

Projet M1: “Prédire qui gagne le prix pour le meilleur défenseur de la NBA” — apprentissage et évaluation d’un modèle prédictif

18 avril 2024

1 La sélection d’attributs

Quand vous avez comparé les classements d’une période/d’une année à une autre pour chaque méthode, vous avez identifié pour quel nombre d’attributs $k \in \{10, 20, 30, 40, 50\}$ ces classements restent relativement stables.

Par conséquent, on va faire des expériences avec chaque méthode de sélection, en exploitant cette valeur k trouvée pour une méthode, ainsi que les valeurs p calculées pour Pearson’s.

2 Préparer les jeux de training

Nous allons utiliser deux types de jeux de training :

- Les 5 à 9 années avant l’année à prédire, p.ex. on prédit 1998 après avoir entraîné le modèle sur 1993-1997, 1999 à partir de 1993-1998 etc.
- Toutes les années précédentes, p.ex. 2010 à partir de 1993-2009

3 Préparer les données

Pour un jeu de training, on calcule le classement de Pearson’s et on sélectionne tous les attributs pour lesquels la valeur $p \leq 0.05$, ça nous donne un nombre k' d’attributs. Puis, on sélectionne les k' meilleurs attributs selon Mutual information et selon Ridge pour le jeu de training et on garde **les mêmes attributs pour le jeu de test**.

On fait la même procédure pour la valeur k “stable” identifiée pour une certaine méthode de sélection d’attributs.

Enfin, pour les méthodes “From model”, on utilise tous les attributs sélectionnés.

4 Évaluation

On utilise chaque année comme jeu de test, sauf les années 1993-1997.

Pour la prédiction elle-même, on ne s’intéresse pas trop à la valeur numérique, mais seulement au classement des joueurs. Donc, p.ex. si les vraies valeurs pour une année donnent

1. Joueur₁ : 400

2. Joueur₂ : 200

3. Joueur₃ : 100

et notre modèle prédit

1. Joueur₁ : 600

2. Joueur₂ : 90

3. Joueur₃ : 100

on considère que le modèle a correctement prédit le gagnant, mais incorrectement le deuxième et troisième, même si la valeur pour Joueur₃ est correcte.

Donc, pour chaque année, on note si

- le modèle a correctement prédit le gagnant,
- quel pourcentage et quel kendall's tau on a avec le top-3
- quel pourcentage et quel kendall's tau on a avec le top-10