Wstęp do sztucznej inteligencji

Laboratorium piąte - modele bayesowskie

Krystian Kamiński nr 304013

Polecenie:

Tematem piątych ćwiczeń są modele bayesowskie. Państwa zadaniem będzie zaimplementować naiwny klasyfikator bayesowski i zastosować go do zbadania załączonego zbioru danych.
Osoby z nazwiskami od G od K 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.
http://archive.ics.uci.edu/ml/datasets/seeds

Założenia:

- w celu wyznaczenia prawdopodobieństw dla danej próbki dla wszystkich klas, pominę część prawdopodobieństwa P(data)
- plik tekstowy z danymi zawiera informacje o 210 próbkach, każda z nich posiada 7 parametrów oraz typ nasiona
  Parametrami są:
  1. powierzchnia A,
  2. obwód P,
  3. zwięzłość C = 4 * pi * A / P ^ 2,
  4. długość ziarna,
  5. szerokość ziarna ,
  6. współczynnik asymetrii
  7. długość rowka jądra.

  Wszystkie te parametry były rzeczywistymi wartościami ciągłymi.

Biblioteki:

- math - wykorzystanie stałych {pi, e}
- matplotlib - używam do tworzenia wykresów
- operator - używam w celu posortowania listy list

Implementacja:

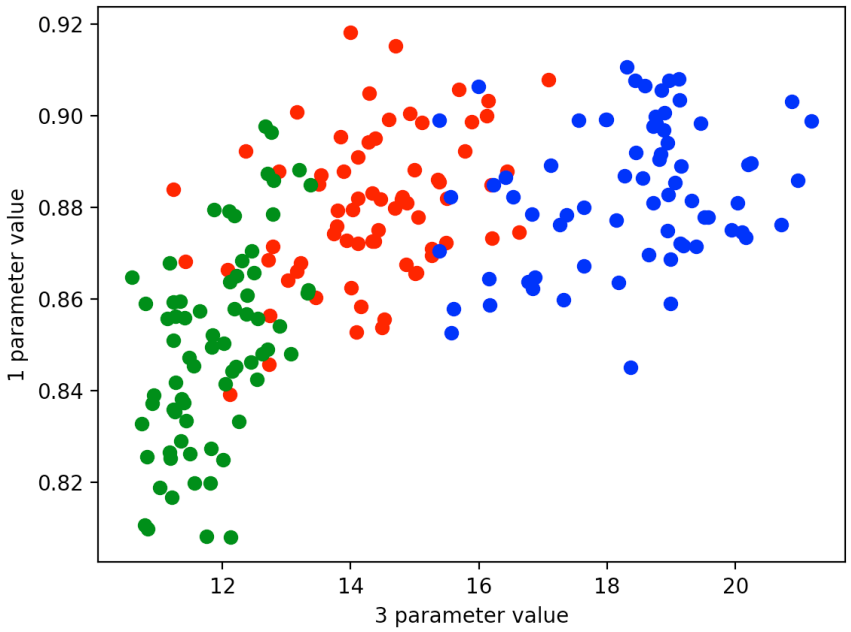
- open_file - wczytywanie danych z pliku
- show_graph - przedstawienie wykresu wybranego zbioru danych
- avg - obliczanie średniej
- stdev - obliczanie odchylenia standardowego
- probability - obliczanie funkcji gęstości prawdopodobieństwa na podstawie wzoru:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

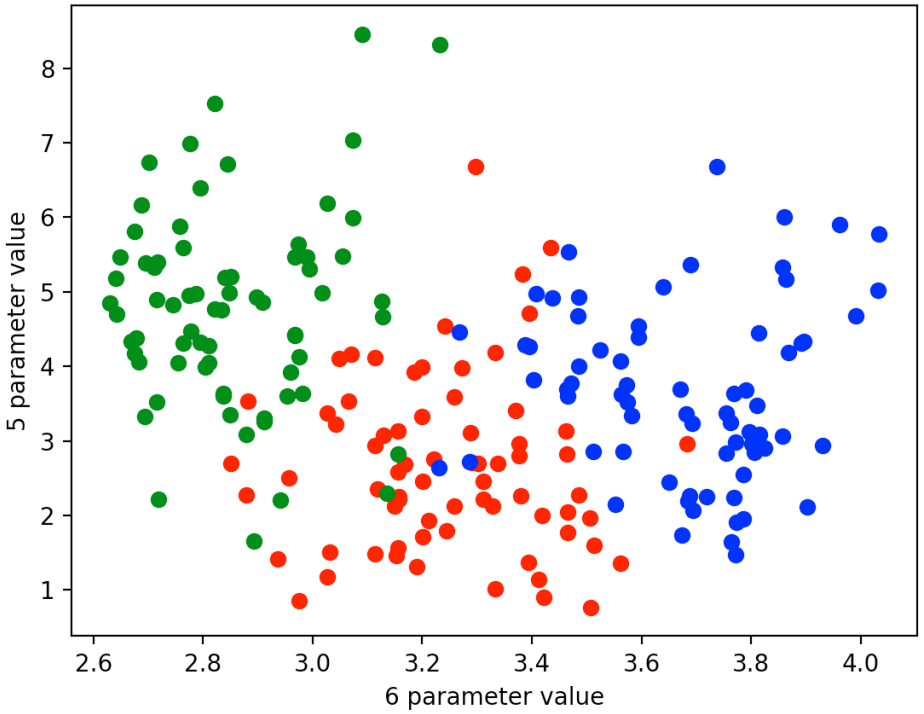
- result_prob - wyznaczenie prawdopodobieństw przynależności danej próbki dla wszystkich klas, z pominięciem dzielenia przez prawdopodobieństwo P(data).
- matrix_correct - obliczanie macierzy błędu oraz liczby poprawnych wyników
- main - znajdują się tam wszystkie testy związane z zadaniem.

Sprawdzając na wykresie rozmieszczenie punktów dla danej pary wybranych parametrów próbek, w wielu przypadkach możemy rozpoznać poszukiwane przez nas grupy nasion, tzn. są one liniowo separowalne, oto przykłady poniżej:

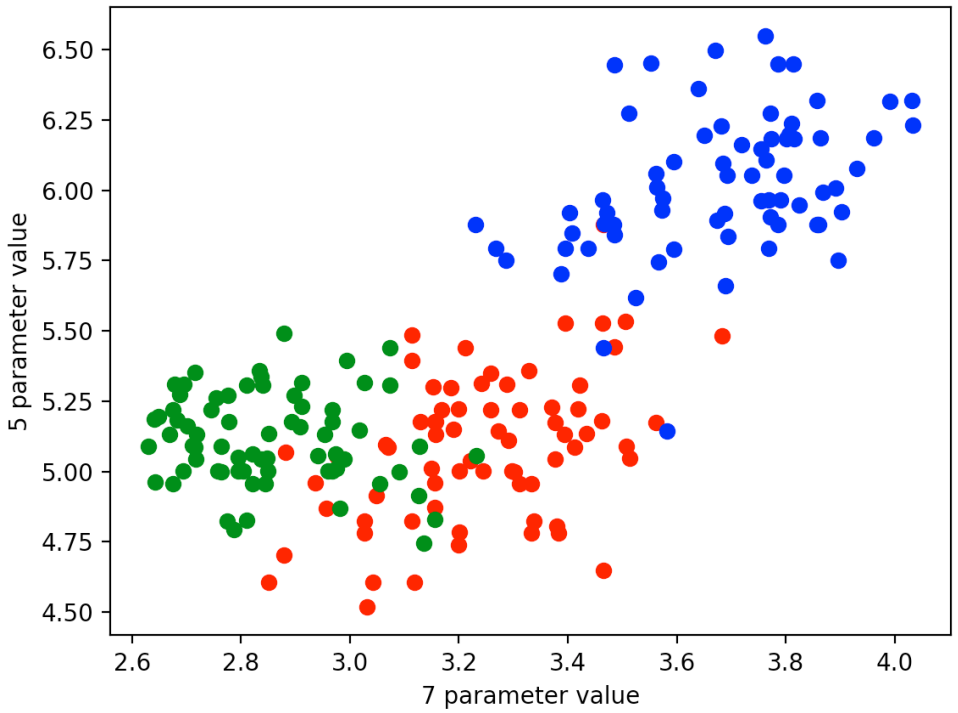
Wykres dla parametru pierwszego i trzeciego:



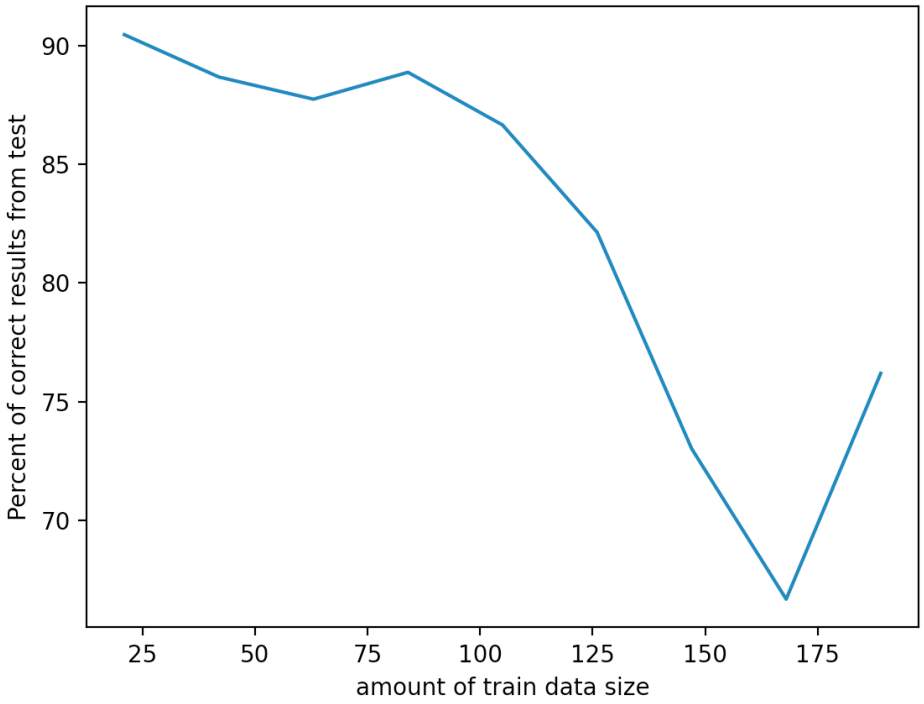
Wykres dla parametru piątego i szóstego:



Wykres dla parametru piątego i siódmego:

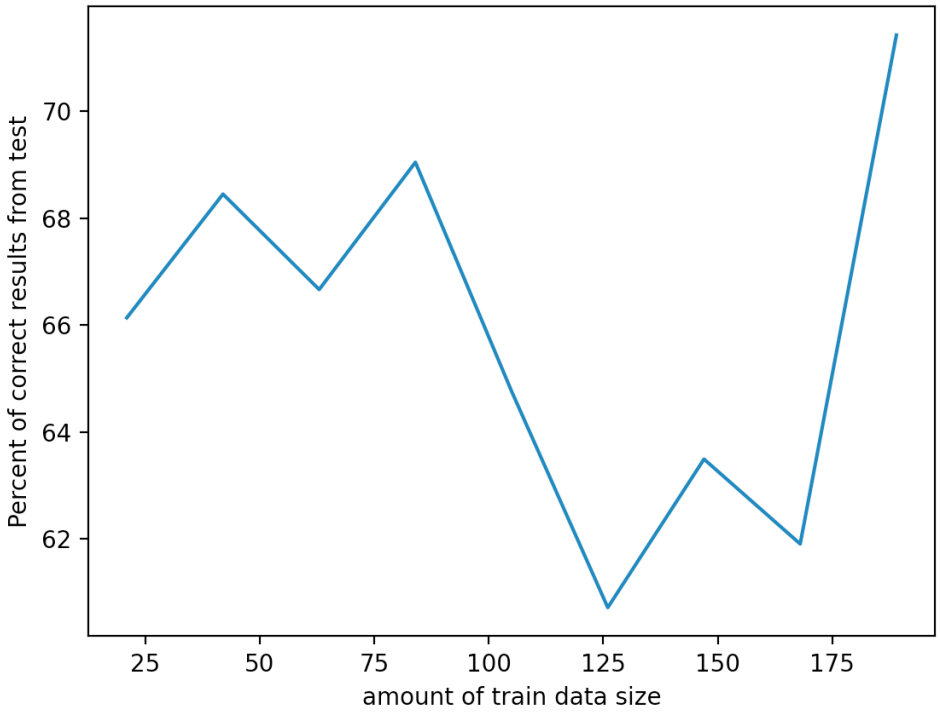


Następnie przetestowałem jak podział na dane trenujące i testujące wpływa na skuteczność modelu.
Można zaobserwować na poniższym wykresie, że im większy zbiór danych trenujących spośród 210 próbek, tym otrzymujemy mniejszą skuteczność modelu.
Zatem widoczne jest wyraźne przeuczenie.
Nie zaobserwowałem zjawiska niedouczenia w tym przypadku.

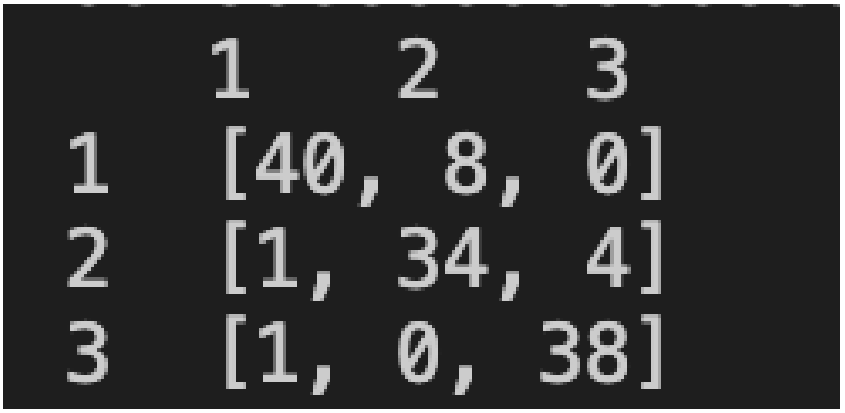


Powyższy przypadek jest rozwiązany gdy grupy są odpowiednio posortowane.
Poniżej znajduje się wykres efektywności algorytmu w zależności wielkości grupy trenującej, czyli tak samo jak powyżej, lecz tym razem dane zostały pomieszane, nie są już posortowane według grup.
Możemy zauważyć, że efektywność zdecydowanie spadła, w najbardziej optymalnej sytuacji efektywność wynosi 71%,

natomiast w poprzednim przypadku maksymalna efektywność wynosiła 90%.
Sortowanie odgrywa zatem bardzo ważną rolę, jest bardzo korzystne.



Następnie obliczam wszystkie wskazane metryki na podstawie wyznaczenia macierzy błędu.



Tablica ma 3 wiersze i 3 kolumny, wiersze przedstawiają klasy predykowane, natomiast kolumny zaś klasy rzeczywiste.

Obliczam wartości TP, FP, FN, TN osobno dla każdej klasy:

dla 1 grupy:
TP = 40
FP = 8 + 0 = 8
FN = 1+1 = 2
TN = 34 + 4 + 0 + 38 = 76

Recall TPR = 40 / (40 + 2) ~ 0.95
Fall-out FPR = 8 / (8 + 76) ~ 0.01
Precision PPV = 40 / (40 + 8) ~ 0.83
Accuracy ACC = (40 + 76) / (40 + 76 + 8 + 2) ~ 0.92
F1-score F1 = (2 * 0.83 * 0.95) / (0.95 + 0.83) ~ 0.89
---
dla 2 grupy:
TP = 34
FP = 1+4= 5
FN = 8 + 0 = 8
TN = 40 + 0 + 1 + 38 = 79
Recall TPR = 34 / (34 + 8) ~ 0.81
Fall-out FPR = 5 / (5 + 79) ~ 0.06
Precision PPV = 34 / (34 + 5) ~ 0.87
Accuracy ACC = (34 + 79) / (34 + 79 + 5 + 8) ~ 0.90
F1-score F1 = (2 * 0.87 * 0.81) / (0.81 + 0.87) ~ 0.84
---
dla 3 grupy:
TP = 38
FP = 1 + 0 = 1
FN = 0 + 4 = 4
TN = 40 + 8 + 1 + 34 = 83

Recall TPR = 38 / (38 + 4) ~ 0.91
Fall-out FPR = 1 / (1 + 83) ~ 0.01
Precision PPV = 38 / (38 + 1) ~ 0.97
Accuracy ACC = (38 + 83) / (38 + 83 + 1 + 4) ~ 0.96
F1-score F1 = (2 * 0.97 * 0.91) / (0.91 + 0.97) ~ 0.94

Na koniec wyznaczam globalną wartość F1:
Makro F1 = (0.94 + 0.84 + 0.89) / 3 = 0.89