Université de Bordeaux 1
Master 1 Mention Informatique
2018/2019

Rapport Compte Rendu 1
ACID

22/10/2018

Encadré par:
Pr. Anne Vialard

Réalisé par :
Jellite Oumayma
Bakir Fatima ezzahra

Analyse, classification et indexation des données: feuille 3
Classifieur bayésien

Objectif :

- Dans ce TD on s'appuie sur l'exemple du cours qui consiste à séparer des poissons en deux classes : **bars** et **saumons**. La classification sera faite à partir d'une seule caractéristique (la longueur) dans un premier temps, puis à partir de deux caractéristiques (longueur et brillance). En d'autres termes, le descripteur sera d'abord de dimension 1 puis de dimension 2.
- Puis nous avons suivi la même démarche pour l'exemple d'**Alzheimer** fourni en TD.

I. Exemple: Saumon/Bar

Exercice 1: Vérité terrain

Dans cet exercice, on visualise les histogrammes des longueurs pour les ensembles des poissons(Saumon, Bar).
La figure ci-dessous représente les résultats obtenus :

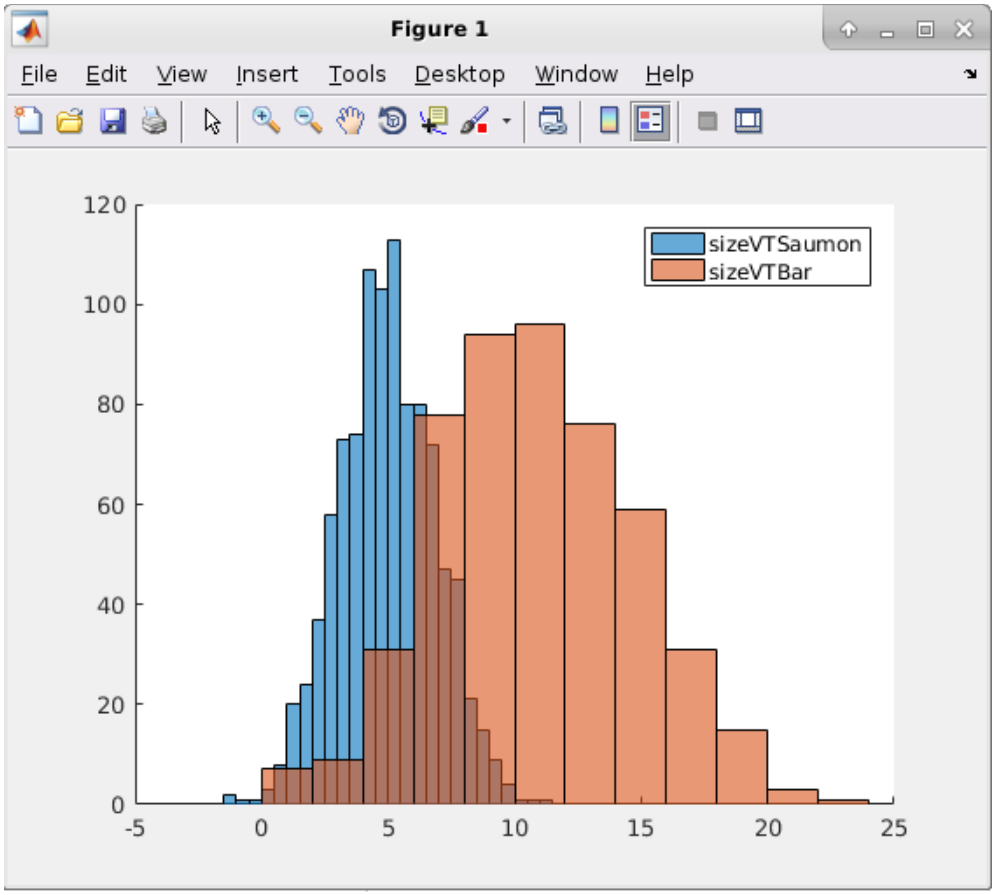

figure1: histogramme vérité terrain

Exercice 2 & 3 : Classifications ML, MAP et Coût

Dans ce compte rendu on va appliquer 3 méthodes de classification :
La classification qui correspond au **Maximum de vraisemblance (ML)** dans un premier temps.
Ensuite l'approche bayésienne basé sur le **Maximum à posteriori (MAP)** sur les données fournis ces derniers comportent deux classes (Saumon/Bar). La classe Saumon contient 1000 valeurs et la classe Bar contient 500 valeurs.
Enfin, on ajoute une classification bayésienne qui prend en compte une fonction de **coût.**

Dans cet exemple nous avons travaillé sur un échantillon de 100 valeurs extraite des données initiales pour les deux classes. Ainsi nous avons calculé l'erreur moyenne total calculé sur les 100 itérations.

Visualisation des échantillons:

«o» Bleus sont les Saumons
«.» Vertes sont les bars

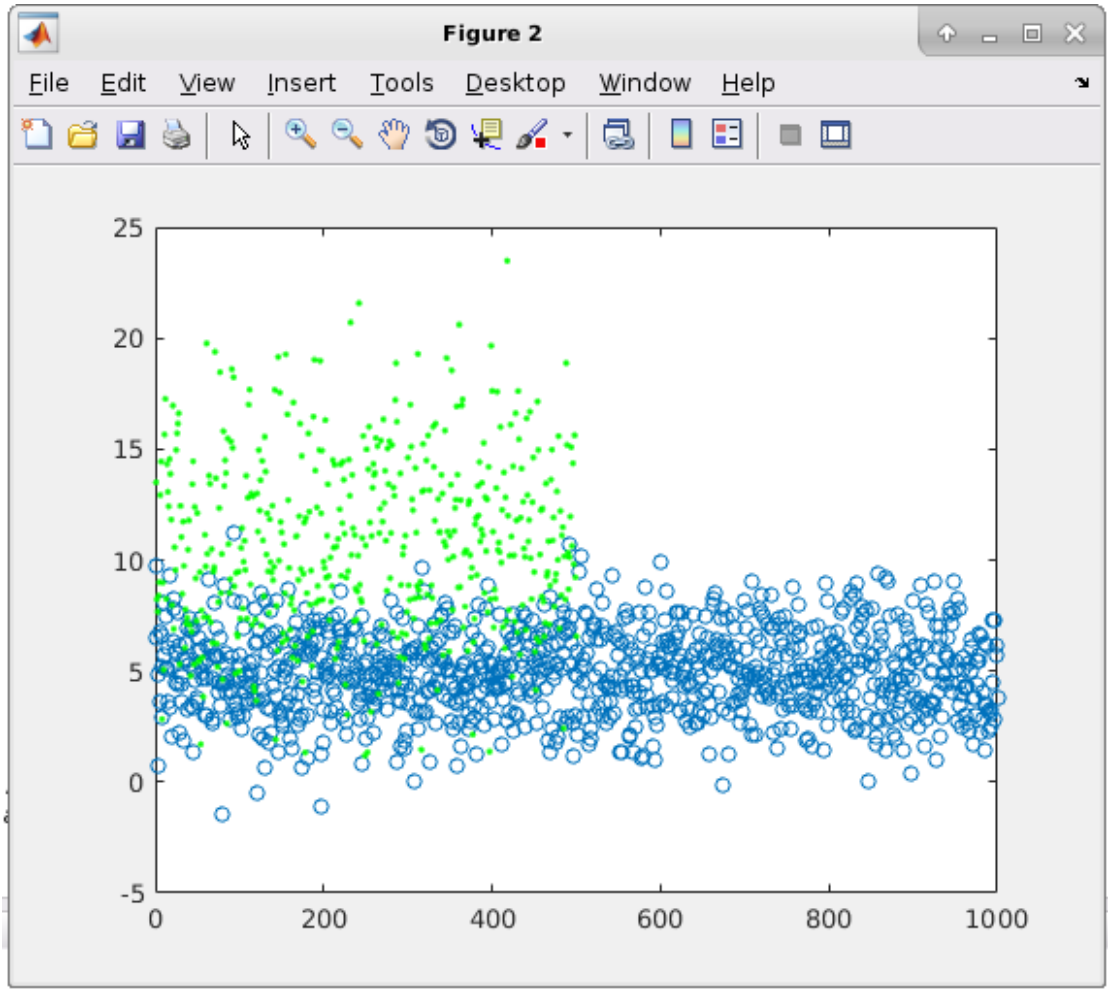

Figure2: Visualisation des échantillons

1) Classifieur ML :

Dans cette première partie, on a essayé d'apparaître les grands lignes fonctionnelles de notre système : Extraction de l'ensemble d'entraînement , entraînement, test et récupération des erreurs. Puis, nous avons découpé notre code en se basant sur ces derniers.
Notre code contient les fonctions suivantes :

- **ExtractTrainAndTest** : qui extrait les données à partir de la vérité terrain donné.
- **TrainModel** : ou on calcule les moyennes et les variances.
- **MyclassifyML**: Cette fonction c'est l la plus intéressante, elle permet de faire la classification en calculant les probabilités de chaque classe et les compare.
- **ComputeError** : cette fonction permet de calculer les erreurs.

On a utilisé une boucle d'évaluation ou on calcule à chaque itération l'erreur pour chaque classe et on la stocke dans un vecteur pour qu'on puisse calculer l'erreur moyenne de chaque classe en sortant de la boucle .
La figure ci-dessous représente la courbe les erreurs en fonction de l'itération pour chaque classe :

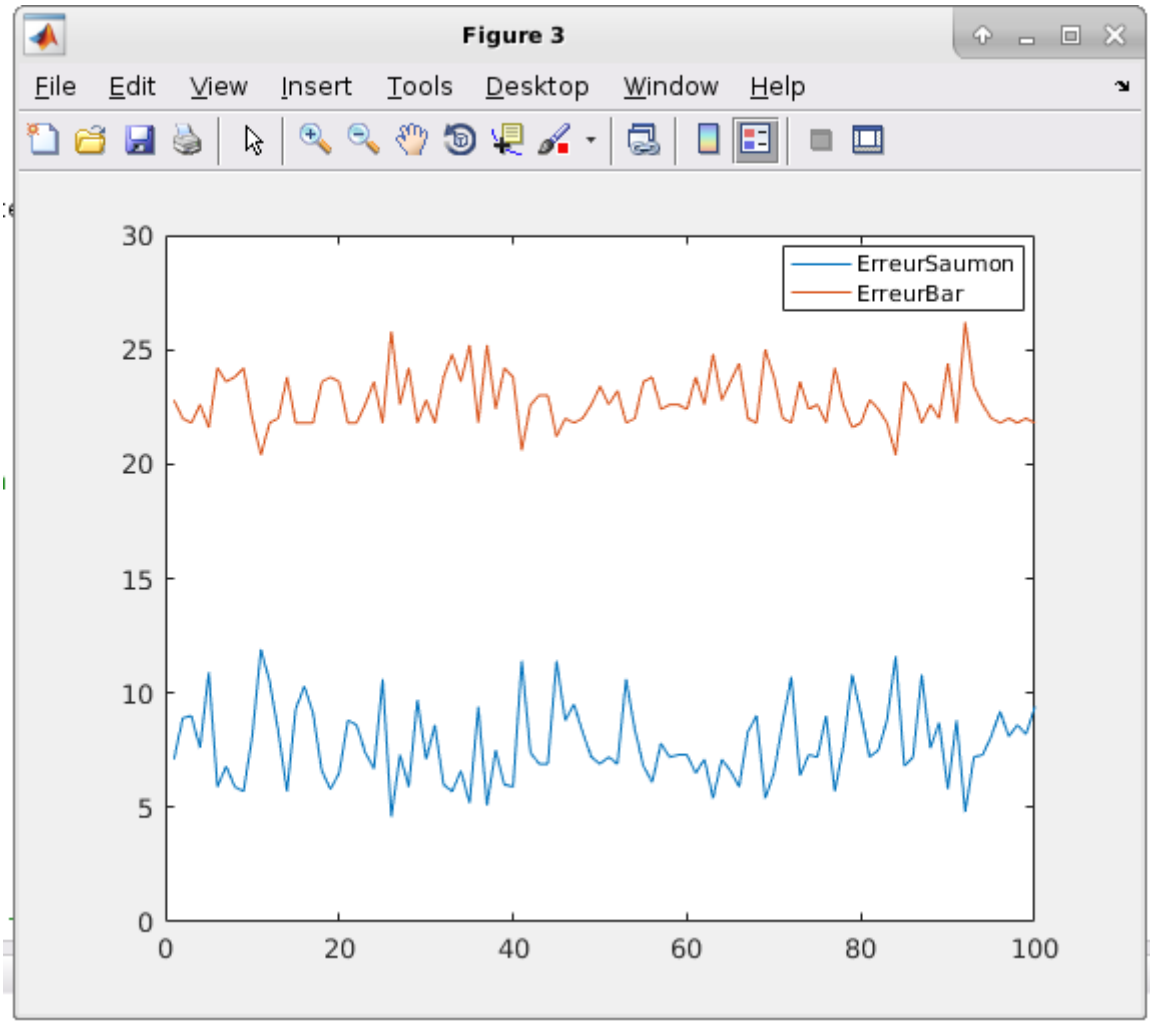

Figure 3 : erreurs ML (1D)

nous avons obtenu une erreur moyenne de 22,81 % pour les Bars et 7.9 % pour les Saumons.

2) Classifieur MAP :

Pour cette classification nous avions qu'à changer la fonction **MyClassify** en multipliant les probabilités obtenues pour les saumons et les bars dans ML respectivement par les valeurs suivantes :

pBar=sizeVTBar/(sizeVTBar+sizeVTSaumon)=1/3
pSaumon=sizeVTSaumon/(sizeVTBar+sizeVTSaumon)=2/3

Après nous avons visualisé les erreurs pour chaque classe en utilisant la même méthode que ML.

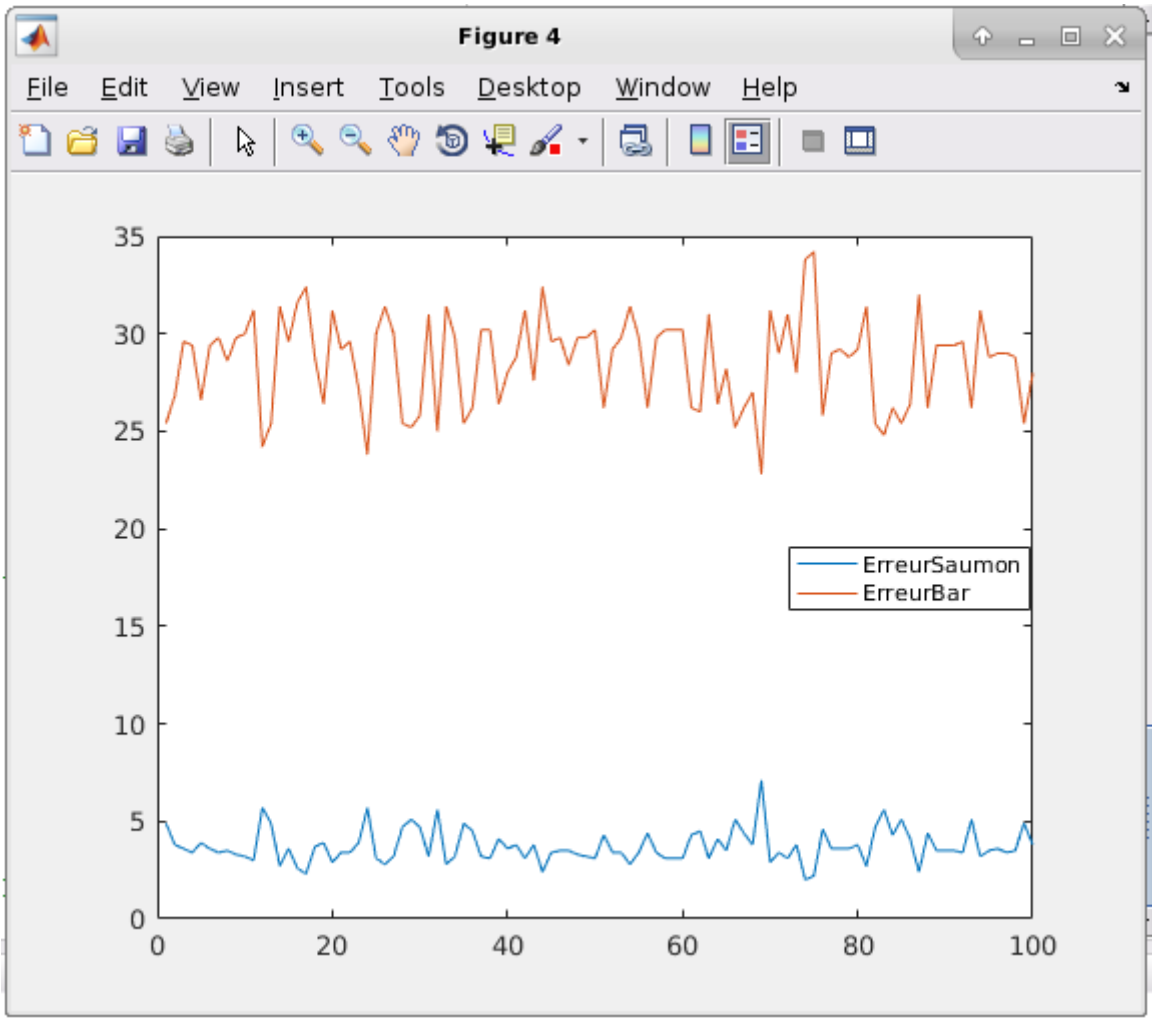

Figure 4: erreurs MAP (1D)

nous avons obtenu une erreur moyenne de 28,59 % pour les Bars et 3 ,73 % pour les Saumons.

On remarque que l'erreur moyenne obtenu par **MAP** se diminue pour les saumons et s'augmente pour les bars cela est justifié par le fait qu'on a multiplié par une probabilité supérieure de la classe Saumon par rapport à la classe Bar.

3) Coût:

Dans cette partie, on a ajouté la notion de coût : le risque de décider" Saumon" alors que l'on a réellement un Bar est de 1, et le risque de décider "Bar" alors que l'on a réellement un Saumon est de 2.
Pour faire cette classification on a modifié notre fonction MyClassifyMAP en multipliant la probabilité des Saumons par le coût 2 et la probabilité des Bars par 1.

Voilà notre visualisation des erreurs :

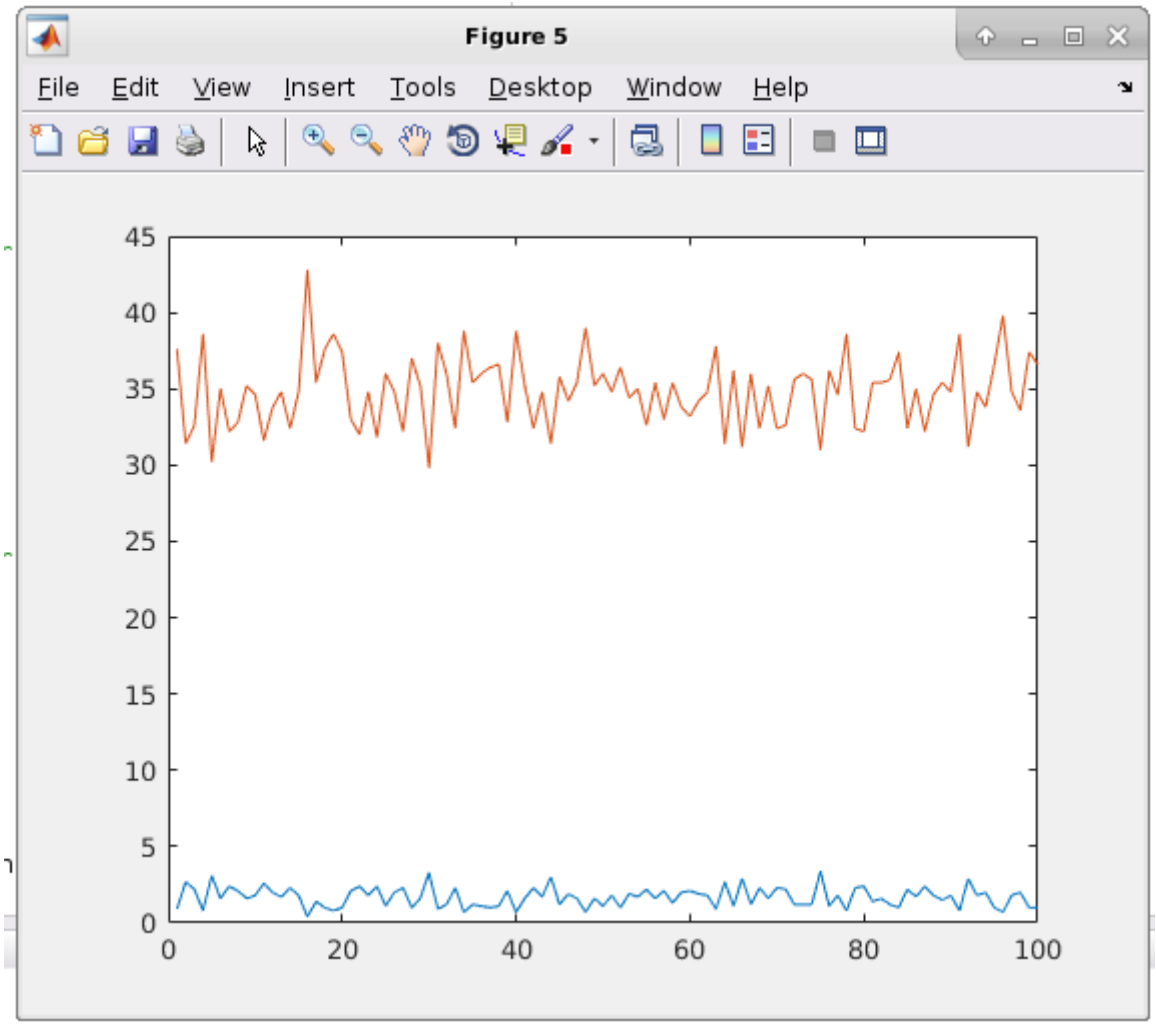

Figure 5: erreurs Coût (1D)

nous avons obtenu une erreur moyenne de 34,82% pour les Bars et 1,68% pour les Saumons.

Conclusion:
le classifieur qui prend en compte la fonction du coût donne des résultats plus précis que le classifieur base sur le maximum a posteriori.

Exercice 4 : Descripteur de dimension quelconque

Dans cette exemple on a deux descripteurs longueur et brillance, c'est pour cela que nous avons effectué quelques modifications au niveau des fonctions suivantes :

ExtractTrainAndTest: On a pris en considération les deux descripteurs fournis dans le fichier «VTSaumonBar2.mat» pour l'extraction des données au niveau des Trains et des Tests.
Par exemple pour les Bars:

```
TrainBar= VTBar(TrainBarIndice,:)
TestBar=VTBar
```

TrainModel: Dans cette fonction, on calcule la matrice de covariance en utilisant la focntion matlab : **COV**

MyClassifyML/MyClassifyCout: Sur ces deux fonctions on remplace la fonction matla**b normpdf** par **mvnpdf.**

On a calculé et affiché les erreurs moyennes des deux classes Saumon/Bar en utilisant les classifieurs **ML**, **MAP** et la classification basée sur la fonction de **coût**.

La visualisation des erreurs se fait de la même manière qu'un seul descripteur, c'est pour cela on va juste afficher ces derniers pour la classification basée sur la fonction de **coût**.

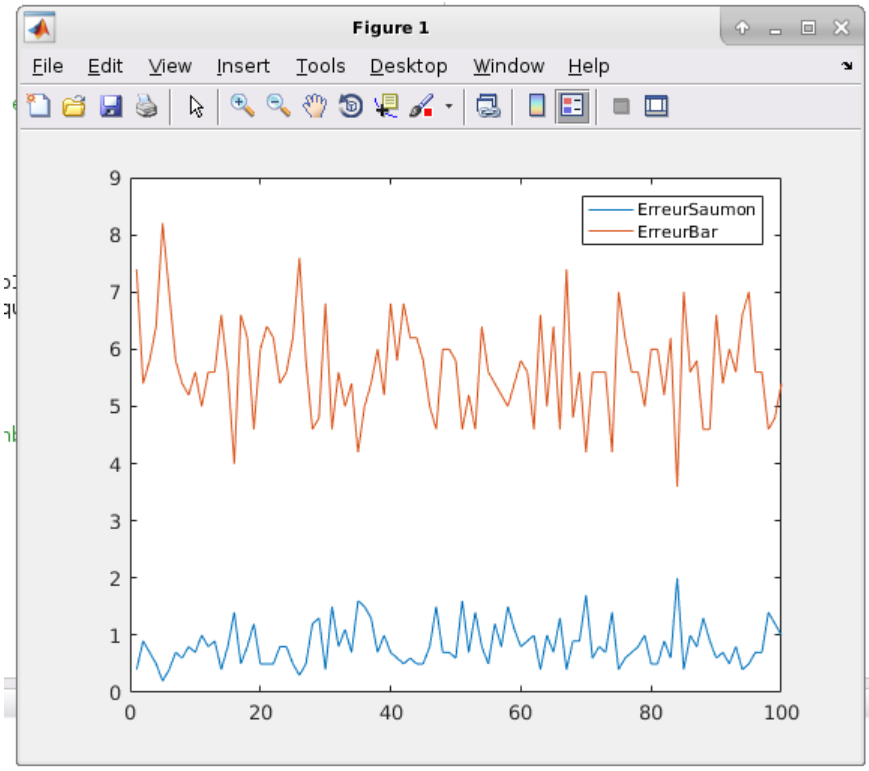

Figure 6 : erreurs Coût (2D)

II. Exemple: Données D'Alzheimer

En suivant la même démarche que l'exemple précédant pour la classification ML, nous avons réussi à afficher les résultats des erreurs pour chaque descripteur **age** et **volume** et pour **les deux** en même temps.

1) Descripteur «Age» :

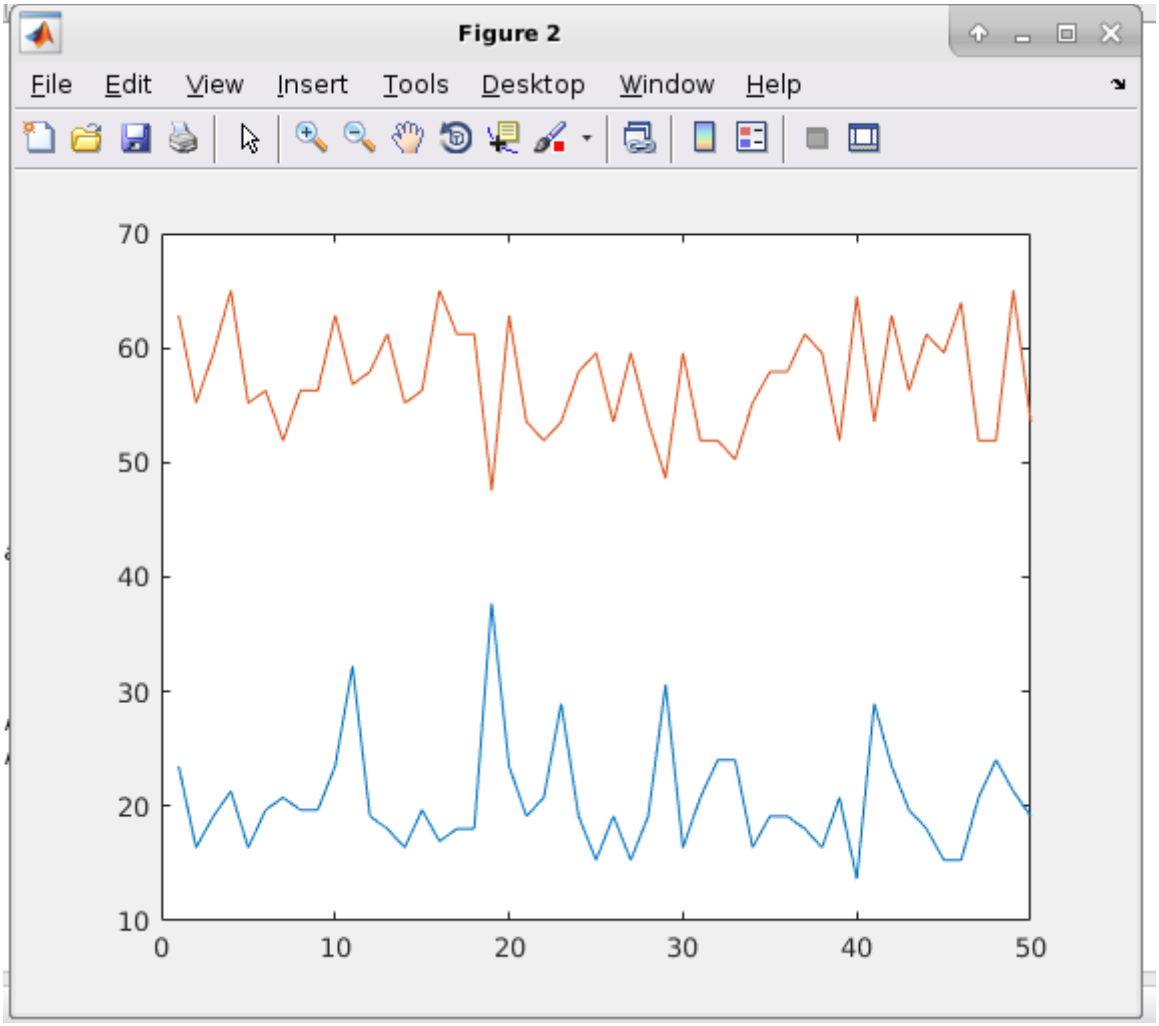

Figure 7: erreurs descripteur Age

nous avons obtenu une erreur moyenne de 57.20% pour les AD et 20.43% pour les CN.

Cela veut dire que les gens les plus agés sont les plus affectés par l'alzheimer et le contrôle de cette maladie devient moins efficace.
Et puisque la valeur moyenne des erreurs pour le groupe de contrôle est petite cela signifie qu'on peut controler la maladie pour les gens qui ont un age inférieure à 40ans.

2) Descripteur «Volume»:

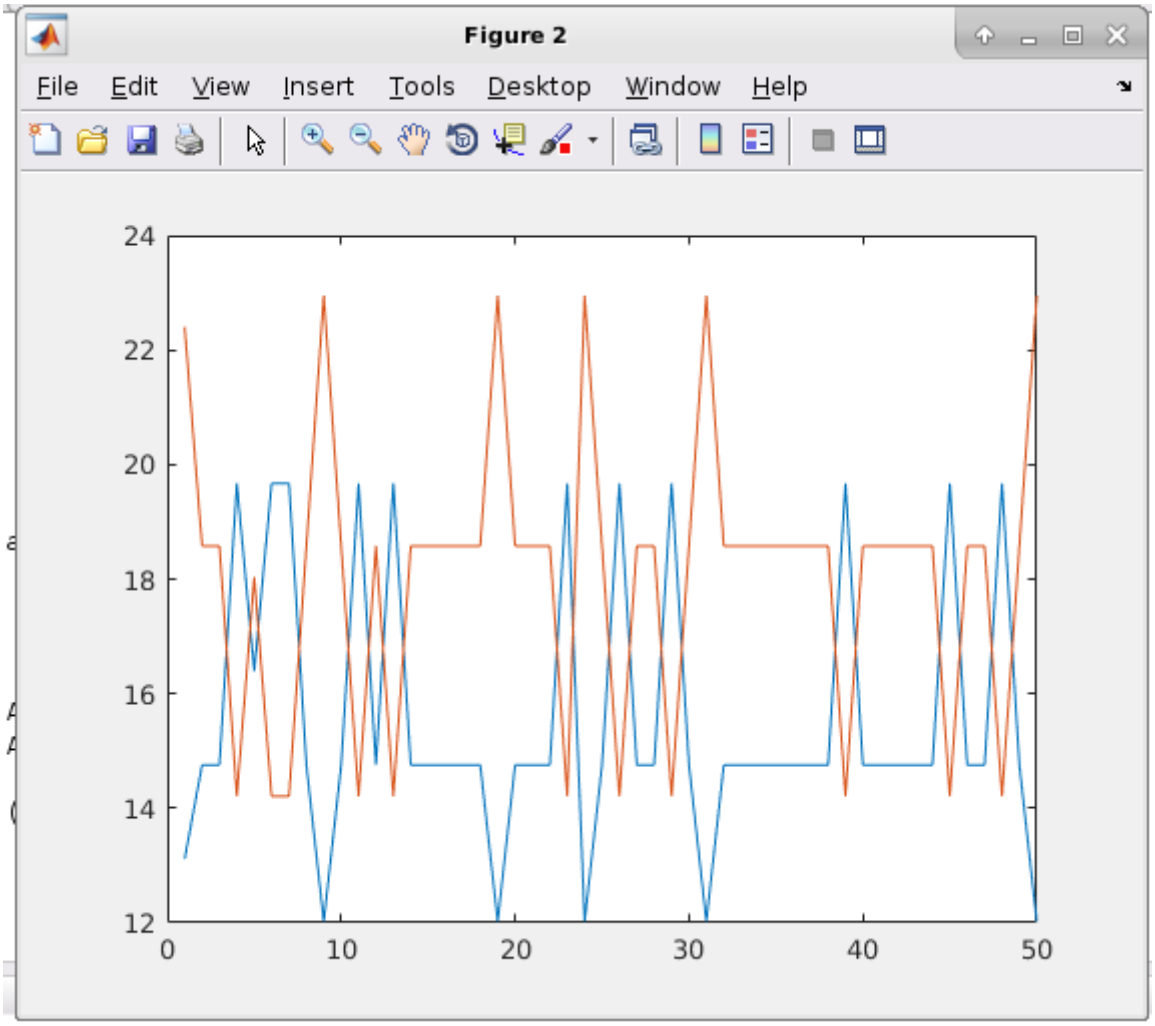

Figure 8: erreurs descripteur Volume

nous avons obtenu une erreur moyenne de 18.12% pour les AD et 15.56% pour les CN.

Par rapport aux résultats obtenus on peut conclure qu'une fois le volume de l'hippocampe est grand on aura plus de chance d'avoir des gens malades et moins de chance pour controler la maladie.

3) Descripteurs «Age et Volume (2D)»:

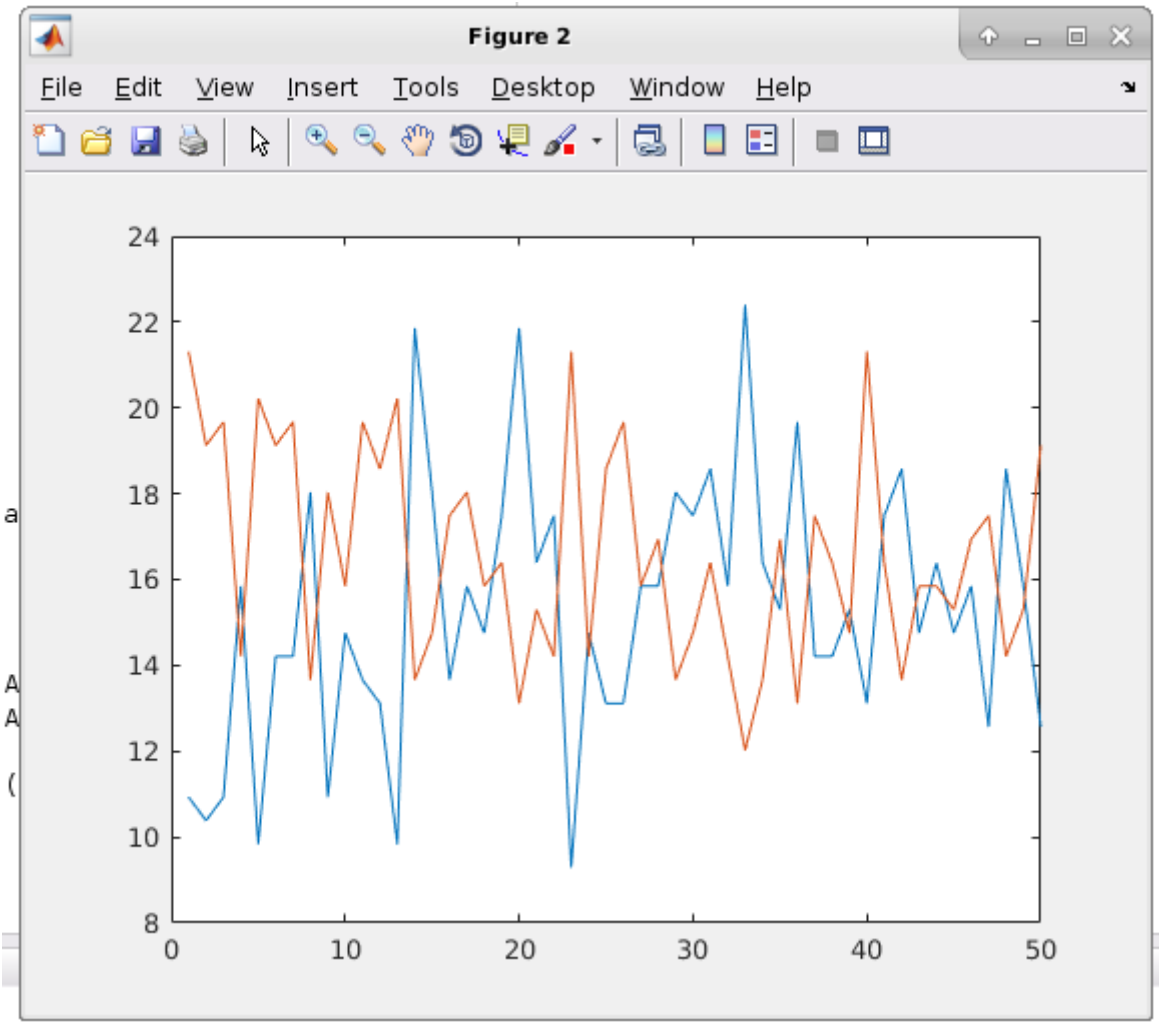

Figure 9: erreurs deux descripteurs (2D)

nous avons obtenu une erreur moyenne de 16.59% pour les AD et 15.27% pour les CN.

Lorsque on voit les résultats par rapport à l'age et le volume de l'hippocampe à la fois on aura plus de chance à controler la maladie si on a des gens ont un age et un volume d'hippocampe petits.