Naiwny Bayes

1. Opis algorytmu

1.1 Tworzenie danych.

Z pliku cars_evaluation.csv wymieszane dane dzielimy na zbiór testowy – 30% i zbiór treningowy 70%.

1.2 Metoda `classify`.

Zadaniem metody *classify* jest zaklasyfikowanie atrybutu decyzyjnego dla każdej obserwacji ze zbioru testowego. Metoda *classify* iteruje po wszystkich wierszach zbioru testowego i wywołuje metodę `calculate_probabilities` (1.3), zwracającą słownik- `probabilities`, z prawdopodobieństwami dla każdego atrybutu decyzyjnego.
Zaklasyfikowany zostaje atrybut z największym prawdopodobieństwem. W międzyczasie rejestrowana jest liczba poprawnych odpowiedzi- `correct_classifications`, do późniejszego policzenia *Accuracy*.

1.3 Metoda `calculate_probabilities`.

Metoda `calculate_probabilities` dla podanego w argumencie wiersza- obserwacji zwraca słownik- `probabilities`, z wartościami prawdopodobieństw zestawu atrybutów dla każdego atrybutu decyzyjnego (pętla po liście `deciosions`):

['**unacc**']: P (X | atrybut = unacc),
['**acc**']: P (X | atrybut = acc),
['**vgood**']: P (X | atrybut = vgood),
['**good**']: P (X | atrybut = good)

Prawdopodobieństwo obliczamy za pomocą twierdzenia Bayesa:

$$P(A \mid B) = (P(B \mid A) * P(A)) / P(B)$$

P (‘unacc’ | X) = P (X | ‘unacc’) * P (‘unacc’) = P (x1 | ‘unacc’) * P (x2 | ‘unacc’) * … * P (‘unacc’)

Informacje czerpiemy z wyznaczonego zbioru treningowego- `train_set`.
Szukane prawdopodobieństwo- `final_probability` rozpoczynamy liczyć od wyznaczenia prawdopodobieństwa wystąpienia atrybutu decyzyjnego- `decision_probability`.
Następnie klasyfikator iteruje po wyznaczonych atrybutach obliczając prawdopodobieństwo każdego z nich dla aktualnego atrybutu decyzyjnego – `attribute_probability`.
W przypadku, kiedy liczba atrybutów spełniających kryterium- `attribute_occ` jest równa 0 stosujemy *wygładzanie*.
Po każdym przejściu pętli atrybutów mnożymy docelowy wynik przez kolejne prawdopodobieństwo. Dodajemy obliczone końcowe prawdopodobieństwa do słownika- `probabilities`.

1.4 Metoda `show_results`.

Wyświetlenie liczby poprawnych kwalifikacji- `correct_anwsers`.
Przedstawienie dokładności- *Accuracy* w procentach.

2. Wyniki klasyfikacji

Rezultaty dziesięciokrotnego uruchomienia algorytmu:

0. Results for a 518x test size:
Number of correct classifications: 433
Accuracy: 83.59073359073359 %

1. Results for a 518x test size:
Number of correct classifications: 451
Accuracy: 87.06563706563706 %

2. Results for a 518x test size:
Number of correct classifications: 440
Accuracy: 84.94208494208493 %

3. Results for a 518x test size:
Number of correct classifications: 431
Accuracy: 83.20463320463321 %

4. Results for a 518x test size:
Number of correct classifications: 445
Accuracy: 85.9073359073359 %

5. Results for a 518x test size:
Number of correct classifications: 435
Accuracy: 83.97683397683397 %

6. Results for a 518x test size:
Number of correct classifications: 436
Accuracy: 84.16988416988417 %

7. Results for a 518x test size:
Number of correct classifications: 446
Accuracy: 86.10038610038609 %

8. Results for a 518x test size:
Number of correct classifications: 440
Accuracy: 84.94208494208493 %

9. Results for a 518x test size:
Number of correct classifications: 448
Accuracy: 86.48648648648648 %

Średnia wyników Accuracy wynosi w przybliżeniu: 85.04%