Structure du projet

S. Perochkin

21 août 2017

[VERSION PRÉLIMINAIRE]

[À modifier après le brainstorm – vendredi ou avant]

Aperçu du projet

Il est difficile de figer le projet lorsqu'il est à peine entrepris, mais on peut quand même dresser son allure générale. Je crois qu'on veut produire un espèce de *hub* où trouver :

— des outils pour analyser/visualiser des données de tennis ; et
— de l'information sur les tendances dans le tennis.

À l'étape où nous sommes, le plus essentiel est un de faire un brainstorm sur ce que nous pensons du projet justement... Il y a toutefois des tâches qui se présente comme incontournables peu importe la direction du projet. En voici quelques-unes auxquelles nous ajouterons le résultat du brainstorm (voir la Table des matières). La dernière section du document concerne ce sur quoi on devrait ce concentrer en fin de semaine.

Table des matières

1 Base de données

1.1 Collecte

Les plus grandes questions tournent autour des sources de données et des scripts pour aller les chercher. Nous avons déjà quelques outils pour ces tâches.

1.2 Nettoyage et entreposage

Comment arranger tout ça ? Quelles nouvelles variables veut-on de déjà calculées ? etc.

2 Visualisation

On aura besoin de plusieurs outils de visualisation, autant pour faire le portrait d'un joueur et comparer deux ou plusieurs joueurs que pour représenter un certain tournoi (son arbre). En analysant/modélisant les données, on verra les outils qui nous serait utile pour travailler et nous ferait gagner du temps. Ces outils peuvent s'avérer intéressant pour mettre sur un site/une app plus tard.

2.1 Arbre d'un tournoi

Un incontournable.

2.2 Interface aux données

Seulement si on planifie, un jour, rendre les données (stats des joueurs) disponibles sur une plateforme quelconque.

3 Modélisation

Le projet le plus ambitieux est de loin celui de prédire efficacement le résultat des matchs. Pour ce faire, il serait bien de procéder par étapes.

3.1 Régression logistique

La plan était (et est peut-être encore) de commencer par effectuer une régression logistique pour prédire le résultat d'un match. Ça nous permettrait aussi de savoir quelles variables semblent utiles et de comprendre les données en général.

3.2 Réseau de neurones

Avec le feedback de la régression logistique, nous pourrons sélectionner une bonnes quantité de variables à utiliser comme première couche d'un réseau de neurones.

3.3 Apprentissage non-supervisé

La qualité des outils de visualisation qu'on aura pour mettre en lumière certains aspects d'un joueur ne passe pas nécessairement directement par les réseaux de neurones. Développer des algorithmes de *clustering* adaptés à nos questions peut nous permettre de résumé l'information de façon efficace, donc de faire des graphiques épurés et facile à comprendre, mais quand même chargés en information.

4 Étape par étape...

Évidemment, nous ne pouvons pas tout faire en 2 mois, encore moins en une fin de semaine. En ce qui attrait à la modélisation, les étapes naturelles sont de commencer avec la régression et ensuite se tourner vers le réseau de neurones. Mais que ce soit pour la régression ou le réseau de neurones, on a une job de nettoyage de données assez intense à faire. La première étape semble donc de se concentrer sur les données, mais travailler un peu sur la régression peut nous donner des idées de comment nous voulons que nos données soient disponible.

Je crois qu'à moyen terme on peut travailler de façon un peu non-structuré pour voir ce qui émerge du projet. Mais, encore une fois, pour ce faire on a au moins besoin d'une base : une base de données en ordre.

4.1 Les données

Nous avons déja du bon data. Il faut le nettoyer. La job de nettoyage n'est pas la même tout dépendant quel jeu de données nous utilisons. À voir !

4.2 Modélisation

Encore une fois, ça dépend du jeu de données, mais aussi du projet qu'on choisit (régression ou réseau ou autre). Une choses est certaine, nous devrons bâtir de nouvelles variables qui sont (trop) implicites dans le jeu de données. Je pense ici à des facteurs de fatigue par exemple, qui nécessitent qu'on calcul (encore une fois un exemple) le nombre de match joué dans les x derniers jours... Il serait bien de commencer par identifier ces variables.