Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Wstęp do Sztucznej Inteligencji
Semestr 24L
Sprawozdanie z ćwiczenia nr 7

Naiwny klasyfikator Bayesowski

Mikołaj Wewiór

Warszawa,
11 VI 2024

1. Opis problemu

Celem ćwiczenia było zaimplementowanie naiwnego klasyfikatora bayesowskiego oraz porównanie go z gotowymi implementacjami z biblioteki `scikitlearn` klasyfikatora SVM oraz drzewa decyzyjnego do zadania klasyfikacji irysów. Do porównania i oceny jakości wykoorzystano metryki: *accuracy*, *precision*, *recall* oraz *F1*.

2. Ładowanie zbioru danych

Klayfikowane były kwiaty irysów, mianowicie trzy gatunki - *Setosa*, *Versicolor* i *Virginica*. Dane były w postaci numerycznych parametrów odpowiadającym fizycznym właściwościom kwiatów. Pobrany zbiór danych zawierał 150 próbek. Dane również zostały pobrane z biblioteki `sklearn`, a następnie próbki zostały pomieszane. Do ewaluacji i zbadania modelu wykorzystano walidację 5-krotną, z zapewnieniem, że podziały były stratyfikowane (tzn. w każdej próbce danych uczących było po równo elementów z każdej z klas)

W celu powtarzalności przeprowadzanych eksperymentów skorzystano z ziarna generacji liczb pseudolosowych. Wykorzystano je również w dalszej części jako `random_state` w klasyfikatorach. Wykorzystanym ziarnem było 318 407.

3. Implementacja

Prawdopodobieństwa, wartości średnie oraz odchylenia warunkowe zostały zaimplementowane jako tablica (macierz) o liczbie wierszy równej liczbie klas i liczbie kolumn odpowiadającej każdej z cech, które reprezentują dane. W przypadku tego zadania jest są to 3 wiersze oraz 4 kolumny. Klasa Klasyfikatora dziedziczy bo klasach z biblioteki `sklearn`: `BaseEstimator`, `ClassifierMixin`. Dzięki temu można było skorzystać z tej implementacji tak jak z pozostałych klasyfikatorów z wspomnianej biblioteki. To pozwoliło w dużo łatwiejszy i bardziej odpowiadający sposób zmierzyć ze sobą różne klasyfikatory i porównać jakość ich działania.

Przy wyborze najbardziej prawdopodobnej klasy w metodzie `predict`, wykorzystano sumę logarytmów prawdopodobieństw zamiast ich iloczynu. To dało o wiele lepsze wyniki, ponieważ iloczyn bardzo małych wartości (rzędu $1e-300$, powodowały niestabilność numeryczną - zwracały wartość zero. Zastąpienie tego logarytmami pozwoliło sumować wyniki, dzięki czemu rozwiązanie nie traciło informacji.

4. Porównanie klasyfikatorów

Do porównania skorzystano z następujących klasyfikatorów:

— **SVM** (funkcja jądra: "rbf:, siła regularyzacji: 0.1, maksymalna liczba iteracji: 25),
— **Drzewo decyzyjne** (kryterium podziału: entropia, technika podziału: najlepsza cecha, maksymalna głębokość: 4),
— **Zaimplementowany naiwny klasyfikator Bayesa**,
— **GNB** (Gaussowski naiwny Bayes) z biblioteki `sklearn`.

Wyniki testu prezentują się następująco:

Miara Jakości	SVM	Drzewo	Naiwny Bayes	GNB
accuracy	95.333 ± 3.399	93.666 ± 5.778	92.000 ± 7.483	94.666 ± 3.399
precision	95.724 ± 3.260	93.650 ± 5.569	93.441 ± 5.972	95.219 ± 3.399
recall	95.333 ± 3.399	93.666 ± 5.778	92.000 ± 7.483	94.666 ± 3.399
f1	95.312 ± 3.414	93.309 ± 5.592	91.701 ± 7.834	94.633 ± 3.416

Tab. 1: Porównanie klasyfikatorów. Wartości w postaci średnia ± odchylenie.

Oprócz trzech opisanych wcześniej klasyfikatorów do badania dodany został Gaussowski naiwny klasyfikator bayesowski. Jest to gotowe rozwiązanie z `sklearn`, analogiczne do implementowanego zadania.

Z danych wyraźnie widać, że najlepszy okazał się klasyfikator SVM. Następnym jest GNB, za nim drzewo decyzyjne i na końcu zaimplementowany naiwny bayes. Zaistniała rozbieżność między omawianą implementacją a rozwiązaniem z biblioteki, może wynikać z niezastosowania w wyliczaniu prawdopodobieństwa warunkowego logarytmów, tak jak zostało to zrealizowane w bibliotece.

Każdy z klasyfikatorów cechuje się wysoką skutecznością. Niestety zaimplementowane rozwiązanie ma ją najniższą z porównywanych opcji oraz posiada największe odchylenie standardowe.

Warto dodać że klasyfikator naiwengo bayesa jest najprostszym z powyższych w implementacji oraz posiada najmniej parametrów. Dzięki temu jest to proste i rozwiązanie, które daje naprawdę satysfakcjonujące wyniki.