Patryk Bandyra

WSI

Ćwiczenie 4

Klasyfikacja

1. Opis zadania

Należało zaimplementować naiwny klasyfikator bayesowski bez użycia dodatkowych bibliotek i zastosować go do załączonego zbioru danych. Moim zadaniem była klasyfikacja 3 gatunków kosaćców (szczecinkowy (setosa), różnobarwny (versicolor) i wirginijski (virginica)) na podstawie długości i szerokości obu okółków ich okwiatów. Zbiór tworzyło 150 obserwacji, po 50 próbek dla każdej klasy. Link do zbioru danych: http://archive.ics.uci.edu/ml/datasets/Iris.

2. Opis algorytmu

Algorytm opiera się na Twierdzeniu Bayesa: Dla pary zmiennych losowych x i y spełniona jest równość:

$$P(y|x) = \frac{P(x|y)p(y)}{P(x)}$$

W przypadku klasyfikacji:

Zakładając, że y to przewidywana klasa, a $x_1, \ldots, x_n$ - atrybuty wejściowe i stosując tw. Bayesa, otrzymujemy:

$$\begin{aligned} P(y|x_1, \ldots, x_n) &= \frac{P(y, x_1, \ldots, x_n)}{P(x_1, \ldots, x_n)} \\ &= \frac{P(x_1, \ldots, x_n|y)P(y)}{P(x_1, \ldots, x_n)} \end{aligned}$$

Naiwny klasyfikator bayesowski:

- Do zbudowania modelu na podstawie tw. Bayesa potrzebujemy estymat:

$$P(x_1, \ldots, x_n|y)$$

- Jeśli („naiwnie") założymy, że atrybuty wejściowe są wzajemnie niezależne:

$$P(x_1, \ldots, x_n|y) = \prod_{i=1}^{n} P(x_i|y)$$

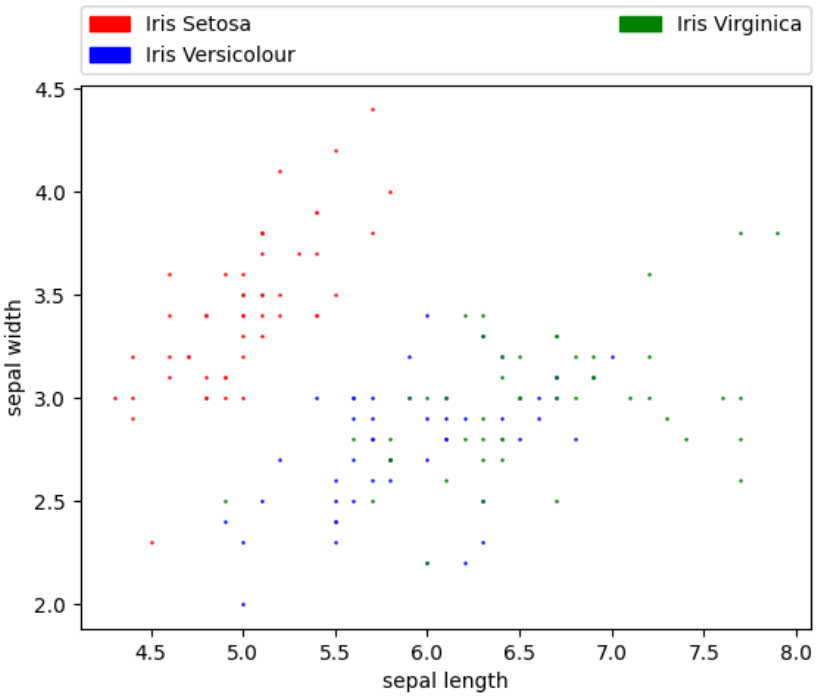
- Klasę dla danego $\mathbf{x} = [x_1, \ldots, x^n]^T$ wyliczamy:

$$\hat{y} = \arg\max_{y \in Y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

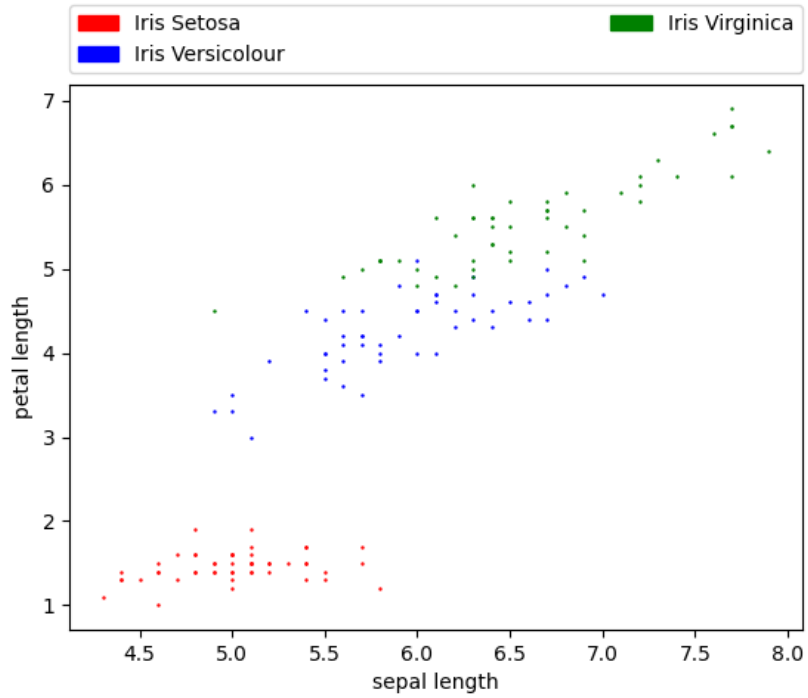
- Wartości $P(u)$ oraz $\forall_{i=1,\ldots,n} P(x_i|y)$ estymujemy na podstawie danych uczących.

3. Analiza zbioru danych

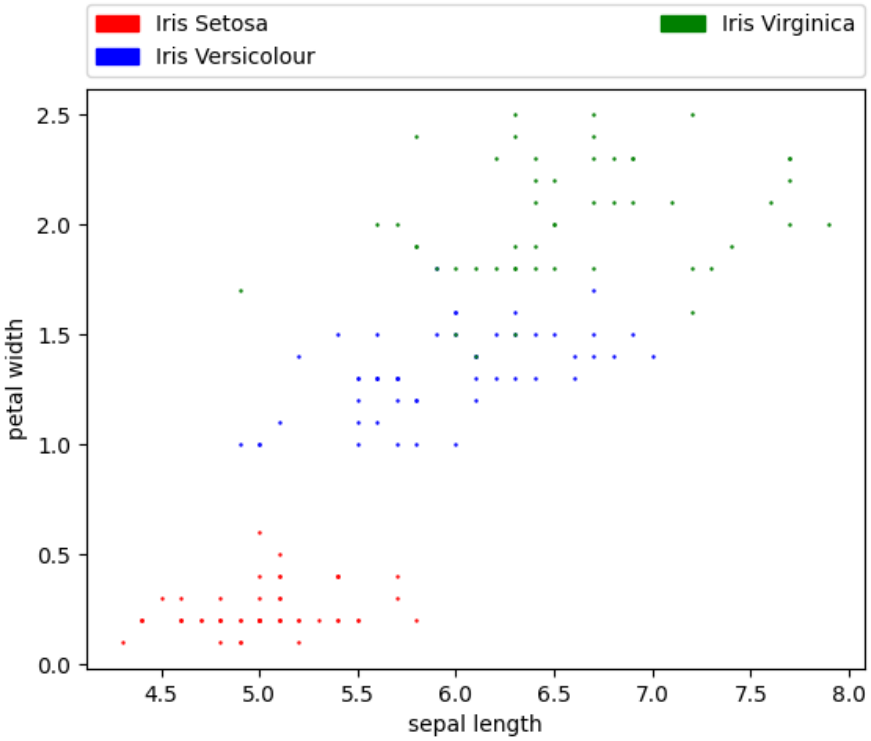
a) Iris Setosa separowalny liniowo od dwóch pozostałych klas



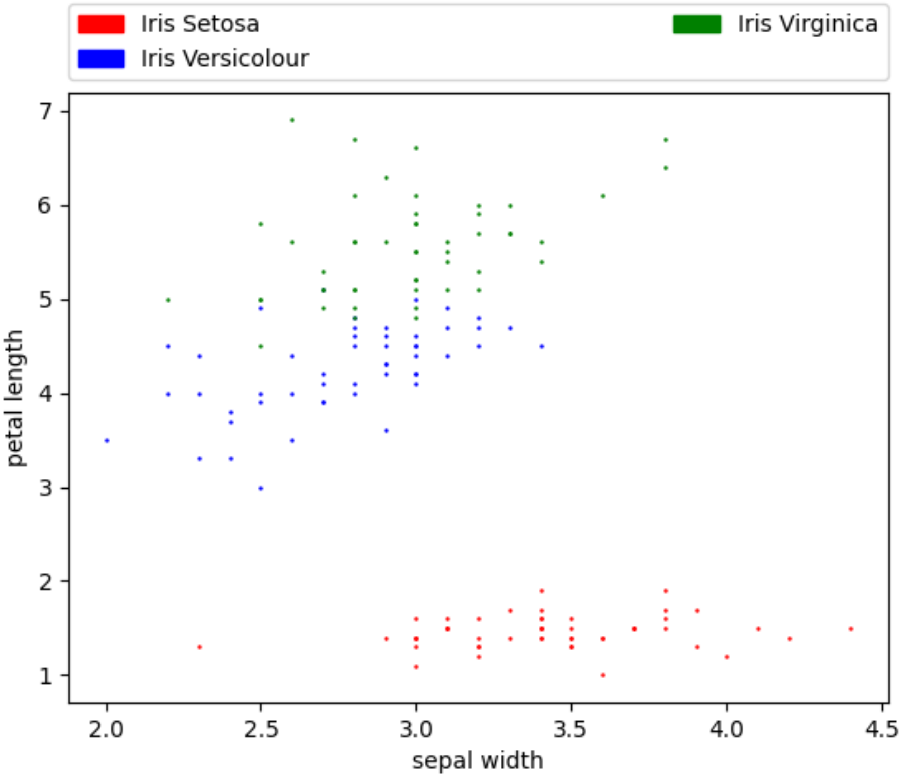
b) Iris Setosa separowalny liniowo od dwóch pozostałych klas



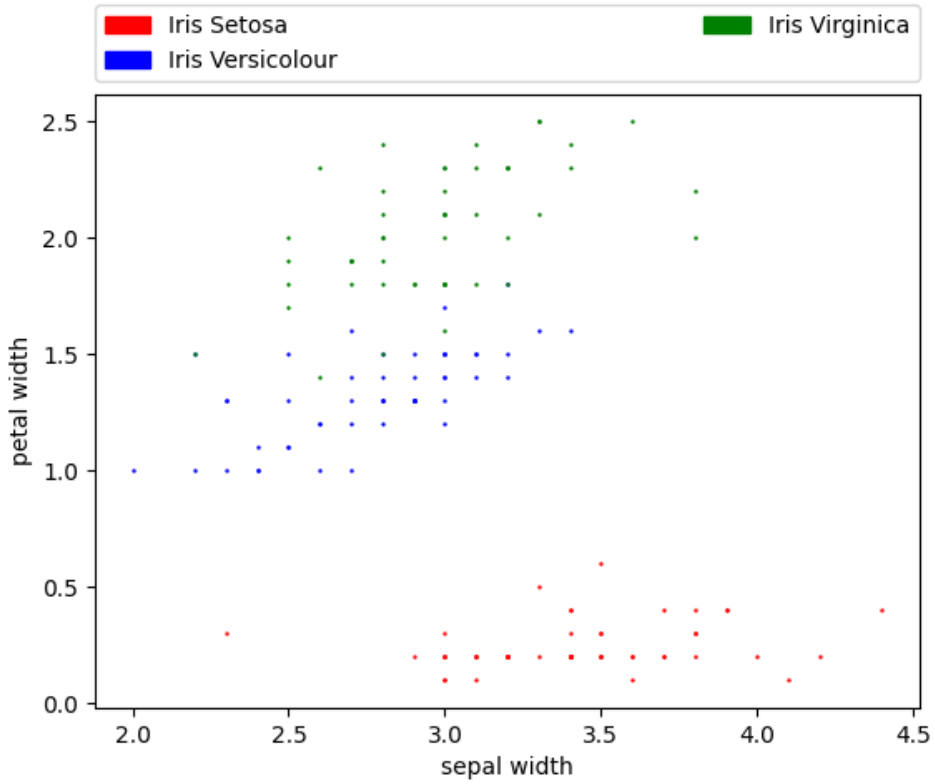
c) Iris Setosa separowalny liniowo od dwóch pozostałych klas



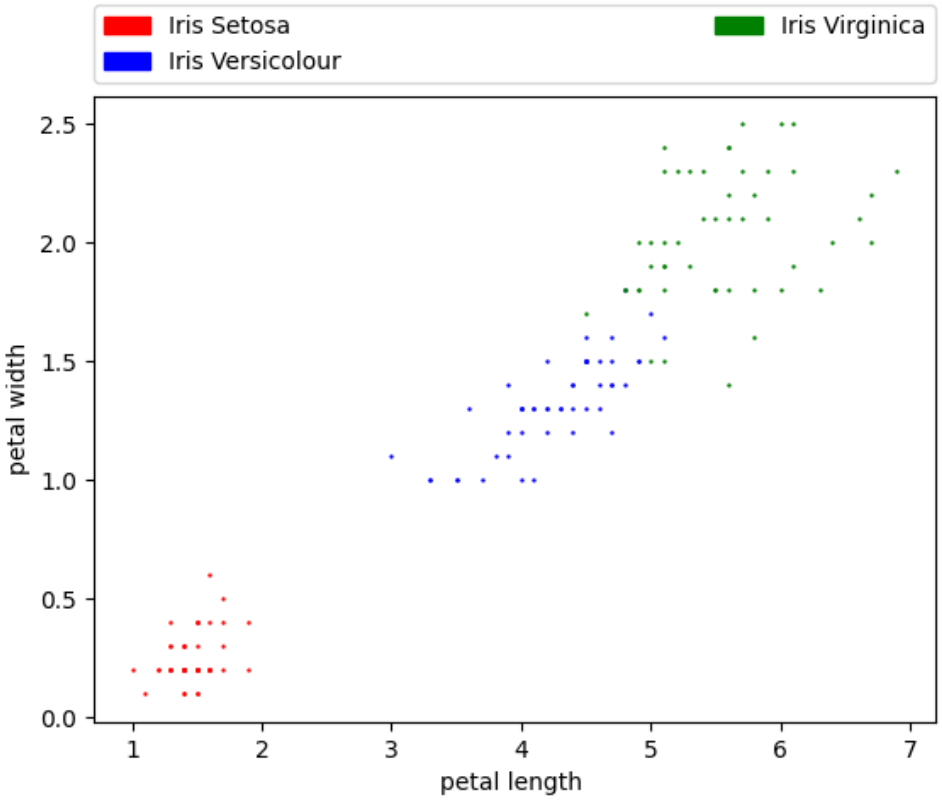
d) Iris Setosa separowalny liniowo od dwóch pozostałych klas



e) Iris Setosa separowalny liniowo od dwóch pozostałych klas



f) Iris Setosa separowalny liniowo od dwóch pozostałych klas

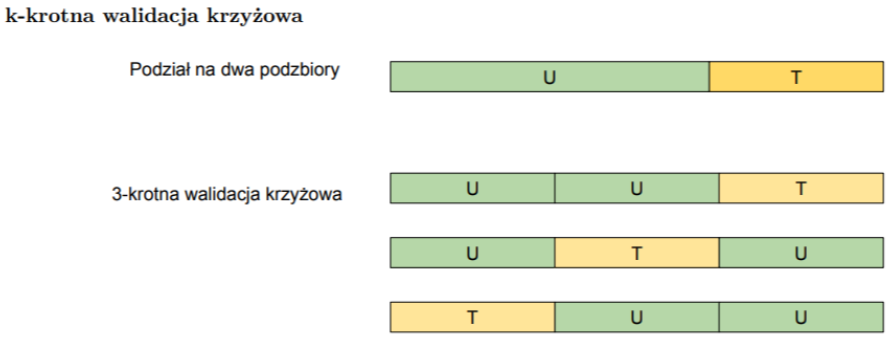


4. Opis i analiza programu

Program podzielony na kilka sekcji:

a) Analiza zbioru danych (funkcja *analyze data*) – wynik: wykresy przynależności do danej klasy w zależności od danych dwóch parametrów. Celem jest stwierdzenie czy dane klasy są liniowo separowalne w zależności od danych atrybutów obserwacji.

b) Klasyfikacja zbioru danych z użyciem k-krotnej walidacji krzyżowej (funkcja *run*) – wynik: porównanie przewidzianych klas z aktualnymi, lista wyników dla danej k-tej walidacji oraz średni wynik algorytmu. Jest to przypadek, gdy dane są losowo rozdysponowane pomiędzy k zbiorów, które w różnych wariantach wchodzą do zbioru uczącego i testowego.

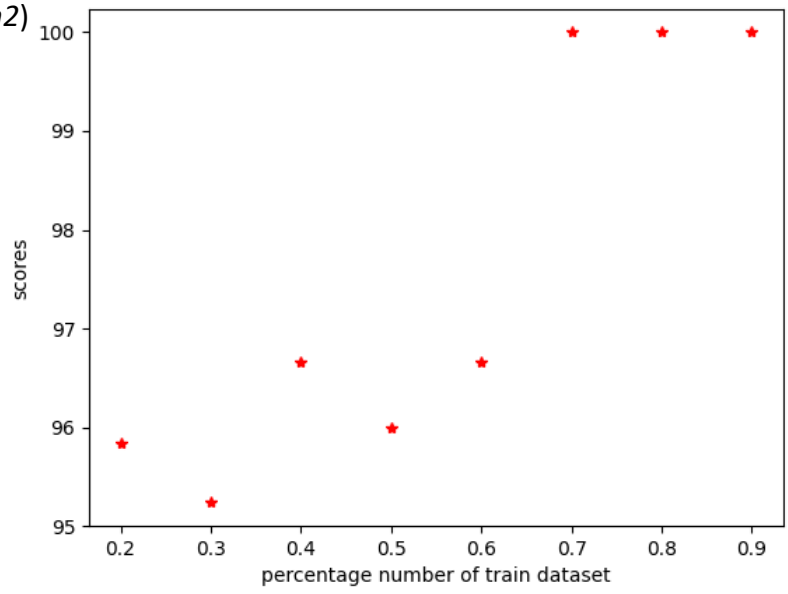


Przykładowy wynik programu dla k = 5:

```
Scores: [96.66666666666667, 100.0, 93.33333333333333, 90.0, 100.0]
Average accuracy: 96.0%
```

Efektywność modelu zależy od doboru liczby walidacji. Zbyt duża liczba walidacji wiąże się z przeuczeniem modelu (zbyt duży zbiór uczący, zbyt mały zbiór testowy), a zbyt mała z niedouczeniem.

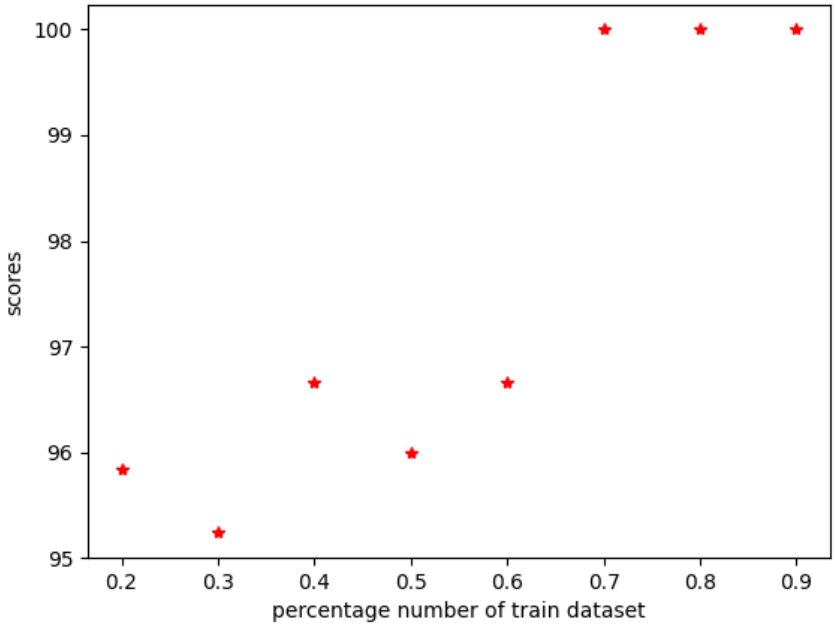
c) Testowanie wpływu proporcji zbioru trenującego do zbioru testowego na wyniki (*funkcja run2*)



Wykres dla zbiorów trenujących wybieranych na zasadzie: z każdej klasy weź x pierwszych obserwacji zgodnie z procentowym podziałem.

Z wykresu wynika, że w takiej sytuacji: im większy zbiór trenujący, tym lepsze wyniki.

- porównać efektywność modelu, kiedy zbiór trenujący jest na wstępie posortowany z przypadkiem, gdy dane zostaną specjalnie pomieszane



Pomieszanie już wybranego zbioru trenującego nie wpływa na efektywność modelu.