Naiwny klasyfikator bayesowski

Marcin Połosak

Styczeń 2024

1 Opis algorytmu

1.1 Ogólne założenia

Naiwny klasyfikator bayesowski to algorytm, który wyznacza prawdopodobieństwa przynależności próbki do określonych klas. Predykcja wyniku opiera się na znajdywaniu klasy, dla której prawdopodobieństwo przynależności jest największe. Warunkiem koniecznym poprawności działania jest fakt, że poszczególne cechy będą od siebie niezależne.

Korzystając z ogólnego twierdzenia, mówiącego że:

$$P(Y = y_k|X = x) = \frac{P(X = x|Y = y_k)P(Y = y_k)}{P(X = x)}$$

można wyprowadzić ogólny wzór metody bayesowskiej zapisując go dla watości parametru x w postaci wektorowej:

$$P(Y = y_k|X = x_{[0,k]}) = \frac{P(Y = y_k)P(X = x_0|Y = y_k)P(X = x_1|Y = y_k)\dots P(X = x_k|Y = y_k)}{P(X = x)}$$

W przypadku konieczności tylko predykcji wyniku nie ma konieczności wyliczania mianownika.

1.2 Wyliczanie wartości

Obliczenie $P(Y = y_k)$ sprowadza się tylko do zbadania liczby elementów danej klasy w o całej próbce testowej.

Pozostałe prawdopodobieństwa warunkowe liczymy zakładając, że badana cecha ma określony rozkład prawdopodobieństwa. Na przykład w przypadku przyjęcia rozkładu normalnego:

$$P(X = x_k|Y = y_k) = \frac{1}{\sqrt{2\sigma^2}} e^{\frac{-(x_k - m)^2}{2\sigma^2}}$$

Aproksymacja parametrów rozkładu jest przeprowadzana dla zadanej próbki treningowej. Po podstawieniu wszystkich wartości do wzoru uzyskamy prawdopodobieństwo przynależności danego osobnika do klasy na podstawie jego wektora cech.

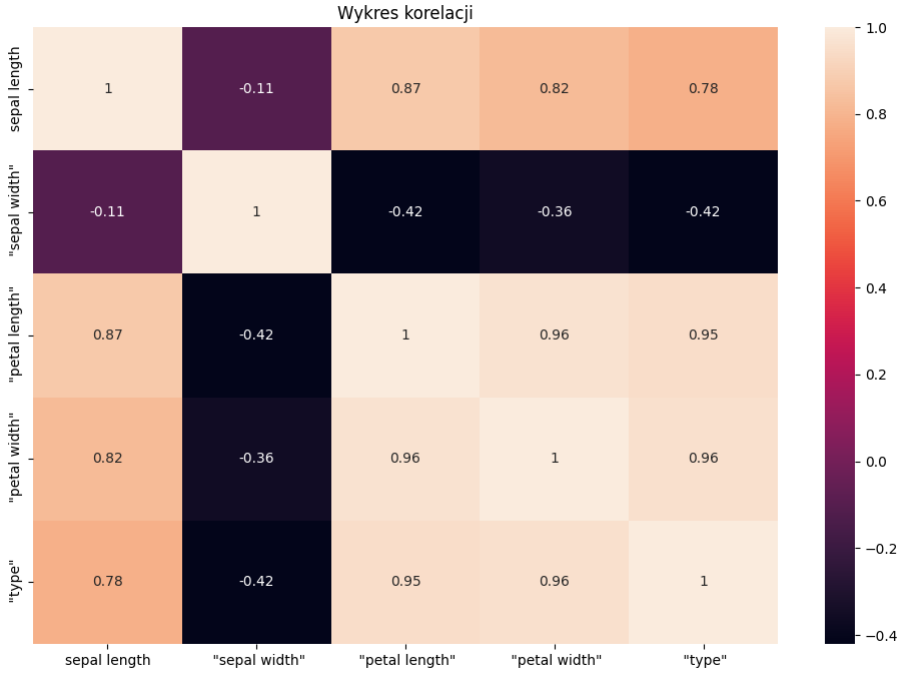
2 Planowane eksperymenty

Z racji, że algorytm ten nie posiada zmiennych hiperparametrów postaram się zaprezentować wpływ proporcji podziału zbioru danych na uzyskany rezultat. Dodatkowo, zaprezentuje, najwny charakter algorytmu, który zakłada, że poszczególne cechy są od siebie niezależne.

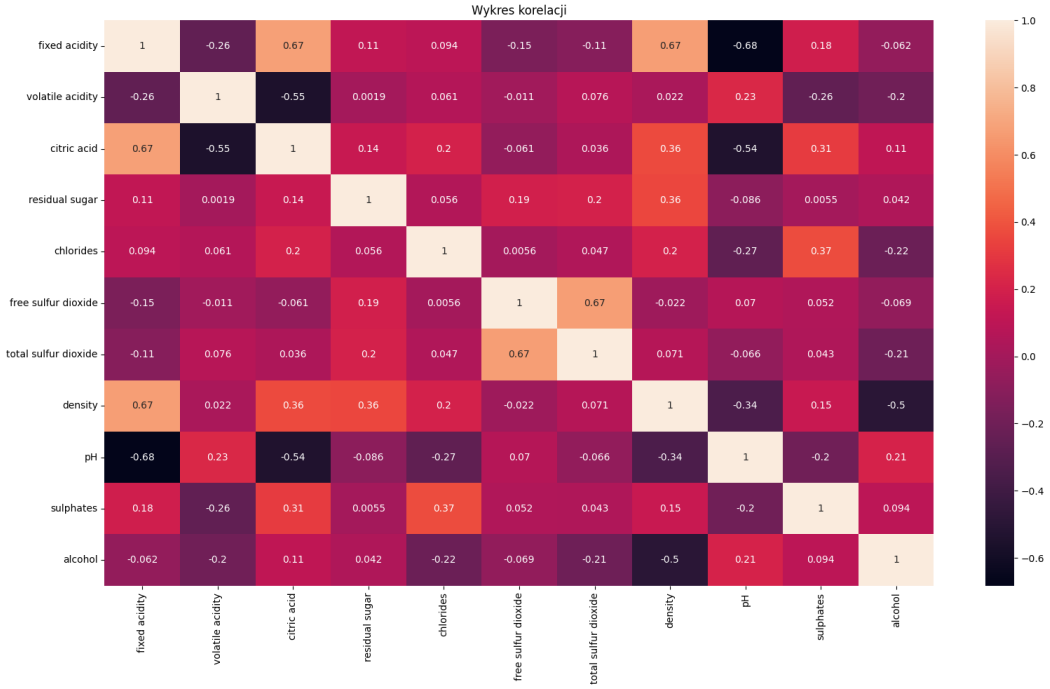
W tym celu spróbuję wykorzystać algorytm dla zbioru treningowego z zadania 4, a uzyskany wynik porównam z dokładnością algorytmu SVM.

Co ważne, do poprawności działania algorytm nie potrzebuje skalowania danych oraz dyskretyzacji wartości oczekiwanych. Ułatwia to znacząco działanie na przyjętych zbiorach.

3 Sprawdzenie założenia

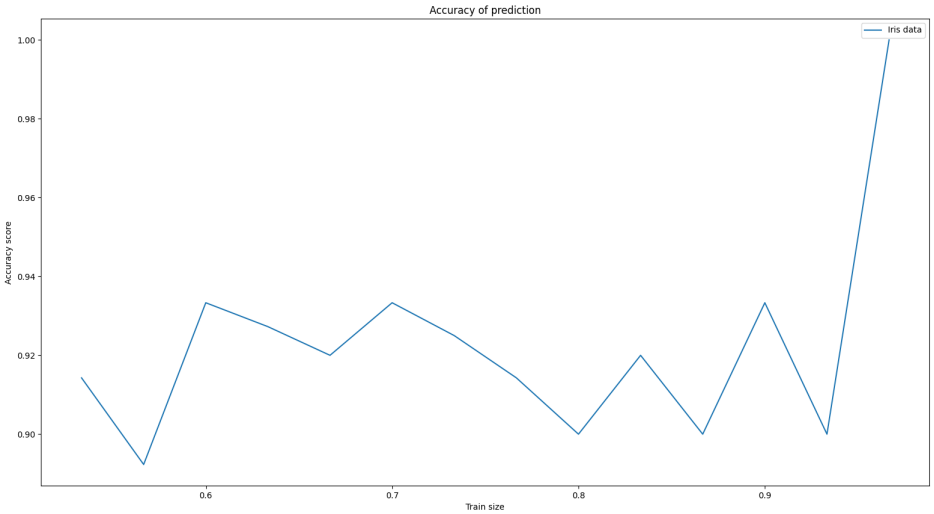


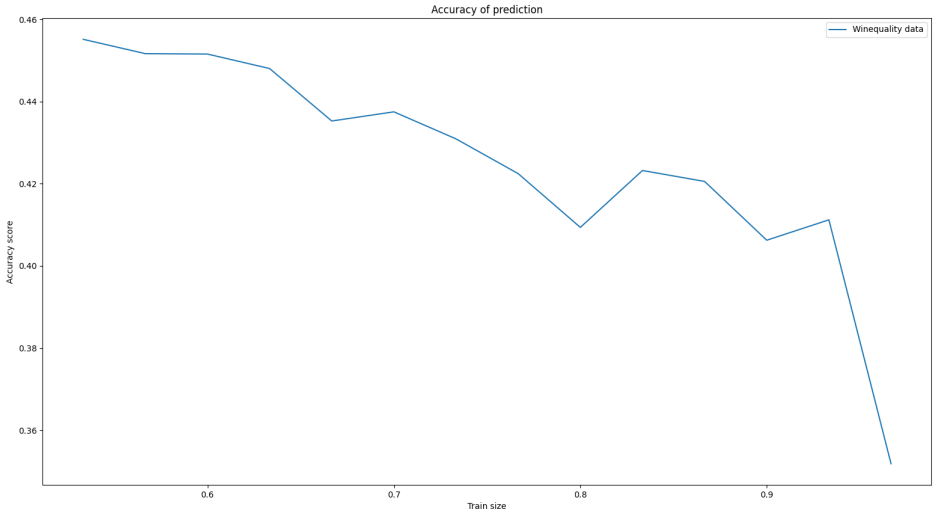
Niestety nie można założyć niezależności badanego zbioru. Ważny jest jednak fakt, że wszelkie badane cechy zachowują się podobnie tj. gdy jedna wartość rośnie większość z nich będzie zawsze rosnąć. Wyjątkiem jest szerokość kielicha, która zawsze będzie maleć.



W przypadku porównywanego zbioru win nie da się już zauważyć podobnej zależności. Najprawdopodobniej większa złożoność zbioru przełoży się na mniejszą dokładność algorytmu, ponieważ dane znacząco odbiegają od przyjętego założenia.

4 Rezultaty





Porównując oba wykresy można zauważyć, że algorytm znacznie lepiej działa dla zbioru "iris", w którym jego charakter ułatwia wykorzystanie bayasowskiego klasyfikatora.

W przypadku drugiego zbioru, dla za dużego zbioru treningowego następuje znaczące przeuczenie.

5 Wnioski

Udało się zaimplementować algorytm, który dla zadanego zbioru danych jest w stanie klasyfikować elementy z dokładnością ponad 90%.

Uzyskane rezultaty pokazują, że ten prosty algorytm jest w stanie bardzo dokładnie przewidywać proste zbiory.

Podczas eksperymentów pokazałem, że w przypadku cięższego zbioru napotyka on już większe problemy.