Tenis Game Prediction

Challenge DataGenius

29 dec 2017

1 Description

Le but de ce challenge est de créer un modèle statistique permettant de prédire quel sera le gagnant d'un match de tennis. Pour cela, vous aurez à votre disposition des sets de données d'entraînement comportant (pour chaque match) :

— le résultat final

— des informations sur les deux joueurs (des statistiques sur leurs matchs passés)

— des informations sur le match (date, terrain...)

Il faudra alors entraîner un modèle qui, en tenant compte des informations relatives aux joueurs et à l'environnement, sera capable de prédire le gagnant pour un match n'ayant pas encore été disputé.

A travers cette épreuve, nous cherchons à évaluer votre capacité à comprendre un problème et à le formaliser dans le but de le résoudre par l'apprentissage d'un modèle statistique. Ne perdez pas de vue que votre démarche nous intéresse plus que les résultats. Nous porterons une attention particulière à votre capacité à :

— extraire les données (feature extraction)

— mettre en forme la donnée (feature engineering)

— sélectionner votre modèle (model selection)

— paramétrer votre modèle (hyper-parameters searching)

— entraîner votre modèle final en évitant le sur-apprentissage

2 Datasets

train.csv :C'est la base d'entraînement. Chaque ligne correspond à un match.

— ID1_G : L'identifiant du joueur qui a gagné le match

— ID2_G : L'identifiant du joueur qui a perdu le match

— ID_T_G : L'identifiant du tournoi (voir tour.csv)

— ID_R_G : L'identifiant du round dans le tournoi

— RESULT_G ; Le résultat du match

— DATE_G : La date du match

stats.csv : C'est un fichier représantant des rencontres anonymes entre deux joueurs avec les stats. Vous pouvez vous servir de ces statistiques afin de modéliser des informations sur chacuns des joueurs.

— ID1 : L'identifiant du joueur qui a gagné le match

— ID2 : L'identifiant du joueur qui a perdu le match

— ID_T : L'identifiant du tournoi (voir tour.csv)

— ID_R : L'identifiant du round dans le tournoi (voir rounds)

— FS_1 : Nombre de premiers services réussis (joueur 1)

— FS_OF1 : Nombre de premiers services (joueur 1)

— ACES_1 : Aces (joueur 1)

— DF_1 : Double Fautes (joueur 1)

— UE_1 : erreurs directes

— W1S_1 : Nombre de points gagnés sur premier service

— W1SOF_1 : Nombre de points joués sur premier service

— W2S_1 : Nombre de points gagnés sur second service

— W2SOF_1 : Nombre de points joués sur second service

— WIS_1 : Nombre de points gagnés en tout

— BP_1 : Nombre de balles de break gagnées

— BPOF_1 : Nombre de balles de break obtenues

— RPW_1 : Nombre de points gagnés

— RPW_OF1 : Nombre de points jouées

— .. 2 : Les mêmes pour le joueur 2

players.csv : Information sur les joueurs individuels.

— ID_P : L'identifiant du joueur

— NAME_P : Nom du joueur

— DATE_P : Date de naissance

— COUNTRY_P : Pays

— RANK_P : Classement

tour.csv Description des différents tournois. NB : la date du tournoi peut être utilisée comme date de match si un match n'est pas associé à une date.

player_rates.csv Classement des joueurs

3 Evaluation

L'évaluation de votre modèle se fera sur le taux de bonne classification.

Le fichier de soumission est test.csv (valeurs séparées par des ‘,’) de 5 colonnes :

ID Joueur 1

ID Joueur 2

ID Tournoi

ID Round

Winner Résultat prédit (1 si joueur gagne, 2 si joueur 2)