Dokumentacja zadania laboratoryjnego - badanie Naiwnego Klasyfikatora Bayesa

Krzysztof Wyrzykowski, nr indeksu 331455

26 stycznia 2025

1 Algorytm Naiwnego Klasyfikatora Bayesa

1.1 Opis działania algorytmu

Jest to algorytm służący do rozwiązywania problemu klasyfikacji oparty na twierdzeniu Bayesa. Jego naiwność wynika z przyjęcia założenia, że badane cechy opisujące dane są od siebie warunkowo niezależne. Działanie algorytmu opiera się na wyznaczeniu prawdopodobieństwa przynależności danej próbki danych do każdej z badanych klas $P(C|X)$, które wyraża się wzorem Bayesa:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)},$$

gdzie:

- $P(C|X)$ – prawdopodobieństwo a posteriori, że dane X należą do klasy C,
- $P(X|C)$ – prawdopodobieństwo, że klasa C generuje dane X,
- $P(C)$ – prawdopodobieństwo a priori klasy C,
- $P(X)$ – ogólne prawdopodobieństwo wystąpienia danych X.

W związku z tym, że $P(X)$ jest stałe dla wszystkich rozważanych klas, możemy je pominąć, gdyż nie wpływa na wynik zadania maksymalizacji. Zatem dla każdej klasy obliczamy wynik wyrażenia proporcjonalnego:

$$P(C \mid X) \propto P(X \mid C) \cdot P(C)$$

$P(X \mid C)$ wyraża się jako iloczyn n cech:

$$P(X \mid C) = P(x_1 \mid C) \cdot P(x_2 \mid C) \cdot \ldots \cdot P(x_n \mid C)$$

Przewidywanie klasy to wybór klasy C o najwyższym prawdopodobieństwie:

$$\hat{C} = \arg\max_C \left(P(X \mid C) \cdot P(C) \right).$$

W celu uniknięcia problemów z utratą precyzji wynikającą z operacji mnożenia na bardzo małych liczbach, zamieniamy iloczyn na sumę logarytmów:

$$\hat{C} = \arg\max_C \left(\log P(X \mid C) + \log P(C) \right).$$

1.2 Funkcja prawdopodobieństwa

Dobór funkcji prawdopodobieństwo zależy od rozwiązywanego problemu. Najczęściej używane rozkłady to:

- rozklad Gaussa - wykorzystywany dla cech ciągłych
- rozklad wielomianowy - odpowiedni dla cech opisanych rozkładem dyskretnym
- rozklad Bernouliego - dla cech binarnych

W tej implementacji wykorzystano rozkład Gaussa, a dokładnie jego funkcję gęstości, jednocześnie umożliwiają zastosowanie innej funkcji prawdopodobieństwa.

2 Planowane eksperymenty numeryczne

2.1 Cel eksperymentów

Celem eksperymentów było zbadanie dokładności Naiwnego Klasyfikatora Bayesa.

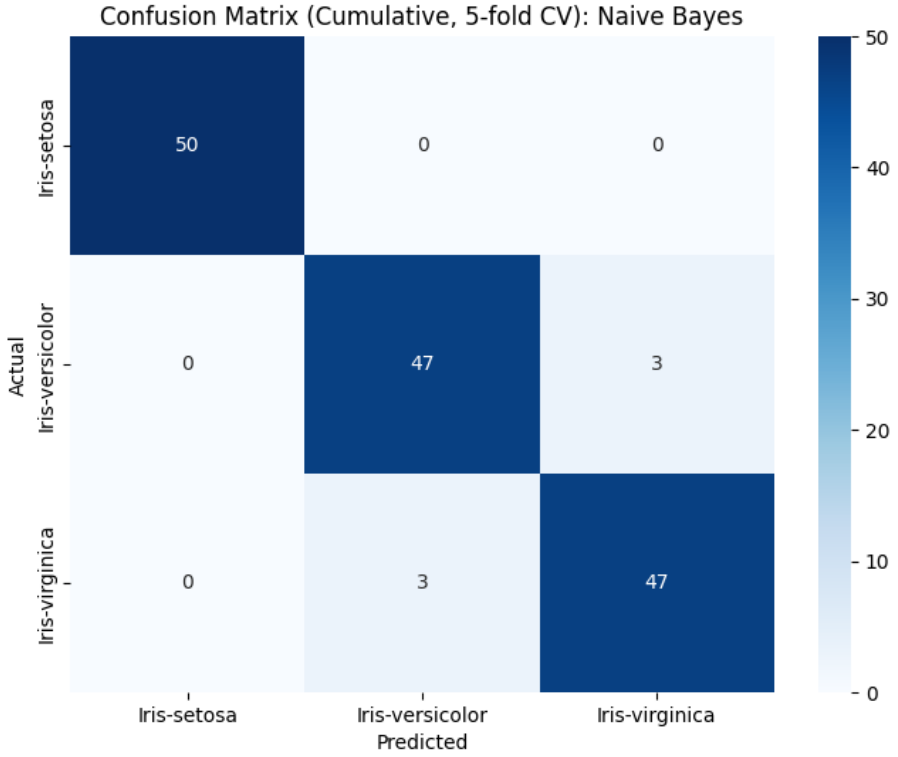
2.2 Przebieg badań

W celu sprawdzenia dokładności klasyfikatora zdecydowałem się na porównanie jego działania z dwoma innymi klasyfikatorami zaimplementowanymi w bibliotece `scikit-learn`: `DecisionTreeClassifier` oraz `KNeighborsClassifier`. Do oceny posłużyłem się metodą walidacji krzyżowej, która polega na podziale zbioru danych na k grup (foldów). A następnie przeprowadzeniu k treningów, w których zbiorem testowym jest jedna z grup, a pozostałe są zbiorem treningowym. Badany zbiór posiadał 150 próbek, które zostały podzielone na 5 foldów po 30 próbek.

2.3 Zbiór danych

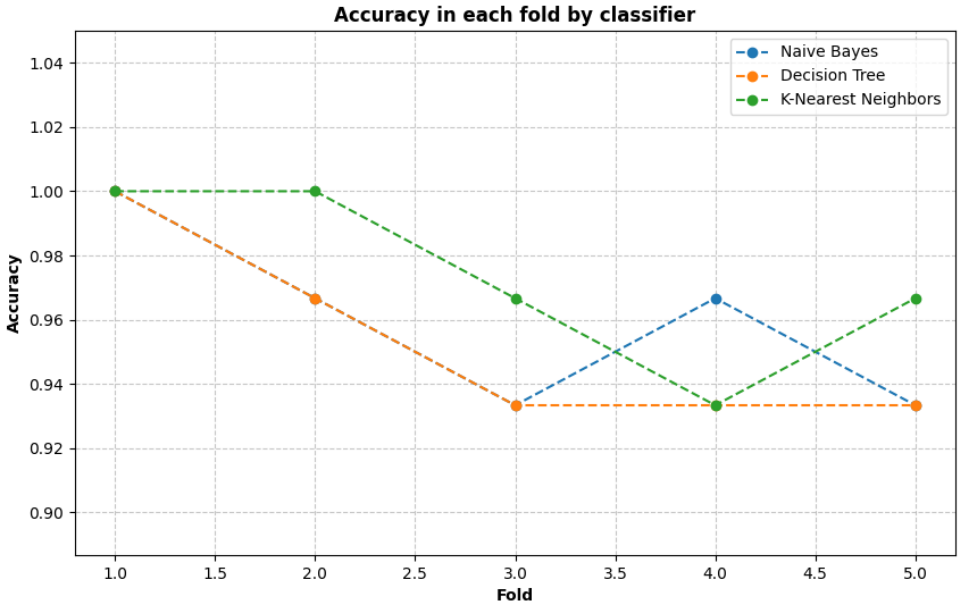
Algorytm został przetestowany na zbiorze Iris Data Set, zawierającym 150 próbek z wartościami 4 zmierzonych parametrów Irysów, należących do 3 różnych gatunków. W badanym zbiorze jedna klasa jest liniowo separowalna od dwóch pozostałych, które nie są wzajemnie liniowo separowalne.

3 Wyniki badań



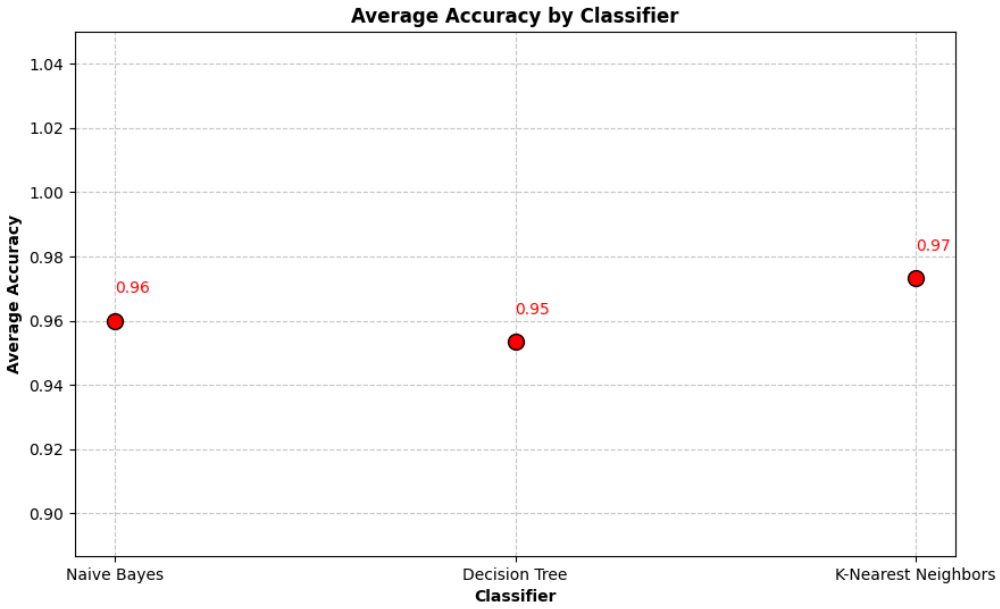
Rysunek 1: Macierz pomyłek dla klasyfikatora Naiwnego Bayesa.

Z macierzy wynika, że klasyfikator poradził sobie w 100% przypadków z klasyfikacją próbek należących do liniowo-separowalnej klasy. Jedynie w 6% przypadków nie udało mu się poprawnie przewidzieć klasy w nieseparowalnych liniowo klasach.



Rysunek 2: Dokładność klasyfikatora w poszczególnych foldach.

Dla każdego foldu wyniki badanego klasyfikatora były porównywalne z innymi klasyfikatorami.



Rysunek 3: Porównanie średniej dokładności między klasyfikatorami.

Różnice średnich wartości dokładności między klasyfikatorami są pomijalnie małe i znajdują się w granicach błędu statystycznego wynikającego z wyboru zbioru danych.

4 Wnioski

Z badań wynika, że Naiwny Klasyfikator Bayesa poradził sobie z zadanym problemem równie dobrze co inne testowane klasyfikatory. Osiągnął dokładność wynoszącą 96%, która jest bardzo dobrym wynikiem zadania klasyfikacji. Warto zwrócić uwagę na 100% skuteczność w klasyfikacji próbek z liniowo-separowalnego zbioru, na podstawie której można wnioskować o wysokiej skuteczności klasyfikatora w prostych zadaniach. W próbkach należących do pozostałych klas również osiągnął wysoką skuteczność. Na podstawie badań wnioskuję, że Naiwny Klasyfikator Bayesa wykazuje bardzo wysoką skuteczność w prostych zadaniach klasyfikacyjnych.