Praca domowa 1

Termin oddania: 23.10.2024

1 Wstęp

Celem pracy domowej jest sprawdzenie jak poszczególne hipeparametry modelu drzewa decyzyjnego wpływają na jego jakość predykcyjną.

2 Zbiór danych

W tym celu posłużymy się zbiorem danych $\mathcal{D} = (X, y)$, gdzie X `= X.csv,` y `= y.csv`. Aby przygotować dane do dalszej pracy należy podzielić zbiór $\mathcal{D}$ na treningowy i testowy w proporcji 7:3 ustawiając parametr `random_state = NUMER_INDEKSU`. Zbiór testowy należy wykorzystać do ostatecznej oceny wybranego modelu.

3 Oczekiwany wynik

Praca domowa składa się z trzech elementów. Pierwszym będzie przygotowanie eksperymentu pozwalającego odpowiedzieć na pytanie, które hipeparametry modelu drzewa decyzyjnego są najlepsze dla zadanych danych. Kolejnym elementem będzie ocena jakości modelu, a ostatnim sprawdzenie czy wielkość próbki danych wpływa na jakość predykcyjną modelu.

3.1 Eksperyment *(8 punktów)*

Przygotuj eksperyment ukazujący miarę AUC drzewa decyzyjnego na zbiorze treningowym i testowym (wykorzystując kroswalidację, co najmniej 5-krotną na danych treningowych z ustawionym parametrem `random_state = NUMER_INDEKSU`) w zależności od parametrów:

- kryterium podziału (*criterion*{ *"gini", "entropy"*}),
- głębokość drzewa (*max_depth*),
- minimalna liczba obserwacji w liściu (*min_samples_leaf*).

Przejrzyj dokumentację dotyczącą budowy drzewa i spróbuj znaleźć inne parametry, które poprawią dokładność.
Opisz przeprowadzony eksperyment oraz wnioski z niego płynące w formie raportu (maksymalnie 2 strony A4).

3.2 Analiza jakości predykcyjnej modelu *(3 punkty)*

Na podstawie wyników z Sekcji 3.1 wybierz Twoim zdaniem najlepszy model, podaj uzasadnienie wyboru. Dla wybranego modelu na danych treningowych i testowych wyznacz:

- macierz pomyłek,
- dokładność (ang. *accuracy, ACC*), czułość (ang. *sensitivity, recall*), precyzja (*ang. precision*),
- krzywą ROC, wartość AUC.

3.3 Wpływ rozmiaru próbki danych na jakość predykcyjną modelu (*4 punkty*)

Dla wybranego modelu w Sekcji 3.2 przeprowadź eksperyment, w którym w losowy sposób wybierzesz `5%, 10%, 25%, 50%, 75%, 90%, 95%` początkowych danych, wytrenujesz model z wybranymi hiperparametrami i ocenisz jego zdolność predykcyjną miarą AUC dla zbioru treningowego i testowego. Czy rozmiar próbki danych wpływa na jakość predykcyjną Twojego modelu?

4 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- folder `Kody` zawierający wszystkie potrzebne kody do odtworzenia rozwiązania zadania domowego,
- plik `NUMERINDEKSU_raport.pdf` opisujący przeprowadzony eksperyment, analizę wybranego modelu oraz badanie wpływu próbki danych na jakoś predykcyjną modelu (maksymalnie 4 strony).

5 Ocena

Łączna liczba punktów do zdobycia jest równa 15, w tym:

- 3.1 Eksperyment (*8 punktów)*
  - jakość kodu (porządek, czytelność) - 1 punkt,
  - jakość eksperymentu - 4 punkty,
  - raport - 3 punkty.
- 3.2 Analiza jakości predykcyjnej modelu (*3 punkty)*
  - jakość kodu (porządek, czytelność) - 1 punkt,
  - wnioski - 1 punkt,
  - raport - 1.5 punktu.
- 3.3 Wpływ rozmiaru próbki danych na jakość predykcyjną modelu (*4 punkty)*
  - jakość kodu (porządek, czytelność) - 1 punkt,
  - wnioski - 1 punkt,
  - raport - 1.5 punktu.

6 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NAZWISKO_IMIE_GR_PD1` (bez polskich znaków), gdzie

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15,} \\ 2 & \text{dla środa, 14:15,} \\ 3 & \text{dla środa, 16:15.} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 23.10.2024 do godziny 23:59. Tytuł wiadomości: *[WUM][PD1] Nazwisko Imię, Numer grupy: GR.*