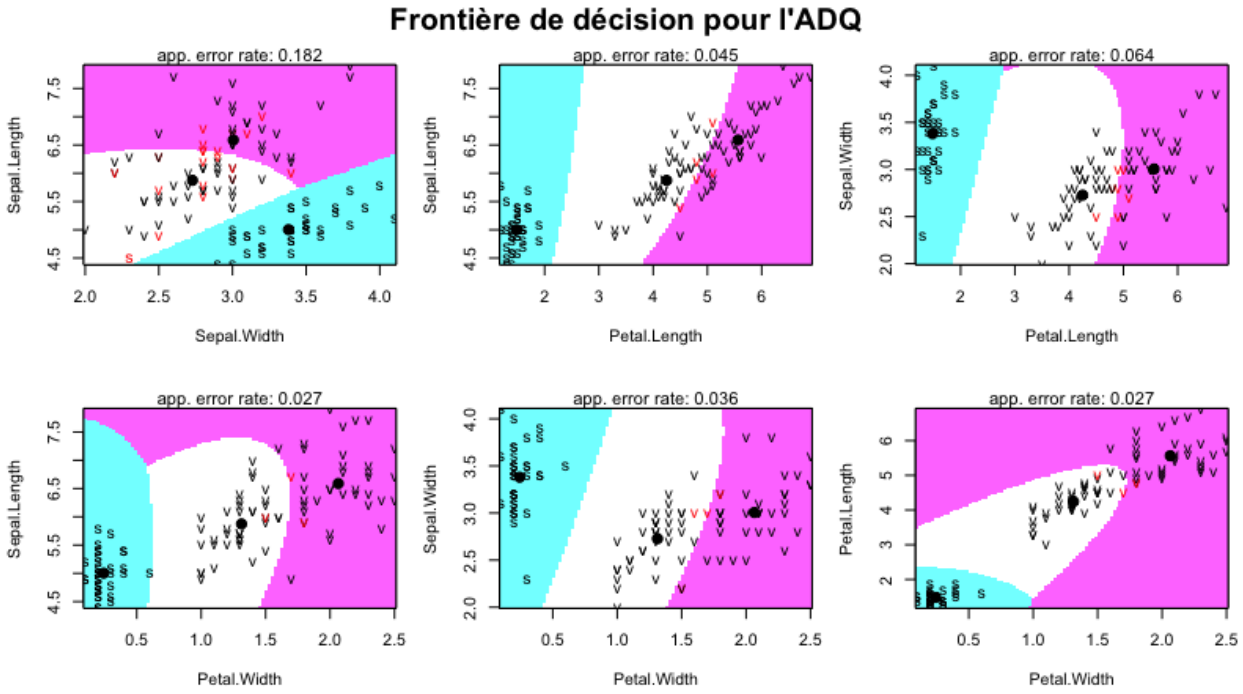
Dans cette partie, nous nous intéressons également à la base de données Iris et nous allons comparer, à travers ce dataset, l'ADQ à l'ADL.

Maintenant, nous allons analyser la distribution des 110 fleurs par rapport à leurs variables explicatives afin de pouvoir se placer dans les hypothèses de l'ADQ (densités gaussiennes et hétéroscédasticité).



On constate que les densités sont gaussiennes et les variances des fleurs "setosa", "versicolor" et "virginica" sont différentes. On peut donc supposer les deux conditions de l'ADQ et ainsi continuer notre analyse.

Afin de représenter les frontières de décision, on utilise la commande “partimat” du package “KlarR”.



Nous constatons que chaque graphique est divisé en trois zones par des fonctions quadratiques:

- la zone bleu clair représente les fleurs setosa
- la zone blanche représente les fleurs versicolor
- la zone magenta représente les fleurs virginica

Commentaires

A l’aide de l’algorithme de l’ADQ qu’on a programmé, on obtient les résultats suivants:

- On trouve qu’il y a . . . setosa qui ont bien été prédites parmi . . . de notre échantillon de 40 fleurs.

```
## [1] 13 13
```

- On trouve qu’il y a . . . versicolor qui ont été bien prédites parmi . . . de notre échantillon de 40 fleurs.

```
## [1] 12 16
```

- On trouve qu’il y a . . . virginica qui ont été bien prédites parmi . . . de notre échantillon de 40 fleurs.

```
## [1] 11 11
```

Notre algorithme nous donne une erreur globale de:

```
## [1] 0.1
```

Conclusion

Tout d'abord, nous avons pu définir dans ce rapport un cadre de travail sur lequel se baser en présentant l'analyse discriminante et plus particulièrement en comparant l'ADL et l'ADQ.

Pour déterminer le maximum de vraisemblance, l'ADL utilise la fonction linéaire discriminante alors que l'ADQ utilise la fonction quadratique discriminante. Donc, la frontière de décision pour chacune des deux méthodes est différente et nous avons bien pu le remarquer sur les graphiques que nous avons tracé.

En ce qui concerne l'efficacité des deux méthodes, l'ADQ est plus performante que l'ADL quand les variances des classes sont différentes. Néanmoins, l'ADL donne des résultats satisfaisants lorsque les hypothèses de gaussianité et d'égalité des variances sont vérifiées.

De plus, nous avons pu constater l'efficacité de chacune des deux méthodes grâce à notre programme effectué à partir de la base de données **Iris**. A travers cet exemple, nous avons pu montrer à plusieurs reprises que c'est l'analyse discriminante quadratique qui a été la plus efficace. En effet, l'ADQ a été plus précise que l'ADL dans la prédiction de l'échantillon choisi, puisque l'hypothèse d'hétéroscédasticité a bien été vérifiée dans l'ADQ, contrairement à l'hypothèse d'homoscédasticité que nous avons admis dans l'ADL.

Enfin, nous n'avons pas eu la possibilité d'étudier et utiliser toutes les techniques d'explication et de prédiction disponibles tel que la régression logistique. Certes, nous aurions pu comparer l'ADL et la régression logistique dans le but de confronter leur performance respectives, mais par manque d'espace dans notre rapport, nous n'avons pas pu étudier cette approche.