Data Science et IA (3I026)

Fumeron Pierre et Taoussi Redouane

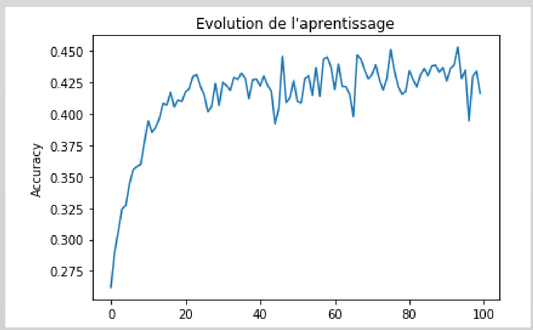
Problématique 1

Description: les films sont notés sur 5. Serait-il possible de prédire cette note en utilisant les catégories des films? Peut on améliorer notre algorithme en ajoutant des données: la moyenne du pire acteur et du meilleur acteur?

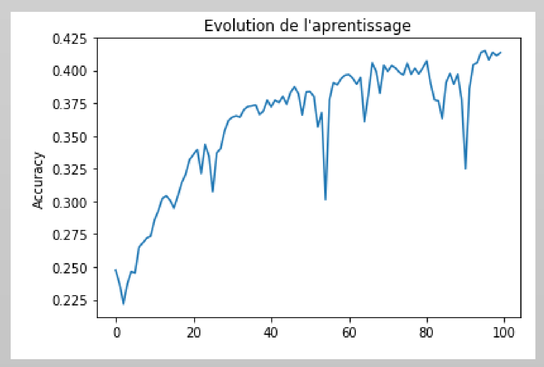
Méthode: Régression supervise grâce à l'algorithme du Moindres Carre vu en cours.

Résultats:

- Evolution de la précision en fonction du nombre de données testées:



- Même évolution après ajout des notes:



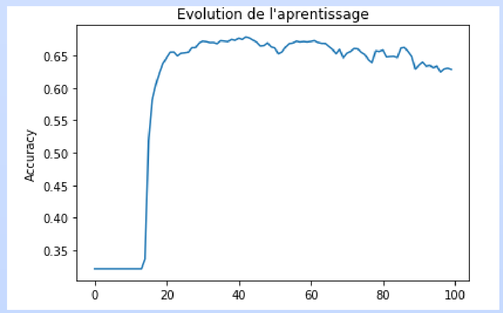
Problématique 2

Description: Serait-il possible à partir de données de prédire si un acteur est un homme ou une femme?

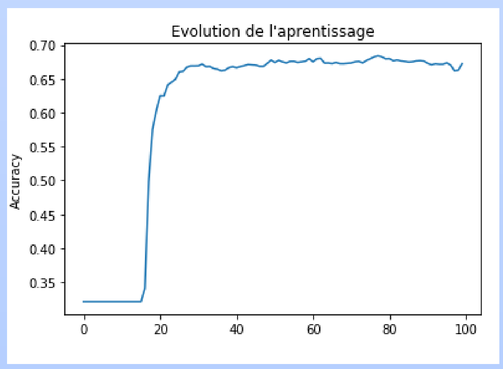
Méthode: Classification supervisée grâce à l'algorithme du gradient stochastique vu en cours.

Résultats:

- Comme pour la régression, nous affichons l'évolution de la précision:



- Et la même évolution après ajout de données:



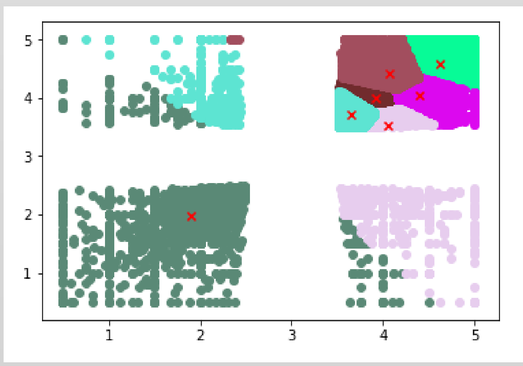
Problématique 3

Description: Peut-on trouver des similarités parmi les utilisateurs qui ont notés les films?

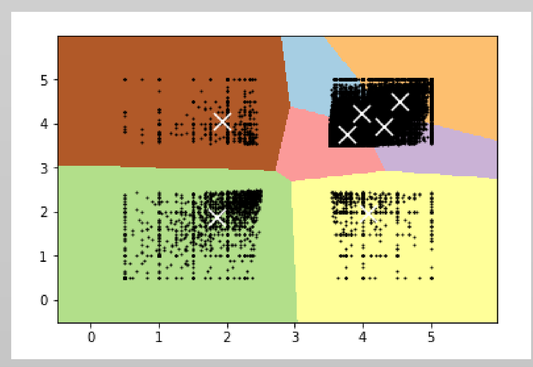
Méthode: Pour cela nous avons utilisé la méthode de clustering non supervisé grâce a l'algorithme de Kmeans .

Résultats:

- Nous utilisons premièrement l'algorithme que nous avons implanté pendant le tme 8:



- Puis nous avons utilise une librairie sklearn pour comparer l'efficacité de notre code:



Problématique 4

Description : Utilisation des résumés de films pour prédire la catégorie d'un film

Méthode : Nous avons utilisé un classifier bayésien qui se base sur la fréquence d'occurrence des mots dans le résumé d'un pour prédire sa catégorie. Pour un résumé ayant k mots, on renvoie la catégorie qui maximise le calcul suivant :

$$\frac{P(mot_1|categorie_i) \times P(mot_2|categorie_i) \times \cdots \times P(mot_k|categorie_i) \times P(categorie_i)}{P(mot_1) \times P(mot_2) \times \cdots \times P(mot_k)}$$

Prétraitement des donnés : Les résumés ont été traités en retirant la ponctuation, les mots vides, certains suffixes afin d'indexer uniquement des mots "intéressants" en minuscule.

Résultats : Plus la base d'apprentissage est importante, plus le nombre de mots indexés est grand et meilleure est la précision de l'algorithme sur la base de test.