Machine learning et scoring à valeurs discrètes

F. Baschenis
sujet proposé par N. Vayatis

École Normale Supérieure de Cachan, département de mathématiques

24 Juin 2010

Bref rappel du problème

- Un film $\longleftrightarrow$ un élément $X \in \mathbb{R}^d$
  Une note $\longleftrightarrow$ un élément $Y \in \{1, 2, 3, 4, 5\}$
  Les données $\longleftrightarrow$ (X_i, Y_i)
- Objectif : Donner une fonction de scoring s sur l'ensemble des $X \in \mathbb{R}^d$ qui les classe : $s(X_1) > s(X_2) \iff Y_1$ est "probablement" plus grand que Y_2

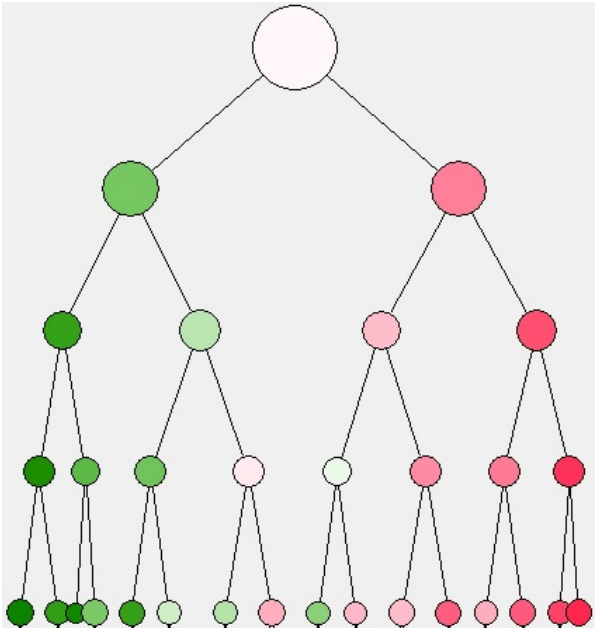
TreeRank (1)

- Fonctionne dans le cas binaire
- méthode de Machine Learning
- Maximise l'AUC (Aera Under Curve)
- implémenté dans R

TreeRank (2)

- Donne un arbre de tri
- On déduit la fonction de score

$$s_D(X) = \sum_{k=0}^{2^D-1} (2^D - k)\mathbb{I}\{X \in C_{D,k}\}$$



Le Discounted Cumulated Gain (DCG)

- Le DCG rend compte :
  - de l'ordre sur les X_i
  - de l'ordre sur les Y_i
  - donner plus d'importance aux éléments bien classés
- En réorganisant les X_i suivant l'ordre induit par s, il est défini par

$$DCG_s(X_i) = \begin{cases} Y_i & \text{si} \quad i = 1 \\ DCG_s(X_{i-1}) + \frac{Y_i}{\log(i)} & \text{sinon} \end{cases}$$

Plan

Comment utiliser le TreeRanking ?

On utilise une technique de régression ordinale.

- On réalise plusieurs fois l'algorithme en changeant les valeurs des labels
- Pour $i \in 1, 2, 3, 4$ on change les valeurs appartenant à $\{1, \cdots, i\}$ par -1, les autres par 1 et on applique le TreeRanking
- On obtient donc 4 fonctions de scoring s_i, i étant le plus grand nombre "rejeté"

Comment utiliser le DCG ?

- s_4 fournit uniquement des informations sur les données de label 5, s_1 sur celles de label 5
- On privilégie donc s_4 par rapport à s_1
- On a une fonction de score

$$s(X) = \sum_{i=1}^{4} \frac{s_i(X)}{\log(6-i)} = s_4(X) + \frac{s_3(X)}{\log(3)} + \frac{s_2(X)}{\log(4)} + \frac{s_1(X)}{\log(5)}$$

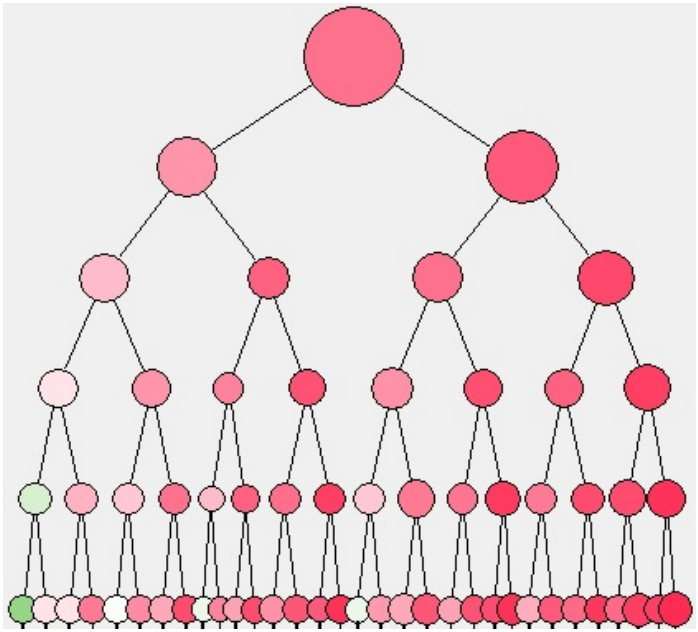
Plan

Quel jeu de données ?

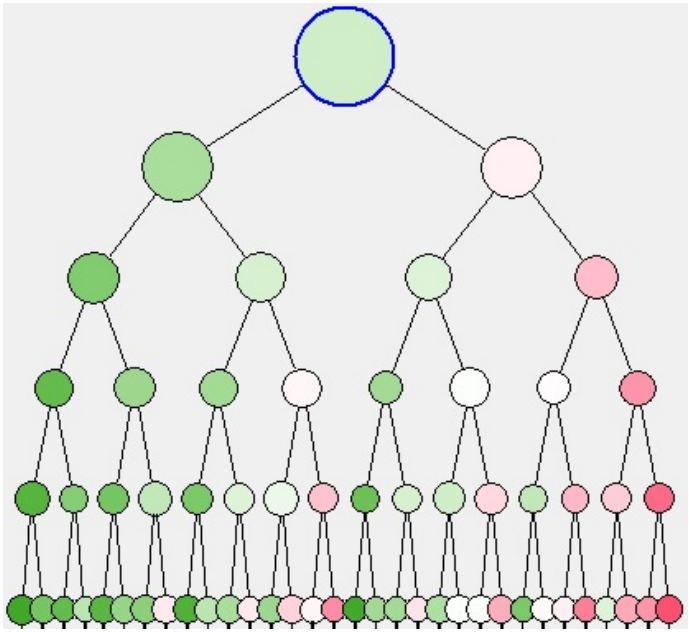
- Movielens : database de 100 000 notations (en fait moins)
- On a mélangé les caractéristiques des films et des utilisateurs
- Un film $\longrightarrow$ un vecteur de booléens
- Un utilisateur $\longrightarrow$ tableaux, avec commes colonnes l'age, le sexe et le métier

Résultats (1)

s_4

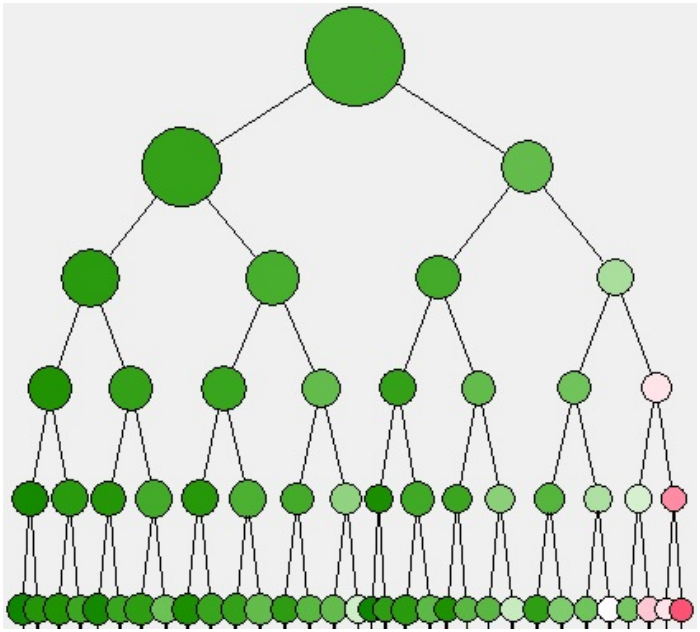


s_3

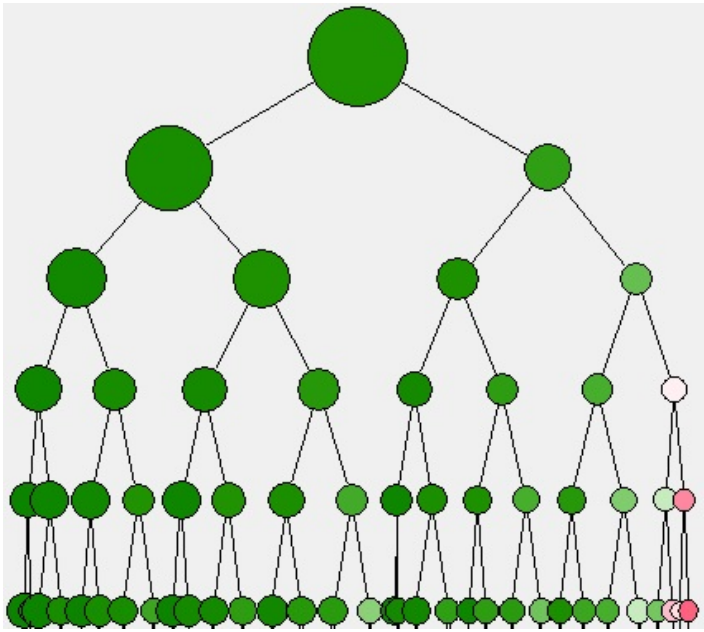


Résultats (2)

s_2



s_1



Plan

Combiner les arbres

- Combiner les 4 Arbres vus précédements, en un arbre correspondant à la fonction de scoring globale s
- Problème : risque d'explosion combinatoire : taille des arbres à la puissance 4 ...

Evaluation des performances

- Perfomance de l'algorithme
- Comparaison des fonctions de scoring avec un autre algorithme
- Problème : peu d'algorithmes traitent de l'ordonnancement discret.

Conclusion

Des Questions ?

Bibliographie

Andrew P. Bradley.
The use of the area under the roc curve in the evaluation of machine learnin algorithms.
Pattern Recognition, 1997.

Pinar Donmez and Jaime G. Carbonell.

Kalervo Järvelin and Jaana Kekäläinen.
Ir evaluation methods for retrieving highly relevant documents.

In *SIGIR '00 : Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.

Nicolas Vayatis StÃ©phan ClÃ©menÃ§on.
Tree-based ranking methods.