Projet M1: “Prédire qui gagne le prix pour le meilleur défenseur de la NBA” — sélection d’attributs

5 décembre 2023

1 Suppression manuelle

Basketball-reference :
— les statistiques offensives :
  — PTS
  — AST
  — “Shooting”
  — FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, eFG%, FT, FTA, FT%, ORB
  — ORtg
  — TS%, 3PAr, FTr, OWS, OBPM
  — “Offense Four Factors”
D’autres sites :
— ORPM (ESPN)
— ESTIMATED +/- OFF (Dunks and Threes)
— “Efficiency”
— “Location”
— OR%
— HANDLE
— pour bball-index, on veut que D-LEBRON

2 Méthodes de sélection

Méthodes à tester :
— Select from Model : `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html#sklearn.feature_selection.SelectFromModel`
— Sequential feature selection : `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html#sklearn.feature_selection.SequentialFeatureSelector`

On va traiter le problème comme un problème de regression pour l’instant, c.-à-d. l’étiquette à prédire est numérique (le nombre de points qu’un joueur à

gagner, p.ex. 391 pour Jaren Jackson Jr. en 2023, 209 pour Brook Lopez etc.).
Modèles à évaluer :

- Bayesian ridge regression : `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html`
- Kernel ridge regression : `https://scikit-learn.org/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html`
- SVM : `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html`
- Gradient Boosted Trees : `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html`
- Multi-layer perceptron : `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html`

Une approche alternative consiste en calculant l'importance d'un attribut et utiliser un seuil (ou un nombre maximal) pour sélectionner les attributs.
Mesures d'importance/de corrélation

- Pearson's R : `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.r_regression.html`
- Mutual Information : `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html`
- χ^2 : `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html`

3 Pour plus tard

À un moment, il sera sans doute plus efficace de traiter le problème comme une regression ordinale, mais il n'y a pas des algorithmes de ce type dans sklearn.

Je vous propose donc de se familiariser avec `https://analyticsindiamag.com/a-complete-tutorial-on-ordinal-regression-in-python/`.