1 Principes de l'apprentissage machine

1.1 Modèles utilisés

1.1.1 Analyses discriminantes

Cette étude proposera plusieurs classifieurs s'appuyant sur des méthodes d'analyse discriminante, en particulier l'analyse discriminante linéaire (LDA : *Linear Discriminant Analysis*).

L'analyse discriminante linéaire est une méthode ayant été proposée par Ronald Fisher en 1936[1] pour résoudre des problèmes de classification taxonomique dans le domaine de la botanique.(1,2) La LDA est basée sur la construction de l'hyperplan de projection permettant de maximiser la distance entre les moyennes projetées des différentes classes et de minimiser la variance intraclasse (voir figure 1).(3) La LDA peut être utilisée à fins de classification, mais aussi pour effectuer des réductions de dimensionnalité ou encore afin de faciliter l'interprétation de l'importance de certaines caractéristiques.

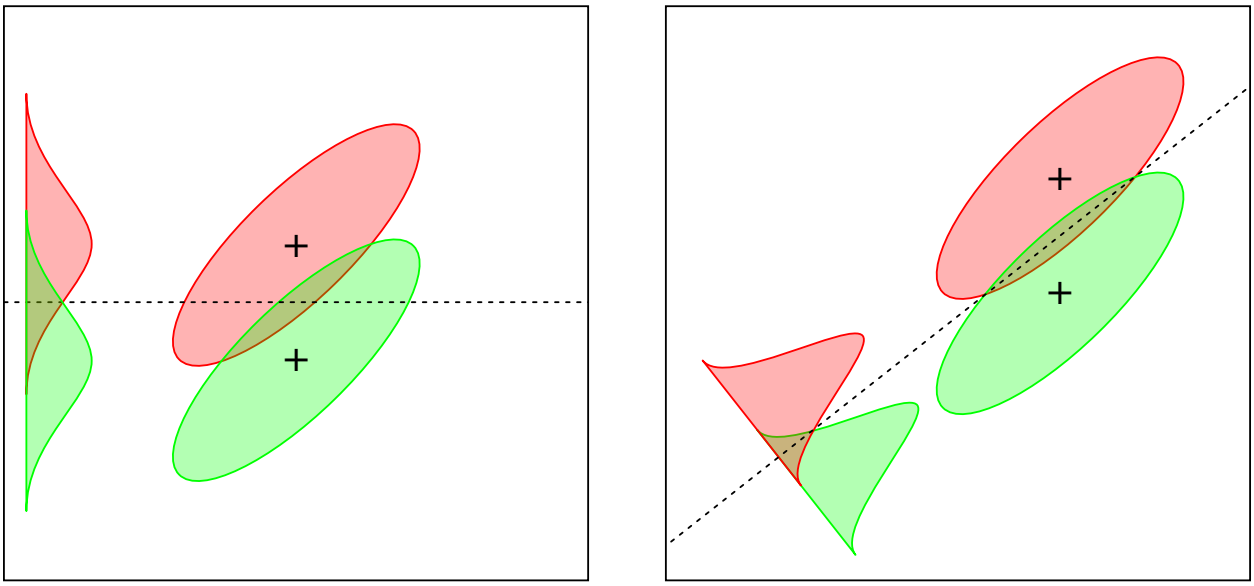

Figure 1: Séparation par distance maximale des moyennes interclasses (à gauche), et par projection sur l'hyperplan optimal tenant compte des variances intraclasses (LDA, à droite)

En pratique, la LDA consiste à construire un indice synthétique, combinaison linéaire des caractéristiques des classes, dont les coefficients permettent de rendre les points du problème originel le plus aisément "séparables". La LDA étant utilisée dans cette étude pour construire un classifieur binaire, c'est ce type de classifieur qui sera présenté dans cette section, et illustré

[1]Cette étude, proposant une méthode de classification des variétés *Iris setosa*, *Iris virginica* et *Iris versicolor* est par ailleurs à l'origine du célèbre jeu de données *Iris*.

avec un exemple extrait du jeu de données *Iris*, dans laquelle nous séparerons les espèces *Iris versicolor* et *Iris setosa*.

Dans ce cadre, la LDA vise ainsi à définir la fonction linéaire en x_i :

$$X = \sum_{i=1}^{n} \lambda_i . x_i$$

avec n le nombre de paramètres caractérisant les individus, x_i les caractéristiques mesurées pour chaque individu et chaque paramètre i, et λ_i des coefficients à optimiser, de sorte que la fonction X maximise le rapport entre les différences des moyennes de chaque classe D et la somme des produits des caractéristiques intraclasses S (proportionnelle à la variance intraclasse), définis par :

$$D = \sum_{i=1}^{n} \lambda_i . d_i \quad \text{avec} \quad d_i = \overline{x_{i,a}} - \overline{x_{i,b}}$$

$\overline{x_{i,a}}$ et $\overline{x_{i,b}}$ étant les moyennes respectives de chaque caractéristique x_i pour les groupes (espèces) a et b, et :

$$S = \sum_{p=1}^{n} \sum_{q=1}^{n} \lambda_p . \lambda_q . S_{pq} \quad \text{avec} \quad S_{pq} = \sum_{i=1}^{n} (x_{p,i} . x_{q,i})$$

$x_{p,i}$ et $x_{q,i}$ étant les caractéristiques mesurées pour les paramètres p et q pour chaque individu i.

L'application sur les espèces *Iris versicolor* et *Iris setosa* nous donne les résultats présentés dans les tables 1 et 2 :

Table 1: Moyennes et différences de moyennes des 4 paramètres d'Iris setosa et versicolor

	Lon.S.	Lar.S.	Lon.P.	Lar.P.
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
difference	-0.930	0.658	-2.798	-1.080

La maximisation du rapport entre les carrés des distances des moyennes interclasses et les variances intraclasses revient à maximiser D^2/S pour chaque coefficient λ_i soit, par dérivation pour chacun des λ_i :

$$\frac{\partial}{\partial \lambda_i} \frac{D^2}{S} = 0 \Leftrightarrow \frac{1}{S} \frac{\partial}{\partial \lambda_i} D^2 + D^2 \frac{\partial}{\partial \lambda_i} \frac{1}{S} = 0 \Leftrightarrow \frac{D}{S^2} \left(2S \frac{\partial D}{\partial \lambda_i} - D \frac{\partial S}{\partial \lambda_i} \right) = 0 \Leftrightarrow \frac{1}{2} \frac{\partial S}{\partial \lambda_i} = \frac{S}{D} \frac{\partial D}{\partial \lambda_i}$$

Table 2: Produits des différences à la moyenne des 4 paramètres d'Iris setosa et versicolor (S_{pq})

	Lon.S.	Lar.S.	Lon.P.	Lar.P.
Lon.S.	19.1434	9.0356	9.7634	3.2394
Lar.S.	9.0356	11.8658	4.6232	2.4746
Lon.P.	9.7634	4.6232	12.2978	3.8794
Lar.P.	3.2394	2.4746	3.8794	2.4604

En supposant que les distributions des classes soient unimodales, cette équation admet une solution unique. Le rapport S/D étant un facteur supposé constant pour tous les coefficients λ_i inconnus, ces coefficients sont donc les solutions du système :

$$\begin{cases} d_1 = S_{11}\lambda_1 + S_{12}\lambda_2 + S_{13}\lambda_3 + S_{14}\lambda_4 \\ d_2 = S_{21}\lambda_1 + S_{22}\lambda_2 + S_{23}\lambda_3 + S_{24}\lambda_4 \\ d_3 = S_{31}\lambda_1 + S_{32}\lambda_2 + S_{33}\lambda_3 + S_{34}\lambda_4 \\ d_4 = S_{41}\lambda_1 + S_{42}\lambda_2 + S_{43}\lambda_3 + S_{44}\lambda_4 \end{cases} \Rightarrow \mathbf{S}.\boldsymbol{\lambda} = \mathbf{D} \Leftrightarrow \boldsymbol{\lambda} = \mathbf{S^{-1}}.\mathbf{D}$$

avec $\mathbf{S}$ la matrice des produits S_{pq}, $\mathbf{D}$ le vecteur des différences des moyennes d_i et $\boldsymbol{\lambda}$ celui des coefficients λ_i.

En indiçant les facteurs :

- $i = 1$ pour la longueur de sépale L_s,
- $i = 2$ pour la largeur de sépale ℓ_s,
- $i = 3$ pour la longueur de pétale L_p,
- $i = 4$ pour la largeur de pétale ℓ_p.

Nous pouvons calculer les coefficients :

$$\begin{cases} \lambda_1 = 0.0311507 \\ \lambda_2 = 0.1839077 \\ \lambda_3 = -0.222104 \\ \lambda_4 = -0.3147364 \end{cases}$$

Soit, après normalisation sur le facteur λ_1 :

$$\begin{cases} \lambda_1 = 1 \\ \lambda_2 = 5.904 \\ \lambda_3 = -7.13 \\ \lambda_4 = -10.104 \end{cases}$$

$$X = L_s + 5.904.\ell_s - 7.13.L_p - 10.104.\ell_p$$

Le seuil de séparation est alors défini par :

$$X_{sep.} = \frac{\overline{X_{ver.}} + \overline{X_{set.}}}{2}$$

Avec $\overline{X_{ver.}}$ et $\overline{X_{set.}}$ les moyennes respectives des X pour *Iris setosa* et *Iris versicolor*.

La valeur absolue des coefficients λ_i calculés précédemment nous indique la pondération de chaque caractère dimensionnel dans l'indice synthétique X permettant d'obtenir une séparation optimale, ainsi que l'illustrent les figures 2 et 3.

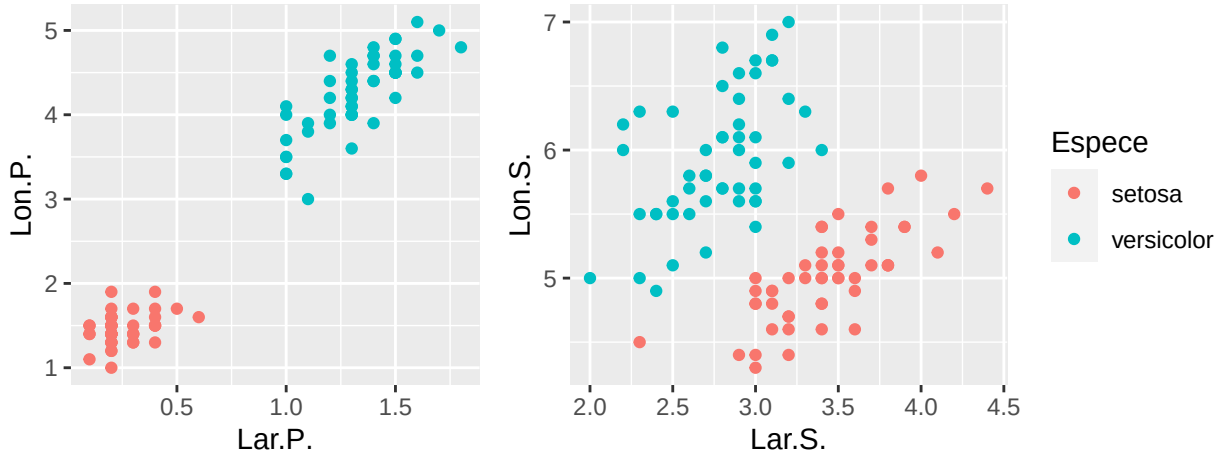

Figure 2: Distribution des variétés setosa et versicolor en fonction de leurs caractéristiques (paramètres fortement pondérés à gauche, faiblement pondérés à droite)

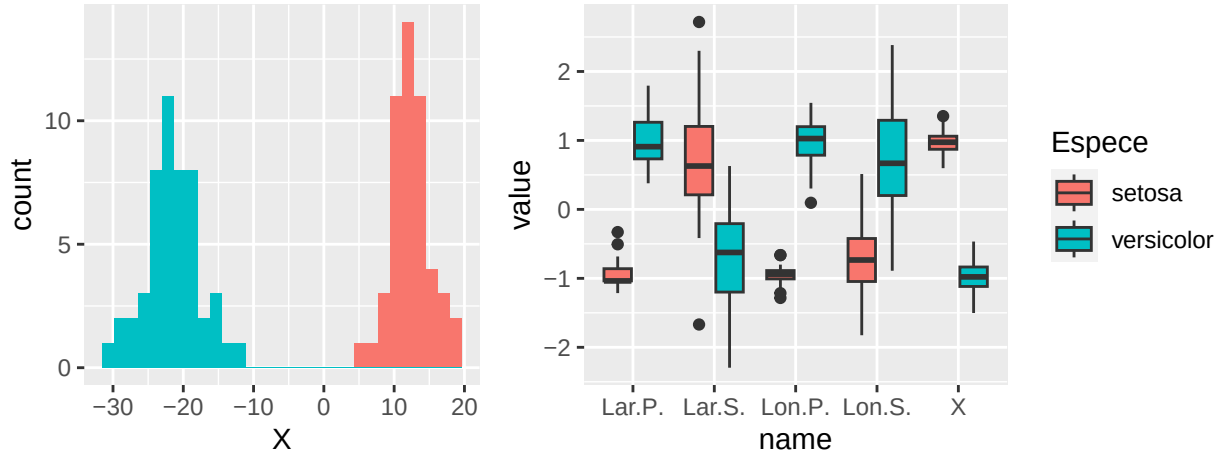

Figure 3: Distribution de X (à gauche) et des paramètres dimensionnels normalisés (à droite) en fonction des espèces

1.1.2 Arbres de décision

etc.

Mini-Bibliographie

1. Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics [Internet]. 1936;7(2):179-88. Disponible sur: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x
2. Anderson TW. R. A. Fisher and Multivariate Analysis. Statistical Science [Internet]. 1996;11(1):20-34. Disponible sur: https://www.jstor.org/stable/2246198
3. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. 2nd éd. Springer; 2016. (Springer Series in Statistics).