Wydział Elektroniki
i Technik Informacyjnych
POLITECHNIKA WARSZAWSKA

Wstęp do Sztucznej Inteligencji

Ćwiczenie nr 7:
Modele bayesowskie

Kaczmarski Robert
293377

Warszawa 2022

Polecenie

Zadanie polega na implementacji naiwnego klasyfikatora bayesowskiego oraz zastosowania go do klasyfikacji trzech gatunków kosaćców.

Użyte narzędzia

Do wykonania tego zadania został użyty Python w wersji 3.9.9 oraz następujące biblioteki i moduły zawarte w pliku requirements.txt:

- numpy==1.22.1
- matplotlib==3.5.1
- seaborn==0.11.2
- sklearn.metrics

Analiza zbioru danych

Zbiór danych zawiera 150 obserwacji po 50 dla każdego gatunku. Każdy rekord zawiera 5 atrybutów:

-szerokość liścia,

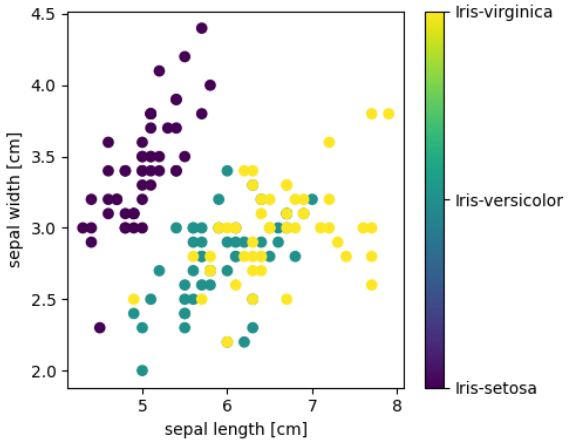
-długość liścia,

-szerokość płatka,

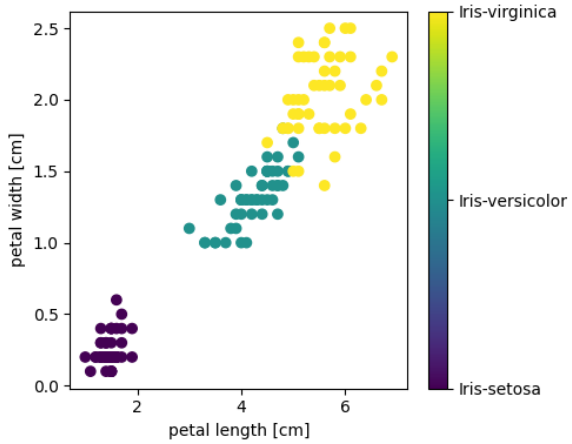
-długość płatka,

-klase np. (Iris-Setosa, Iris-Versicolour, Iris-Virginica).

Na rysunkach 1-2 została przeprowadzona analiza mająca na celu sprawdzenie, czy na podstawie obserwacji można stwierdzić istnienie poszukiwanych klas.



Rys. 1. Zależność gatunku Irysa od wymiarów liscia.



Rys. 2. Zależność gatunku Irysa od wymiarów płatka.

Na podstawie rysunków 1-2 widać dużą zależność gatunku od rozmiarów płatka. Na podstawie długości i szerokości płatka można łatwo stwierdzić, czy dany gatunek to *setosa*. Trudniej może być z rozróżnieniem gatunków *virginica* i *versicolor*, ponieważ chmury punktów należących do tych gatunków częściowo nakładają się na siebie.

Analiza algorytmu

Zadany zbiór danych posiada atrybuty ciągłe, dlatego do twierdzenia Bayesa został użyty rozkład normalny gęstości prawdopodobieństwa. Dla każdej instancji klasy wyznaczane jest prawdopodobieństwo warunkowe zaobserwowania zdarzenia Y=k pod warunkiem, że zaobserwowano zdarzenia $X_1 \ldots X_n$, zgodnie ze wzorem:

$$P(Y = k|X_1 \ldots X_n) = \frac{P(X_1|Y = k) \ldots * P(X_n|Y = k) * P(Y = k)}{P(X_1) \ldots * P(X_n)}$$

Przy czym dla zadanego zbioru danych X to obserwacje {sepal_length, sepal_width, petal_length, petal_width}, Y to klasa, a k to gatunki kosaćców {Iris-setosa, Iris-versicolor, Iris-virginica}

Wyniki

W zaimplementowanym algorytmie został zbadany wpływ różnych parametrów wejściowych. Wyniki zostały przedstawione w tabelach 1-2. Zbiór danych był wstępnie uporządkowany według gatunku. Pierwsze 50 obserwacji należy do gatunku *Iris*-setosa, kolejne 50 to *Iris-versicolor*, a ostatnie 50 to *Iris-virginica*. W przypadku badania zbioru uporządkowanego nie ma sensu przeprowadzać wiele testów, ponieważ w tym przypadku nie ma losowości. Przy badaniu wstępnie pomieszanego zbioru danych dla każdego testu ustawione ziarno losowe odpowiadało numerowi testu, tzn. od 0 do 99.

Tabela 1. Badanie wpływu różnych proporcji zbiorów na wynik klasyfikacji dla zbioru uporządkowanego.

Proporcje	Pomieszany	Dokładność[%]
0,1 / 0,9	Nie	25,93
0,2 / 0,8	Nie	16,67
0,3 / 0,7	Nie	4,76
0,4 / 0,6	Nie	44,44
0,5 / 0,5	Nie	33,33
0,6 / 0,4	Nie	16,67
0,7 / 0,3	Nie	77,78
0,8 / 0,2	Nie	93,33
0,9 / 0,1	Nie	100,00

Dla uporządkowanego zbioru danych, gdy zbiór trenujący jest stosunkowo mały, widać oznaki niedouczenia się modelu. Przy zbiorze trenującym zawierającym 90%*150=135 obserwacji dokładność modelu osiągnęła 100%.

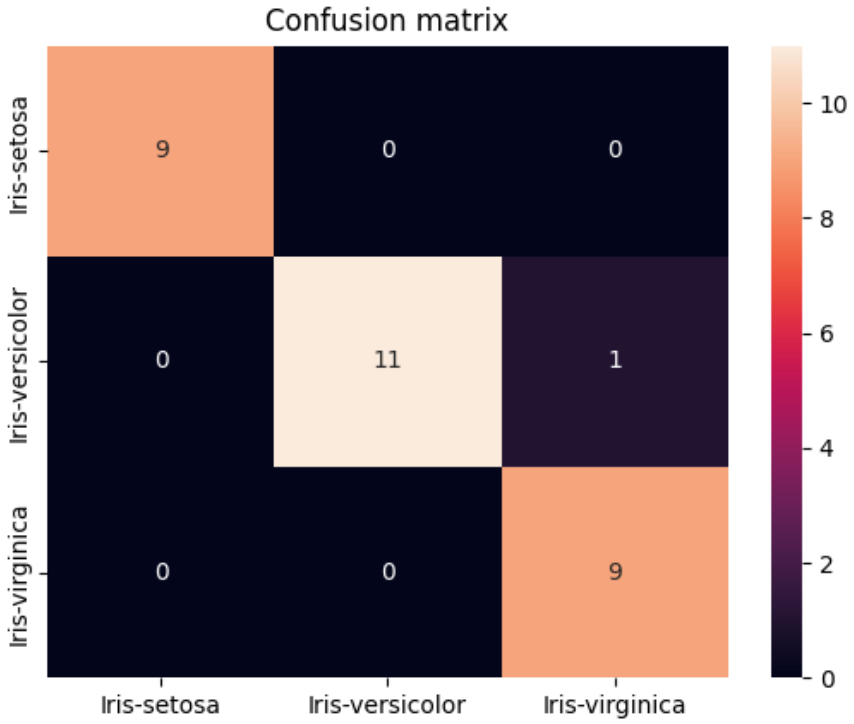
Tabela 2. Badanie wpływu różnych proporcji zbiorów na wynik klasyfikacji przy wstępnie pomieszanym zbiorze.

Liczba testów	Proporcje trenujący / testujący	Pomieszany	Minimum[%]	Średnia[%]	Maksimum[%]	Odchylenie[%]
100	0,1 / 0,9	Tak	31,11	84,09	97,04	305,82
100	0,2 / 0,8	Tak	85,00	94,43	98,33	5,55
100	0,3 / 0,7	Tak	91,43	94,98	99,05	2,48
100	0,4 / 0,6	Tak	90,00	94,93	98,89	3,22
100	0,5 / 0,5	Tak	90,67	95,12	100,00	4,13
100	0,6 / 0,4	Tak	90,00	95,20	100,00	4,96

100	0,7 / 0,3	Tak	86,67	95,16	100,00	7,74
100	0,8 / 0,2	Tak	86,67	95,60	100,00	10,64
100	0,9 / 0,1	Tak	80,00	94,80	100,00	31,63

Dla wstępnie pomieszanego zbioru danych średnia dokładność utrzymywała się na poziomie ok. 95% z wyjątkiem podziału na proporcje 0,1 / 0,9. Zwiększanie zbioru trenującego od 50% pozwala uzyskać mniejszą wartość minimum oraz wartość maksimum równą 100%.

Przykładowa macierz błędów została zamieszczona na rysunku 3.



Rys. 3. Macierz błędów dla ratio=0,8, shuffle=True, random.seed(0).

Wyznaczone zostały miary jakości klasyfikacji.

Tabela 1. Miary jakości klasyfikacji dla ratio=0,8, shuffle=True, random.seed(0).

	Iris-setosa	Iris-versicolor	Iris-virginica
Precyzja	100%	100%	90%
Czułość	100%	91,67%	100%
Dokładność	96,67%		

W badanym przykładzie widać, że jedyny błąd przy klasyfikacji wystąpił między gatunkami *versicolor* a *virginica*, co zgadzałoby się ze wstępnie przeprowadzoną analizą danych.

Podsumowanie

Udało się zaimplementować naiwny klasyfikator Bayesa z całkiem dobrymi wynikami. Następnym krokiem będzie nauka do egzaminu z WSI.