Praca domowa 3

Termin oddania: 20.12.2023

1 Wstęp

W tej pracy domowej przyjrzymy się modelowi k najbliższych sąsiadów. Celem pracy jest implementacja metody oraz jej przetestowanie.

2 Część 1

Przygotuj implementację algorytmu k – najbliższych sąsiadów (funkcję `knn()`), który przyjmuje jako argumenty kolejno:

1. macierz rzeczywistą $\mathbb{X}$ typu $n \times p$, reprezentującą n punktów w $\mathbb{R}^p$ (zbiór treningowy),
2. n – elementowy obiekt $\mathbf{y}$, gdzie y_i reprezentuje etykietę odpowiadającą obserwacji $\mathbb{X}[i,]$,
3. macierz rzeczywistą $\mathbb{Z}$ typu $m \times p$, reprezentującą m punktów w $\mathbb{R}^p$ (zbiór testowy),
4. liczbę całkowitą $1 \leqslant k \leqslant n$, oznaczającą liczbę najbliższych sąsiadów biorących udział w poszukiwaniu etykiety odpowiadającej punktom ze zbioru testowego,
5. wartość rzeczywistą $p \geqslant 1$, domyślnie równą 2, określającą, która metryka Minkowskiego L_p będzie używana do poszukiwania najbliższych sąsiadów. Uwaga, możliwe jest, by $p = \infty$.

Funkcja ma zwracać m – elementowy obiekt $\mathbf{w}$, gdzie w_i reprezentuje etykietę odpowiadającą obserwacji $\mathbb{Z}[i,]$.
Dla $i = 1, \ldots, m$, etykieta w_i wyznaczana jest w następujący sposób:

1. Niech d_j oznacza odległość L_p między `Z[i,]` a `X[j,]`, tj. $d_j = \|\texttt{Z[i,]} - \texttt{X[j,]}\|_p$, $j = 1, \ldots, n$.
2. Niech $(j_1, \ldots, j_k)$ oznaczają indeksy k najbliższych `Z[i,]` punktów z $\mathbf{X}$, tj. $d_{j_1} \leqslant d_{j_2} \leqslant \cdots \leqslant d_{j_k} \leqslant d_j$ dla każdego $j \notin \{j_1, \ldots, j_k\}$.
3. Wyznacz modę (dominantę) z ciągu etykiet $(y_{j_1}, \ldots, y_{j_k})$ i przypisz jako wartość w_i. Jeśli moda nie jest określona w jednoznaczny sposób, zwróć losową najczęściej występującą wartość (rozkład jednostajny - każda z tą samą miarą prawdopodobieństwa).

3 Część 2

W tej części przeprowadź test poprawności implementacji funkcji `knn()` na przynajmniej dwóch zbiorach danych z $\mathbb{R}^2$. W szczególności należy sprawdzić, czy 1 – nn w przypadku, gdy próba ucząca i testowa są tożsame, odtwarza idealnie wektor prawdziwych etykiet.

Przetestuj algorytm k – najbliższych sąsiadów z metryką L_1, L_2 oraz L_∞ dla różnych k.

4 Część 3*

Wykorzystując gotowe implementacje metody k – nn porównaj otrzymane wyniki ze swoją implementacją. W tym celu możesz wykorzystać na przykład `Annoy` (https://github.com/spotify/annoy), `ScaNN` (https://github.com/google-research/google-research/tree/master/scann) lub inną.

5 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- folder `Kody` zawierający wszystkie potrzebne kody do przygotowania rozwiązania zadania domowego,
- plik `NUMERINDEKSU_raport.pdf` opisujący wyniki testów (maksymalnie 4 strony).

6 Ocena

Łączna liczba punktów do zdobycia jest równa 10 + 3, w tym:

Część 1 (6 punktów)

- implementacja metody - 5 punktów,
- jakość kodu (porządek, czytelność) - 1 punkt.

Część 2 (4 punkty)

- testy metody - 2 punkty,
- raport - 2 punkty.

Część 3* (3 punkty)

- testy metody - 2 punkty,
- raport - 1 punkt.

7 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NUMERINDEKSU_GR_PD3`, 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 20.12.2023 do godziny 23:59. Tytuł wiadomości: *[WUM][PD3] Nazwisko Imię, Numer grupy: GR.*