Mikalai Stelmakh, 316951

WSI lab 7 – Modele Bayesowskie

Zadanie

Zaimplementować naiwny klasyfikator bayesowski bez użycia dodatkowych bibliotek i zastosować go do zbadania załączonego zbioru danych.

Zbior

Zadanie klasyfikacji 3 odmian ziaren pszenicy Kama, Rosa i Canadian na podstawie ich wielkości geometrycznych. Zbiór tworzy 210 próbek, w skład których wchodzą 3 grupy 70 elementowe.

Uruchomienie programu

python3.10 NaiveBayes.py [-h] [--test_size TEST_SIZE] [--shuffle] file

Żeby dowiedzieć się więcej na temat argumentów należy użyć flagi "-h".

Wyniki

Po jednokrotnym uruchomieniu programu uzyskano następujące wyniki:

Macierz błędów

Spodziewana/Otrzymana	Kama	Rosa	Canadian
Kama	**9**	2	2
Rosa	3	**15**	0
Canadian	1	0	**10**

Klasa	Kama	Rosa	Canadian
Precyzja	0.69	0.88	0.83
Dokładność	0.8		
Czułość	0.7	0.83	0.9

Algorytm daje stosunkowo dobre wyniki, chociaż mogło by być lepiej, jeśli by nie było zależnych wartości obserwacji (pole, obwód, szerokość/długość ziarna).

Badanie wpływu parametrów

Sortowanie zbioru

Wyniki programu, kiedy zbiór jest na wstępie posortowany:

Klasa	Kama	Rosa	Canadian
Precyzja	0	1	0
Dokładność	0.26		
Czułość	0	0.8	0

Jak widać, wyniki są niezadowalające. Ponieważ zbiór trenujący zawiera wszystkie osbserwacje klasy „Kama” oraz większość obserwacji klasy „Rosa”, algorytm nie jest nauczony na klasyfikowanie klasy „Canadian” oraz nie jest testowany na klasie „Kama”. Można zmienić proporcje zbiorów testowych i trenujących tak, aby zbiór testowy zawierał kilka obserwacji klasy „Kama”, ale wtedy algorytm nie będzie nauczony na klasyfikowanie obserwacji klas „Rosa” i „Canadian”.

W celu obejścia tego problemu warto ten zbiór na początku pomieszać.

Wyniki wykonania programu, kiedy zbiór jest na wstępie pomieszany (w tabeli umieszczone są średnie dane po 50 uruchomieniach programu):

Klasa	Kama	Rosa	Canadian
Precyzja	0.87	0.94	0.86
Dokładność	0.89		
Czułość	0.82	0.93	0.95

Otrzymane wyniki są znacznie lepsze niż na posortowanym zbiorze, ponieważ dla każdego podziału zbioru na zbiory testowe/trenujące zawsze znajdzie się kilka obserwacji dla każdej klasy w obu tych zbiorów.

Podział zbioru na zbiory testowe/trenujące

Klasa	Kama				Rosa				Canadian			
Proporcja	0.1	0.3	0.5	0.9	0.1	0.3	0.5	0.9	0.1	0.3	0.5	0.9
Precyzja	0.9	0.86	0.87	0.23	0.93	0.95	0.95	0.08	0.89	0.9	0.89	0.006
Czułość	0.83	0.85	0.83	0.72	0.95	0.92	0.92	0.26	0.95	0.9	0.95	0.02

Proporcja	0.1	0.3	0.5	0.9
Dokładność	0.91	0.9	0.9	0.33

Oczywiste jest to, że przy zmniejszeniu rozmiaru zbioru testowego i zwiększeniu trenującego, dokładność klasyfikacji będzie rosła, zaś w przeciwnym przypadku, przy zmniejszeniu zbioru trenującego liczba obserwacji może być niedostateczna do nauczenia modelu, co może skutkować niską dokładnością klasyfikacji.