Metody Odkrywania Wiedzy - Dokumentacja wstępna projektu

Rafał Okuniewski, Maciej Zaborek

11 kwietnia 2017

1 Temat projektu

Temat analityczny dotyczący regresji na zbiorze danych nr 14, dotyczącychpredykcji popytu w systemie rowerów miejskich.

2 Szczegółowa interpretacja tematu

Zbiór danych dostępny pod adresem `https://www.kaggle.com/c/bike-sharing-demand/data` Zawiera on dane zebrane na przestrzeni dwóch lat. Każda próbka to opis każdej godziny pracy systemu w ciągu dnia. Atrybuty to kolejno:

- data
- pora roku
- czy dany dzień jest świętem
- czy dany dzień jest dniem roboczym
- pogoda - jedna z czterech wartości:
  - 1 - bezchmurnie, lekkie zachmurzenie, częściowe zachmurzenie
  - 2 - zamglenie + zachmurzenie, zamglenie + duże zachmurzenie, zamglenie + lekkie zachmurzenie, zamglenie
  - 3 - lekkie opady śniegu, lekki deszcz + błyskawice + duże zachmurzenie, lekki deszcz + ciężkie zachmurzenie
  - 4 - mocny deszcz + marznący deszcz + błyskawice + mgła, śnieg + mgła
- temperatura
- odczuwalna temperatura
- wilgotność

- prędkość wiatru
- liczba rozpoczętych wypożyczeń niezarejestrowanych użytkowników
- liczba rozpoczętych wypożyczeń przez zarejestrowanych użytkowników
- **całkowita liczba wypożyczeń** - atrybut ciągły podlegający pomiarowi błędu

W każdym miesiącu mamy do czynienia z danymi z 19 dni, będącymi danymi uczącymi. Celem przeprowadzanej regresji jest przewidzenie liczby wypożyczonych rowerów w każdej godzinie 20. dnia, który stanowi zbiór danych testowych, na podstawie dotychczasowych obserwacji.

3 Charakterystyka zbioru danych

Dostępne dane są w formie nie wymagającej dalszego przetwarzania lub uzupełniania.

4 Opis wykorzystywanych algorytmów oraz parametrów wymagających strojenia

- Regresja liniowa (linear regression) - metoda lm
- Regresja lokalna (local regression) - metoda regresji wykorzystująca równania wielomianowe w otoczeniu danego punktu do wyznaczenia wartości atrybutu docelowego, metoda loess. Parametry - degree - stopień wielomianu, span - stopień wygładaania
- Robust regression z M-estymacją - rodzaj regresji odporny na zanieczyszczone dane trenujące wykorzystywany w sytuacji (takiej jak potencjalnie w zbiorze dotyczącym predykcji popytu w systemie rowerów miejskich), gdy pomimo istnienia wartości odstających, nie należy traktować ich jako błedne czy nie pochodzące z populacji. Opiera się na ważeniu obserwacji ze zbioru treningowego - zaimplementowane w pakiecie MASS w języku R
- Drzewa regresyjne (Regression trees) - drzewa decyzyjne wykorzystujące do stworzenia podziałów sumę kwadratów błędów, zaimplementowane w pakiecie rpart. Parametry - głębokość przycinania.

5 Algorytmy selekcji atrybutów

- Stepwise Regression - stepAIC
- Regsubsets(leafs)

6 Plan badań

1. **Analiza zbioru trenującego** mająca na celu redukcję wartości skrajnie odstających od pozostałych w zbiorze dla algorytmów, które tego wymagają, badanie korelacji pomiędzy atrybutami, wyodrębnienie potencjalnie przydatnych dla potrzeb regresji podzbiorów atrybutów. Wykonana zostanie również normalizacja atrybutów poprzez ich standaryzację.

2. **Analiza algorytmów regresji** - analiza mająca na celu ustalenie najlepszego dla danego problemu algorytmu regresji, umożliwiającego zminimalizowanie błędu. Analizie poddane zostaną również parametry wykorzystywanych algorytmów celem odnalezienia optymalnych wartości parametrów

3. **Miary jakości i procedury oceny modeli** - miarą jakości każdego algorytmu będzie błąd średniokwadratowy (RMSE) i błąd średniokwadratowy logarytmiczny (RMSLE) - celem możliwości porównania wyników wewnętrznej walidacji z tą przeprowadzoną na zaślepionym zbiorze testowym udostępnianym przez Kaggle.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i + 1) - log(a_i + 1))^2} \tag{2}$$

   Wielkość udostępnionego zbioru treningowego pozwala na przeprowadzenie wewnętrznej walidacji modelu z wykorzystaniem tego zbioru. Modele regresji zostaną ocenione z uwzględnieniem trzech metod:

   (a) walidacji krzyżowej k-krotnej na zbiorze treningowym

   (b) walidacji krzyżowej leave-one-out na zbiorze treningowym

   (c) pomiaru błędu przy wykorzystaniu aplikacji Kaggle na pobranym zbiorze testowym

4. **Analiza istotności atrybutów** - przeprowadzona zostanie analiza istotności atrybutów na podstawie wyjścia klasyfikatorów, np. drzewa regresji.

7 Otwarte kwestie wymagające doprecyzowania

Doprecyzowania wymaga ilość podzbiorów tworzonych w wyniku walidacji krzyżowej k-krotnej.