Université Paris Descartes

Module : Apprentissage Supervisé

PART I : Données Synthétiques

Réalisé par :

Ait Bachir Samy

I. Introduction

Dans cette partie du projet il s'agit de comparer plusieurs méthodes d'apprentissage supervisé sur 3 base de données différentes. Les différences entre les bases de données sont essentiellement des différences de structure de classes, cependant le nombre de classes et le nombre d'observation varient également d'une base de données à l'autre.

II. Méthodes utilisées

Il existe plusieurs méthodes d'apprentissages supervisé, chacune d'entre elle ayant ses avantages et ses défauts. Dans ce projets les méthodes testées et comparées sont :

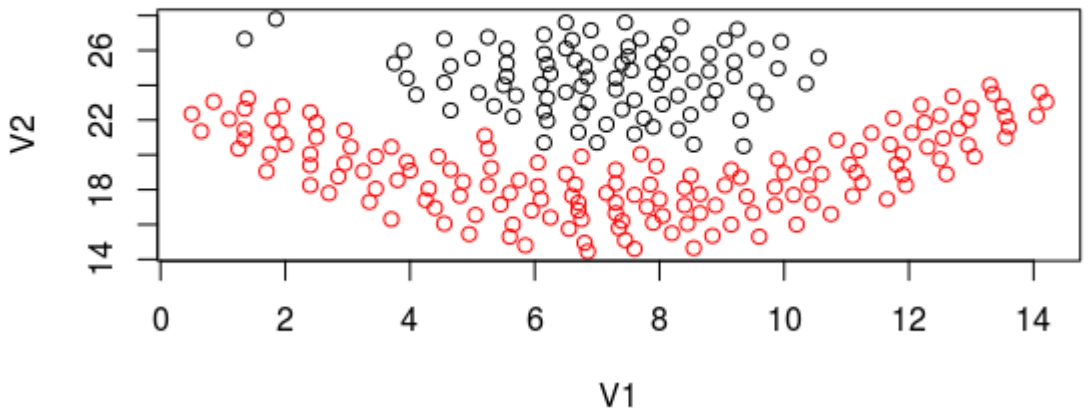
- L'analyse discriminante factorielle, ou LDA.
- Les K plus proches voisins, ou KNN.
- Le classificateur Baysien naïf.
- L'analyse discriminante quadratique, ou QDA.
- Les machines à support vecteur, ou SVM.

III. Base de données

Les bases de données sur lesquels ses méthodes seront utilisées sont au nombre de 3. Dans ce qui suit un bref descriptif de chacune d'entre elles.

III.1 Base de données « flame »

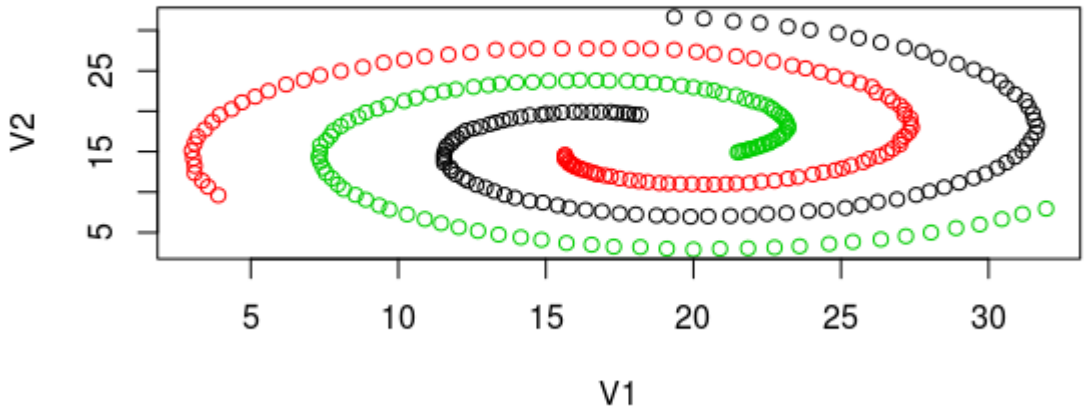
Cette base de données contient 240 individus, décrits par deux variables et divisés en deux classes. On peut représenter cette base données sous forme de nuage de points ou « scatterplot », ce faisant, on obtient la figure suivante :



Sur cette figure on voit bien que la séparation entre les deux classes pas une séparation linéaire, mais plutôt une séparation sous forme de parabole. Il n'y a aucun doute que les méthodes linéaires d'apprentissage ne pourront pas séparer les deux classes.

III.2 Base de données « spiral »

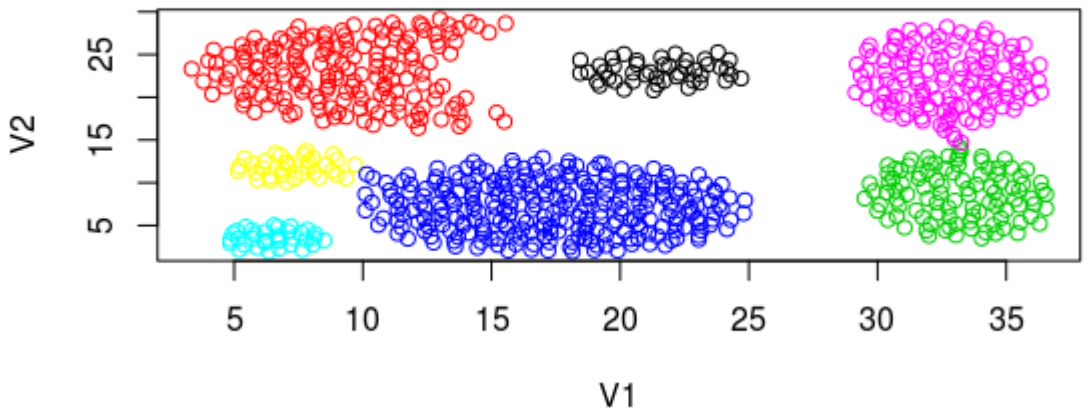
Cette base de données est aussi de dimension 2, mais elle contient un peu plus d'observation. 312 précisément. De plus le nombre de classes, supérieur à celui de la base de données précédente, est 3. De la même manière, pour y voir plus clair, un *scatterplot* des données est établi dans la figure suivante :



Sur cette figure il est visible que la séparation linéaire est complètement impossible, la séparation quadratique elle aussi semble compliquée.

III.3 Base de données « Aggregation »

La dernière base de données utilisée pour ce comparatif est de loin la plus grande. Elle comporte 788 individus décrits par 2 variables et séparés en 7 classes distinctes. Un *scatterplot* est aussi utile ici pour visualiser la structure des différentes classes :



Nous voyons bien 7 classes très bien séparées sauf mise à part quelques rares points ou un doute est de mise. Les séparateurs linéaires devraient pouvoir détecter les classes de cette base de données.

IV. Comparaison des Méthodes

Dans ce chapitre, les méthodes citées précédemment, seront comparées entre elles pour chacune des bases des données et les résultats obtenus représentés sous forme de petits tableaux commentés par la suite.

IV.1 Apprentissage sur « flame »

Les résultats sont obtenus par validation sur un échantillon d'individus différents des individus sur lesquels les algorithmes se sont basés pour créer le modèle. Ici la taille de l'échantillon de test est de 20 individus.

Les paramètres des modèles (Le nombre de plus proches voisins dans KNN, et le type de kernel pour SVM) sont estimés de manière empirique. (boucle à la main pour KNN, et fonction tune de R pour SVM).

Le tableau suivant montre la précision de chacune des méthodes exécutées sur la base de données flamme. :

Méthode	LDA	KNN (K=4)	MBN	QDA	SVM (Radial Gaussien)
Précision	0,25	0,9	0,55	0,55	0,995

Étonnement le modèle quadratique n'arrive pas à bien séparer les classes. Et le SVM obtient les meilleurs résultats. Sans surprise le modèle discriminant linéaire n'arrive pas du tout à séparer les deux classes de la base de données.

IV.2 Apprentissage sur « spiral »

De la même manière que pour la base de données « flame », on sépare les individus utilisés pour le test des individus utilisé pour l'entraînement du modèle (280 pour l'entraînement, 32 pour le test), et obtient les résultats suivants pour celle-ci :

Méthode	LDA	KNN (K=1)	MBN	QDA	SVM (Radial Gaussien)
Précision	0,33	1	0,36	0,33	0,985

Sur cette base de données, 1NN (KNN avec un seul plus proche voisin votant) arrive à une précision de 100 %, cela peut s'expliquer graphiquement. En effet, on voit bien une séparation notable entre individus appartenant à deux classes différentes, le plus proche voisin est donc à chaque fois un individu de la même classe. Cependant si on avait des individus se trouvant quelque part entre les spirales représentant les classes, KNN aurait certainement du mal à déterminer à quel classe ces individus appartiennent, de plus, si de tel individus étaient dans l'ensemble d'apprentissage, cela aurait

pu complètement fausser la phase d'apprentissage du modèle. SVM n'aurait pas eu ce type de problèmes, car il possède une très bonne faculté à généraliser les résultats pour de nouvelles données, et le noyau gaussien est tout à fait adéquat à cette structure en spirale.

Et encore une fois sans surprise, la LDA ne parvient pas du tout à séparer les classes, de même que la QDA ou le modèle bayésien naïf.

IV.3 Apprentissage sur « Aggregation »

Finalement pour la dernière table, l'ensemble d'apprentissage contient 688 individus, et l'ensemble de test contient 100 individus. On obtient les résultats montrés dans le tableau suivant :

Méthode	LDA	KNN (K=1)	MBN	QDA	SVM (Radial Gaussien)
Précision	0,98	1	0,98	0,99	0,9

Comme les classes sont bien distinctes, et linéairement séparables, tous les modèles arrivent à de bon résultats. Cependant, SVM obtient le plus bas des scores de précision. On peut donc en déduire que sur les structures de classes relativement simples, il est préférable d'utiliser des modèles de données linéaires simples.

V. Comparaison et Conclusion

Après avoir étudié les différents résultats des différents modèles sur les trois base de données, on peut dire que les modèles linéaires ne sont utilisables que lorsque les classes sont bien séparées et forment des masses distinctes, dans ces cas-là il est même préférable d'utiliser une LDA.

Cependant, dès que les classes commencent à avoir des formes moins séparable par des droites, ce genre de modèle est très peu fiable, il est mieux dans ces situations de faire confiance à un SVM, même si ce dernier est plus gourmand en temps de calcul, il est plus apte à détecter les classes.

Les modèles quadratiques, bien que fait pour localiser des classes de la forme de la base de données « flame », n'ont pas donné de bons résultats, sauf dans le cas ou déjà la LDA donnait une précision excellente. Il en va de même pour les modèles bayésiens naïfs, pas très satisfaisant dans les cas observés.

Par ailleurs, l'algorithme des K plus proche voisin, donne ici de bons résultats quelle que soit la base de données. Cela est sans doute dû au fait de la bonne séparation de classe en termes de distance euclidienne, ainsi qu'a la basse dimension à laquelle nous l'avons confronté (2 dimensions).

Pour finir, on peut dire qu'il n'existe pas de méthode magique capable de résoudre tous les problèmes de classement, aussi il faut étudier ses données et essayer de voir quelle serait la meilleure solution, non seulement en précision de résultat, en généralisation, mais aussi en temps de construction du modèle.