
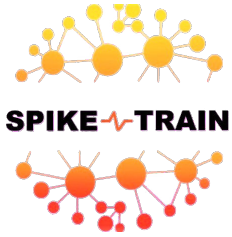


# Rapport de TP

Évaluation des performances d'algorithmes de classification et de réseaux de neurones

Théo FIGINI
L3 Informatique
Année universitaire 2021-22

Syoan ODOUHA
L3 Informatique
Année universitaire 2021-22

Organisme d'accueil : *Université des Antilles*

Enseignant :
Vincent PAGÉ

Dimanche 3 Mars 2022

# Table des matières

# Chapitre 1

# Introduction

La qualité d'un vin est un paramètre qui varie en fonction du goût de tout un chacun. Est-il possible de déterminer la qualité d'un vin de façon plus objective ? Pour cela, nous pouvons utiliser les caractéristiques physiques et chimiques du vin. Tout d'abords, nous devons déterminer la nature du problème sur lequel nous allons travailler, s'il s'agit d'un problème de classification, de régression ou de clustering.

Après avoir déterminé qu'il s'agissait à la fois d'un problème de classification et de régression, nous avons mis en place des réseaux de neurones adaptés à chacun de ces problèmes. Nous avons également mis en place différents algorithmes tels que les **K plus proches voisins** (*K nearest neighbors* en anglais) et les **machines à vecteur de support** (abrégé *SVM* en anglais) et. Nous avons évalué les performances de ces réseaux sur les données de vins blancs, car il s'agit de la base contenant le plus grand nombre d'exemples. Nous avons également évalué les performances en utilisant un nombre réduit de caractéristiques. Et enfin, nous allons conclure ce rapport en présentant ce que nous avons pu tirer de ces expérimentations.

# Chapitre 2

# Travail réalisé

Nous commencerons par utiliser la base de données des vins blancs, car celles-ci contiennent le plus grand nombre d'exemples. Puis nous allons utiliser la base de données des vins rouges qui contient moins d'exemples, afin de vérifier l'impact du nombre d'exemples sur les performances.

Chaque base de données est séparée en deux parties : la partie d'entrainement et la partie de test. Les données ont également été normalisées.

## 2.1 K-nearest neighbors

### Caractéristiques complètes

Le principe des K-nearest neighbors est de classer un objet par rapport aux objets qui sont le plus proche de lui. Appliqué à la classification de vins, un algorithme KNN observera tous les vins ayant des propriétés similaires. Il quelle qualité est attribuée à ces vins similaires afin de déterminer la qualité du vin observé.

Lors des tests, nous avons obtenu une précision de 0.62 sur la base d'apprentissage et une précision de 0.55 sur la base de test en prenant 10 voisins. En diminuant le nombre de voisins à 3, nous obtenons une précision de 0.77 sur la base d'apprentissage et 0.56 sur la base de test. Au contraire, augmenter le nombre de voisins à 20 diminue la précision à 0.58 sur la base d'apprentissage et 0.54 sur la base de test.

Nous constatons qu'augmenter le nombre de voisins diminue les performances en apprentissage. En revanche, la précision en validation ne varie que très peu. L'algorithme KNN reste cependant relativement simple, pour améliorer ces résultats, nous pouvons changer de modèle.

### Caractéristiques limitées

En limitant le nombre de caractéristiques, on obtient une précision de 0.76 sur la base d'apprentissage et 0.51 sur la base de test avec 3 voisins. Avec 10 voisins, nous obtenons une précision de 0.61 sur la base d'apprentissage et 0.53 sur la base de test. Et avec 20 voisins, nous obtenons une précision de 0.59 sur la base d'apprentissage et 0.55 sur la base de test.

Nous constatons que le nombre de voisins a toujours la même influence sur les performances, mais diminuer le nombre de caractéristiques a permis d'améliorer considérablement les performances en apprentissage et de les améliorer légèrement en validation.

## 2.2 SVM

### Caractéristiques complètes

Le principe des SVM consiste à ramener un problème de classification ou de discrimination à un **hyperplan** (*feature space*) dans lequel les données sont séparées en plusieurs classes dont la frontière est la plus éloignée possible des points de données (ou "marge maximale"). D'où l'autre nom attribué aux SVM : les séparateurs à vaste marge. Le concept de frontière implique que les données soient linéairement séparables. Pour y parvenir, les support vector machines font appel à des noyaux, c'est-à-dire des fonctions mathématiques permettant de projeter et séparer les données dans l'espace vectoriel, les "vecteurs de support" étant les données les plus proches de la frontière. C'est la frontière la plus éloignée de tous les points d'entraînement qui est optimale, et qui présente donc la meilleure capacité de généralisation.

Nous avons obtenu une précision de 0.54 en apprentissage et 0.53 en validation. Les performances obtenues sont légèrement inférieures à celles l'algorithme KNN, cela est peut-être dû au grand nombre de caractéristiques. Nous retesterons cet algorithme avec un nombre inférieur de caractéristiques.

### Caractéristiques limitées

En diminuant le nombre de caractéristiques, nous obtenons une précision de 0.53 sur la base d'apprentissage et 0.53 sur la base de test. Ce résultat est quasi identique à celui obtenu avec les données complètes, le nombre de caractéristiques n'a donc aucun impact sur les performances des SVM.

## 2.3 Réseau de neurones

### Classification

Les réseaux de neurones sont une technique d'apprentissage supervisé pouvant répondre à différents problèmes tels que la classification ou la régression.

Tous les entrainements ont été effectués sur 300 époques avec un modèle composé de 2 couches cachées de 12 neurones chacune.

#### Caractéristiques complètes

Avec en utilisant toutes les caractéristiques, nous obtenons une précision de 0.6 en apprentissage et 0.55 en validation (*Fig.* 1) avec une perte de 1.0 en apprentissage et 1.1 en validation (*Fig.* 2). Ces résultats sont assez moyens, mais ne semble pas indiquer un sur-apprentissage.

#### Caractéristiques limitées

En réduisant le nombre de caractéristiques, nous obtenons des performances similaires, nous obtenons une précision de 0.58 en apprentissage et 0.55 en validation (*Fig.* 3) avec une perte de 1.0 en apprentissage et 1.1 en validation (*Fig.* 4). Cela montre qu'il y a eu un sous-apprentissage.

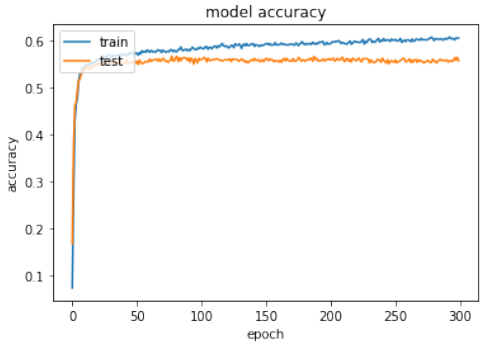


FIGURE 1 – **Précision du réseau avec toutes les caractéristiques**.

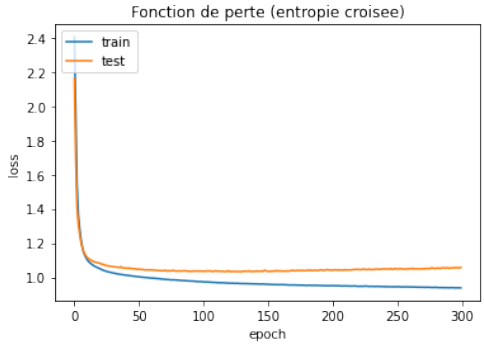


FIGURE 2 – **Fonction de perte du réseau avec toutes les caractéristiques**.

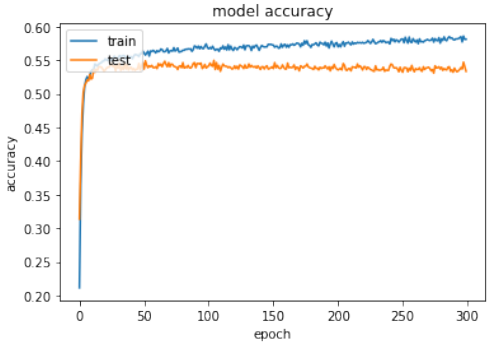


FIGURE 3 – **Précision du réseau avec toutes les caractéristiques**.

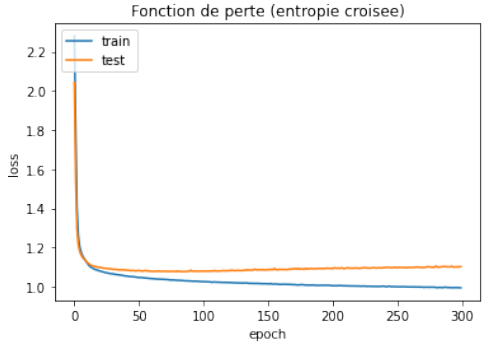


FIGURE 4 – **Fonction de perte du réseau avec toutes les caractéristiques**.

## Régression

En utilisant l'ensemble des caractéristiques, nous obtenons Erreur Absolue Moyenne (EAM) de 0.5 en apprentissage et en validation. Nous obtenons les mêmes résultats en utilisant moins de caractéristiques. Ces résultats indiquent clairement qu'il y a une erreur dans le modèle.

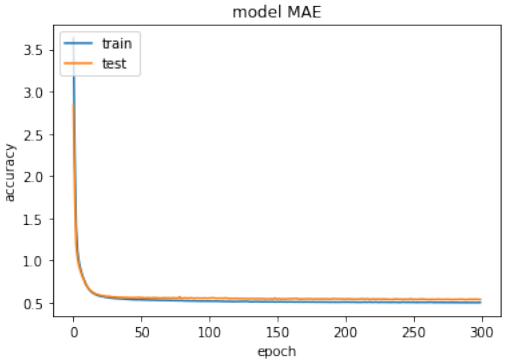


FIGURE 5 – **MAE du réseau avec toutes les caractéristiques.**

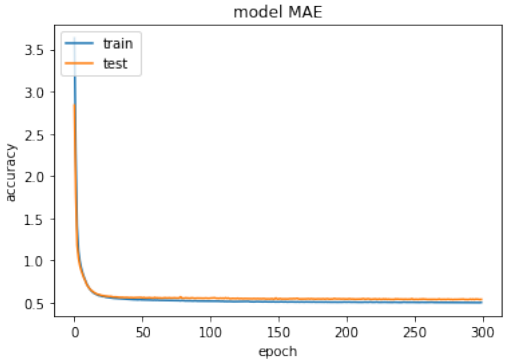


FIGURE 6 – **MAE du réseau avec un nombre de caractéristiques limité.**

# Chapitre 3

# Conclusion

Au cours de ces travaux pratiques, nous avons étudié les différents algorithmes de classification et de régression. Nous avons pu en déduire la régression n'était pas une méthode adaptée pour traiter ce problème. Les méthodes ayant eu les meilleures performances en apprentissage sont le KNN et les SVM, en revanche, le réseau de neurones pour la classification a obtenu les meilleures performances en validation.

Afin d'améliorer ces résultats, nous aurions pu essayer de choisir des caractéristiques plus pertinentes ou améliorer le traitement de nos données. Nous pouvions également essayer de complexifier notre modèle en augmentant le nombre de couches ou le nombres de neurones par couches.