DM2020 Projekt 2

Opis projektu

1. Celem projektu jest praktyczne sprawdzenie metod klasyfikacji.
2. Celem jest identyfikacja klasy 1 (zmienna `poor="poor"`).
3. W pliku *train_data.txt* i *train_labels.txt* znajdują się dane treningowe (6000 obserwacji). W pliku *validation_data.txt* dostępne są dane walidacyjne (3000 obserwacji), których etykiety zostaną udostępnione 31 maja. Dostępne są również dane *individual_data.txt* (dla obserwacji ze zbiorów treningowego, walidacyjnego i testowego razem).
4. Należy dokonać predykcji dla danych ze zbioru *test_data.txt* (3000 obserwacji), każdej obserwacji przypisujac prawdopodobieństwo klasy 1.
5. Wyniki nalezy zapisać do pliku, którego nazwa to: piewsza litera imienia + dwie pierwsze litery nazwiska + ".txt". Przykładowo, student Jan Kowalski zapisze wyniki do pliku JKO.txt. W pliku w każdym wierszy powinny się znaleźć prawdopodobieństwa odpowiadajace klasie 1. Przykładowy plik: MLA.txt.
6. Należy przetestowac co najmniej 4 metody klasyfikacji.
7. Projekty sa wykonywane indywidualnie.
8. Ocena na podstawie:

- jakości klasyfikacji mierzonej jako AUC (20 pkt),
- prezentacji (5-7 minut) podsumowującej wyniki (10 pkt),
- raportu (maksymalnie 3 strony A4), który zawiera: podsumowanie eksperymentów, uzasadnienie wyboru końcowej metody, opis wyboru i przekształceń zmiennych (15 pkt),
- kodów źródłowych (5 pkt).

9. Prezentacje odbędą się na ostatnich zajęciach 8 czerwca (odrabianych za 12.03.2020).
10. Wyniki, raport, i kody należy wysłac do 5 czerwca (godzina 22:00) na adres: m.lazecka(at)mini.pw.edu.pl.
11. Projekt można wykonywać w R lub Python.

Terminy

07.05-21.05 - dostępne pliki *train_data.txt*, *train_labels.txt*, *validation_data.txt*, *individual_data.txt*, *test_data.txt*

21.05-31.05 - dostępna aplikacja, w której można sprawdzić AUC dla predykcji dla obserwacji ze zbioru *validation_data.txt* (będzie to można zrobić anonimowo, format pliku wsadowego powinien być taki sam jak opisany wyżej)

31.05-5.06 - dostępne pliki *train_data.txt*, *train_labels.txt*, *validation_data.txt*, *validation_labels.txt*, *individual_data.txt*

5.06 do 22:00 - oddanie raportu, pliku zawierającego predykcje dla obserwacji z pliku *test_data.txt* i przesłanie kodu

8.06 (pniedziałek) godz. 9.15 - prezentacja wyników

Dane

Opis danych dostępny jest w dwóch plikach pdf. Celem jest przewidzenie, czy dane gospodarstwo dotyka bieda (zmienna `poor` - informacje o tej zmiennej dostępne są w plikach zawierających "labels"; na początku dostępny jest tylko plik *train_labels.txt*, później również *validation_labels.txt*). Można do tego wykorzystać zarówno dane na poziomie gospodarstwa (wszystkie pliki oprócz *individual_data.txt*), jak i dane indywidualne dla mieszkańców. Dane można połączyć korzystając z kolumny `hid`.