Raport

Tomasz Żelawski - predykcja wyników ucznia w szkole średniej

Opis danych wejściowych

Dane zawierają 30 atrybutów oraz 3 wartości.

Atrybutami są dane socjoekonomiczne uczniów, takie jak:

- poprzednia szkoła ucznia,
- wykształcenie rodziców,
- zawód rodziców itd.

Wartościami są oceny końcowe uczniów w szkole średniej,

- z pierwszego semestru,
- z drugiego semestru,
- z trzeciego semestru.

Dane są zarówno ilościowe jak i kategorialne.

Dla uproszczenia problemu rozpatrywany był problem binarny - czy uczeń zda, czy nie. W pierwszej iteracji modelu celem treningowym jest zaliczenie pierwszej klasy (G1) a predykcja jest robiona dla danych testowych pierwszej, drugiej (G2) i trzeciej klasy (G3).

W drugiej iteracji G1 oraz G2 zostały przesunięte do zbioru atrybutów, żeby sprawdzić jak duży wpływ na wyniki ucznia mają jego poprzednie rezultaty. Predykcja jest robiona tylko dla danych testowych G3.

Strategia podziału danych

Dane zostały zasadniczo podzielone na dwa zestawy:

- Zestaw treningowy: 80% danych
- Zestaw testowy: 20% danych

Nie zostały podjęte żadne kroki gwarantujące jednorodność danych w zestawie treningowym i testowym - ze względu na skomplikowanie danych, byłby to najprawdopodniej osobny problem klasyfikacji.

Wyniki testowe i treningowe

Model AdaBoost z Decision Tree Classifier

Ze względu na to, że nie jest oczywiste jakie metryki mają znaczenie w przypadku tego problemu, zdecydowałem się polegać na *accuracy*.

Dokładne dane oraz więcej metryk można znaleźć w pliku `model.ipynb`.

- **Wyniki Treningowe:**
  - Accuracy: 0.834
- **Wyniki Testowe:**
  - G1:
    - Accuracy: 0.808
  - G2:
    - Accuracy: 0.785
  - G3:
    - Accuracy: 0.831

Model Keras Sequential

- **Wyniki Treningowe:**
  - Accuracy: 0.808
- **Wyniki Testowe:**
  - G1:
    - Accuracy: 0.854
  - G2:
    - Accuracy: 0.815
  - G3:
    - Accuracy: 0.846

Jak widać, model sieci neuronowej wyprodukował lepsze wyniki, ale nieznacznie.

Model AdaBoost z Decision Tree Classifier na rozszerzonych danych

- **Wyniki Treningowe:**
  - Accuracy: 0.956
- **Wyniki Testowe:**
  - Accuracy: 0.892

Model Keras Sequential na rozszerzonych danych

- **Wyniki Treningowe:**
  - Accuracy: 0.998
- **Wyniki Testowe:**
  - Accuracy: 0.938

Jak widać, oba modele doznały overfittingu, zwłaszcza sieć neuronowa, która wciąż jednak daje lepsze wyniki. Oba modele istotnie się poprawiły.

2. Uzasadnienie wyboru modeli

W oryginalnej pracy naukowej (patrz źródło) użyte zostały lasy decyzyjne. W związku z tym, że praca pochodzi z 2008 roku, brak zastosowania modeli sieci neurowonych nie jest wielkim zaskoczeniem. Postanowiłem sprawdzić model lasu decyzyjnego i model sieci neurowonej a następnie porównać ich wyniki.

Analiza wyników

Moim zdaniem osiągnięte wyniki są zdecydowanie zbyt słabe, aby utworzone tutaj modele mogły służyć do realnych zastosowań czy wiarygodnych predykcji. Dla większego zestawu danych oba modele doznały overfittingu.

Nie jest jednak do końca jasne, czy niesatysfakcjonujący rezultat jest efektem źle dobranych modeli. Problem taki jak predykcja wyników w szkole ucznia jest zdecydowanie bardzo złożony i nie wszystkie istotne parametry mogą być dostępne w danych - widać to zwłaszcza po tym, jak wynik się poprawił po przesunięciu G1 i G2 do atrybutów.

Dla poprawienia wyników trzeba by przede wszystkim zmodyfikować metody regularyzacji, aby zredukować duży overfitting. Myślę też, że dobrym wyborem byłoby częściowo sklasyfikować dane wejściowe, aby mieć pewność, że pomiędzy danymi treningowymi i testowymi nie ma zbyt dużego zróżnicowania.

Źródła

- USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE, Paulo Cortez and Alice Silva