Wstęp do Uczenia Maszynowego

Projekt

1 Wstęp

Celem jest zaproponowanie metody klasyfikacji, która pozwoli zbudować model o jak największej mocy predykcyjnej. Dysponujemy sztucznie wygenerowanym zbiorem danych *artificial*, w którym zostały ukryte istotne zmienne. Należy dokonać klasyfikacji do dwóch klas. Dokładność modelu będzie mierzona za pomocą miary zrównoważonej dokładności (*balanced accuracy*).

Projekt jest wykonywany samodzielnie!

2 Zbiór danych

Dane do projektu to sztucznie wygenerowany zbiór, który zawiera 30 zmiennych objaśniających (część z tych kolumn może być zbędna). Zbiór treninowy zawiera 2000 obserwacji, natomiast zbiór testowy 600.

Dostępne są następujące pliki:

- zbiór treningowy: `artifical_train_data.csv`,
- etykiety zbioru treningowego: `artifical_train_labels.csv`,
- zbiór testowy: `artifical_test_data.csv`.

Aby wczytać zbiór danych w języku Python wystarczy użyć funkcji `read_csv` z pakietu `pandas`.

3 Oczekiwany wynik

Na przygotowanie rozwiązania projektu będą składały się następujące elementy:

- jakość predykcji na zbiorze testowym mierzona przez *balanced accuracy*

$$BA = \frac{1}{2}\Big(\frac{TP}{P} + \frac{TN}{N}\Big),$$

- raport opisujący wykorzystane metody i wyniki eksperymentów (maksymalnie 4 strony A4),
- krótka prezentacja podumowująca rozwiązanie (maksymalnie 4 minuty).

4 Szczegóły rozwiązania

Zbiór treningowy oraz etykiety do zbioru treningowego należy wykorzystać do przygotowania modelu. Oczekiwany wynik to wektor prawdopodobieństw przynależności do klasy 1 dla obserwacji ze zbioru testowego.

Rozwiązanie powinno zawierać pliki:

- `NUMERINDEKSU_artifical_prediction.txt` - prawdopodobieństwo przynależności do klasy 1 dla danych testowych, gdzie `NUMERINDEKSU` zastępujemy swoim numerem indeksu (przykładowy plik `example_artifical_prediction.txt`),
- folder `Kody` zawierający wszystkie potrzebne kody do przygotowania rozwiązania projektu,

- plik `NUMERINDEKSU_raport.pdf` opisujący wykorzystane metody i wyniki eksperymentów (maksymalnie 4 strony),
- plik `NUMERINDEKSU_prezentacja.pdf` krótka prezentacja podsumowująca rozwiązanie (maksymalnie 4 minuty).

5 Ocena

Łączna liczba punktów do zdobycia jest równa 40, w tym:

- jakość kodu (porządek, czytelność, obszerność eksperymentów) - 12 punktów,
- jakość predykcji rozwiązania* - 8 punktów,
- raport - 15 punktów,
- prezentacja - 5 punktów.

* - jakoś predykcji jest oceniana miarą *balanced accuracy* na zbiorze testowym. Wyniki zostaną ustawione w ranking (od najlepszego do najgorszego). Osoba z najlepszym wynikiem (najblizszym wartości 1) zyskuje 8 punktów. Osoba z najgorszym wynikiem (najbliższym wartości 0) zyskuje 0 punktów. Pozostałe wyniki zostaną przeskalowane i zaokrąglone do wartości 0.5.

6 Oddanie projektu

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NUMERINDEKSU_GR_projekt`, gdzie

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15,} \\ 2 & \text{dla środa, 16:15.} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 17.01 do godziny 23:59. Tytuł wiadomości: *[WUM][Projekt] Nazwisko Imię, Numer grupy: GR.*

7 Terminy

1. Oddanie projektu - 14 tydzień zajęć (17.01.2024),
2. Wyniki projektu oraz prezentacje na zajęciach laboratoryjnych - 15 tydzień zajęć (24.01.2024).