MAT-MEX

C++

Полная документация к проекту по прогнозированию футбольных матчей.
Оформлено при помощи LaTeX.

Студент: Шаламов Иван.

Группа: 251.

Санкт-Петербург
2017

Содержание

1 Вступление.

- Данный проект содержит в себе модели прогнозирования футбольных матчей.

- Основные цели проекта: создание и развитие рабочих моделей на основе машинного обучения, с помощью которых возможно реализовать прогнозирование исходов футбольных матчей.

- Подготовка данных для обучения производилась на основе функционала библиотеки scikit learn.

- Визуальное взаимодействие с данными производилось на основе функционала библиотеки pandas.

2 Препроцессорные директивы.

Препроцессорные директивы отвечают за импортирование заголовочных файлов. В данном проекте представлены следующие из них:

- # **import pandas as pd** (Библиотека для работы с данными и их отображением).
- # **import xgboost as xgb** (Библиотека одного из классифаторов).
- # **from sklearn.linearmodel import LogisticRegression** (Библиотека одного из классифаторов).
- # **from sklearn.ensemble import RandomForestClassifier** (Библиотека одного из классифаторов).
- # **from sklearn.svm import SVC** (Библиотека одного из классифаторов).
- # **from IPython.display import display** (Библиотека для вывода данных на экран).

3 Основные принципы работы.

- Основной принцип реализации: построить рабочии модели, которые на основе имеющихся данных могли бы предсказать исход матча (атрибут FTR).

- Для того, чтобы пользователю было комфортно работать с имеющимися данными, было решено оставить 12 атрибутов, которые по нашей модели помогут предсказать победителя матча:

  HTP, ATP, HM1, HM2, HM3, AM1, AM2, AM3, HTGD, ATGD, DiffFormPts, DiffLP.

- Имеющиеся стандартизированные данные мы разделяем на обучающуюся и тестируемую части при помощи функций scikit learn, чтобы начать работу с классифаторами.

- Мы создаем три модели, три классифатора, которые будут работать с нашими данными, пытаясь на их основе предсказать исход матча: Logistic Regression, Support Vector Machine, XGBoost.

4 Классифаторы и методы их использования.

Проект содержит следующие классифаторы:

Logistic Regression – Применяется для предсказания вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится так называемая зависимая переменная y, принимающая лишь одно из двух значений — как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами) — вещественных x1,x2,...,xn, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

SVM – Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

XGBoost – XGBoost - это алгоритм, который недавно доминировал в прикладном машинном обучении и соревнованиях Kaggle для структурированных или табличных данных. XGBoost - это реализация градиентных деревьев принятия решений, предназначенных для скорости и производительности. Он создает несколько деревьев, на основе которых строит модель для предсказаний.

5 Источник, благодаря которому был составлен файл с данными:

- http://football-data.co.uk/data.php.

6 Заключение.

Данный проект был собран на базе Jupyter Notebook.