Analyse de données de films

Hakim CHEKIROU

Introduction:
Pour analyser et extraire des informations pertinentes des données fournies, nous prétraitons d'abord les données. Pour decider quelles informations utiliser , nous analysons statistiquement les données avant d'appliquer les algorithmes d'apprentissage.

PRE-TRAITEMENT:

> Reduction de la taille de Ratings

> Supression des attributs non significatifs dans toutes les tables.

> Supression des valeurs nulles.

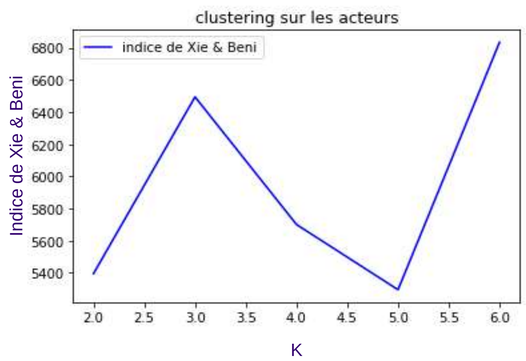
> Rajout de nouvelles colonnes a partir des données brutes.

> **167** Relier les differentes tables entres eux en utilisant la base Links.

Aprentissage non supervisé : K-moyennes

Nous voulons ici trouver des groupes d'acteurs en se basant sur le type de films dans lesquels ils jouent et leurs performances dans ceux ci. Les roles ont été groupés par acteur et de nouvelles colonnes ont été rajoutées comme la qualité moyenne du film par acteur. Du fait du temps de calcul important nous nous somme limité a **10000 acteurs**

Recherche du meilleur nombre de clusters



Selon l'indice Xie & Beni, les Valeurs de K les plus intéressantes sont 2 et 5

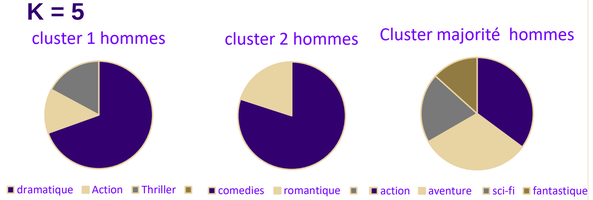
Analyse des clusters trouvés

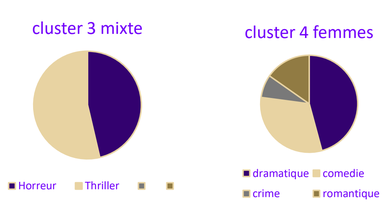
K = 2

Cluster 1 composé d'hommes

Cluster 2 composé uniquement de femmes.

K = 5

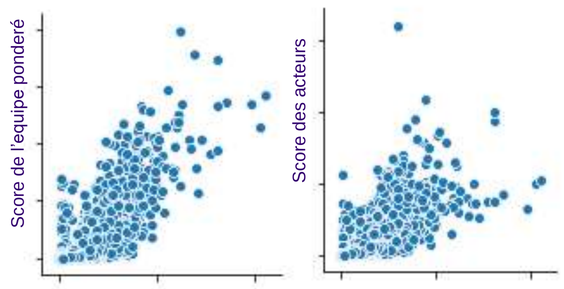




Regression

Dans cette section on cherche a prédire le score d'un film.

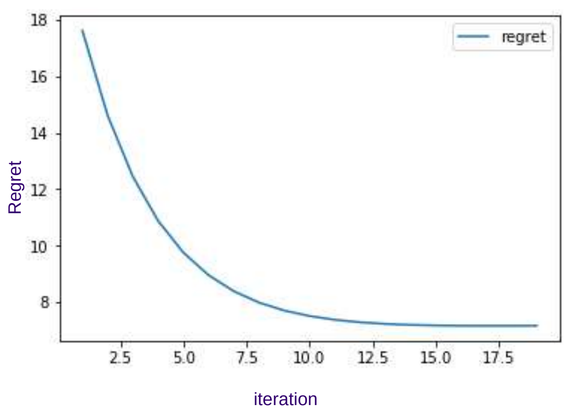
Correlation



Perceptron batch

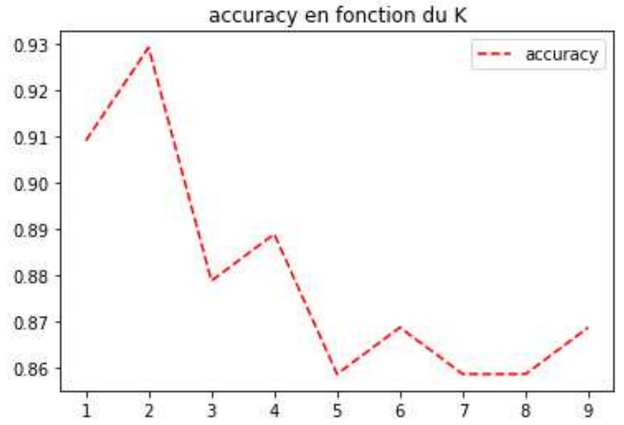
Nous avons entrainé l'algorithme sur 25000 films, avec un taux d'apprentissage de 0,002

Regret en fonction des iterations



Classification Supervisée

Notre but ici est de prédire le type du film.



Arbre de decision

Classifieur qui predit si la personne est un realisateur ou non.
Epsilon = 0.01,