

Wydział Elektroniki

Metody techniki systemów w medycynie 2

Komputerowe wspomaganie diagnozowania nowotworów piersi z wykorzystaniem naiwnego algorytmu bayesowskiego.

Autorzy:

Kamil Nowak 234959  
Filip Kądziołka

Prowadzący:

dr inż. Andrzej Żołnierek

Wrocław 2019

# Algorytm Bayesowski

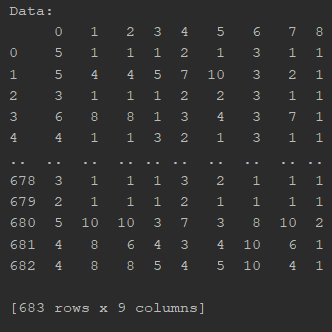
Wrzucić wzory, definicję książkową co to Bayes, to co tłumaczył nam Żołnierek i jak tak naprawdę działa, względem naszego programu xD

# Materiał badań

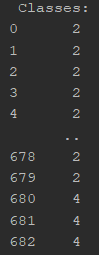
W celach badawczych korzystaliśmy z datasetu o nazwie „Breast Cancer Wisconsin (Original)” dostępnego na stronie archive.ics.uci.edu. Składa się on z 699 instancji przy czym niektóre z nich zawierają niekompletne dane. Brak wartości atrybutu jest oznaczony znakiem „?”. Każda instancja składa się z atrybutu:

* Sample code number: id number
* Clump Thickness (0)
* Uniformity of Cell Size (1)
* Uniformity of Cell Shape (2)
* Marginal Adhesion (3)
* Single Epithelial Cell Size (4)
* Bare Nuclei (5)
* Bland Chromatin (6)
* Normal Nucleol (7)
* Mitoses (8)
* Class: (2 for benign, 4 for malignant)

Wszystkie atrybuty oprócz pierwszego (id) oraz ostatniego są dyskretne oraz są opisywane przez liczbę z zakresu od 1 do 10.



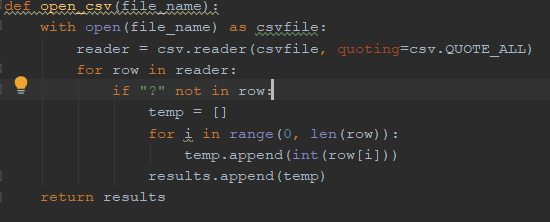
Ostatni atrybut oznacza klasę czyli w naszym przypadku opisuje czy kobieta przy podanym zestawie atrybutów ma łagodnego raka(określony przez liczbę 2) czy też złośliwego(określony przez liczbę 4).



# Zaimplementowany algorytm

* 1. Odczyt danych

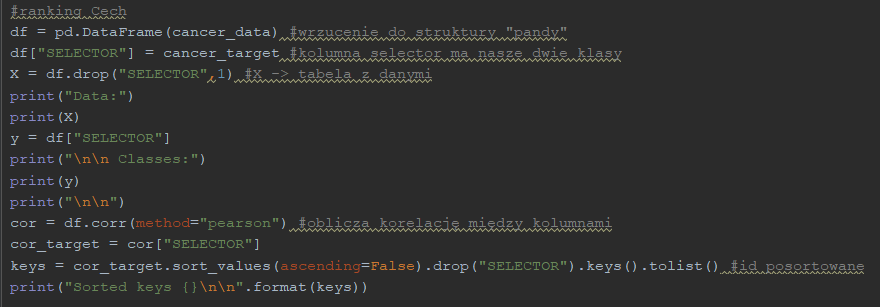
Dane pobrane ze strony były dostępne w formacie „.data”, lecz w celu łatwiejszego dostępu do pliku przenieśliśmy wszystkie dane do pliku CSV. Wybraliśmy taki format pliku z powodu zapisu danych (wszystkie atrybuty były oddzielone znakami „,”. Do odczytu pliku została wykorzystana biblioteka CSV. Przy odczycie danych zostały pominięte linie w których występuje znak „?”.



* 1. Wyznaczanie rankingu cech

Do przeprowadzenia rankingu cech skorzystaliśmy z współczynnika korelacji Pearsona(Pearson Product Moment Correlation – PPMC) Określa on poziom zależności liniowej między zmiennymi losowymi.

Powyższy wzór prezentuje jak wyliczyć współczynnik korelacji Pearsona. W naszym przypadku x jest naszym konkretnym atrybutem, y wartością klasy, a n jest liczbą instancji przez które musi przejść algorytm czyli w naszym przypadku jest to wartość 682.

Do zaimplementowania algorytmu skorzystaliśmy z biblioteki Pandas. Na początku wrzuciliśmy nasze dane do jednej ze struktur dostępnych w bibliotece Pandas – DataFrame. Dodatkowo dodaliśmy kolumnę która posiada nasz selektor, czyli naszą klasę. Metoda corr(method = „perarson”) wylicza korelację metodą pearson’a między parami kolumn. Ostatnim punktem jest posortowanie wyników korelacji. W naszym przypadku kolejność atrybutów od najbardziej istotnych do najmniej wygląda następująco: [5, 2, 1, 6, 7, 0, 3, 4, 8] przy czym każda liczba odpowiada konkretnemu atrybutowi. W punkcie 2 zostały przyporządkowane id do konkretnych atrybutów.

* 1. Badanie

Badanie przebiega dla różnej liczby cech. Dodajemy pojedynczo do badania kolejne atrybuty wedle kolejności wyznaczonej w rankingu cech(od najbardziej przydatnych argumentów do klasyfikacji do tych najmniej istotnych). Za każdym razem testujemy klasyfikator z wykorzystaniem 5 razy powtarzanej metody dwukrotnej walidacji krzyżowej. Owa walidacja polega na tym, że dzielimy nasz zbiór na zbiór uczący oraz testujący, Następnie zamieniamy miejscami zbiory przez co zbiór uczący staje się zbiorem testujący, a testujący zbiorem uczącym.

Ranking cech pomógł nam określić jaka liczba konkretnych cech jest najlepsza dla naszego algorytmu w celu jak największej efektywności algorytmu.

Po skończeniu algorytmu wyświetlamy jakość klasyfikacji podanej w procentach która jest w formie uśrednionej względem 5 razy powtórzonej 2-krotnej walidacji krzyżowej. Najlepszy przypadek, czyli z największą jakością klasyfikacji jest przedstawiony jako macierz konfuzji. Jest to podstawowe narzędzie wspomagające ocenę jakości klasyfikacji. Rozróżniamy w niej dwie klasy:

* Pozytywną
* Negatywną

Możemy przez to dostać wyniki klasyfikacji takie jak:

* + - Prawdziwa pozytywna
    - Prawdziwa negatywna
    - Fałszywa pozytywna
    - Fałszywa negatywna

|  |  |  |
| --- | --- | --- |
|  | Zaklasyfikowany do klasy pozytywnej | Zaklasyfikowany do klasy negatywnej |
| Należy do klasy pozytywnej | Prawdziwa pozytywna | Fałszywa negatywna |
| Należy do klasy negatywnej | Fałszywa pozytywna | Prawdziwa negatywna |

# Wyniki

# Wnioski

# Bibliografia

<https://www.ii.pwr.edu.pl/~zieba/W5_PMWZI.pdf>

<https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn>

<https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstnaiveb.html>

<https://pl.wikipedia.org/wiki/Naiwny_klasyfikator_bayesowski>

<http://zsi.tech.us.edu.pl/~nowak/si/w3a.pdf>

<http://zsi.ii.us.edu.pl/~nowak/smad/SMAD_w2.pdf>

<https://medium.com/@joseph.magiya/pearson-coefficient-of-correlation-using-pandas-ca68ce678c04>

<https://en.wikipedia.org/wiki/Pearson_correlation_coefficient>

<https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-coefficient-formula/>

https://www.python-course.eu/naive\_bayes\_classifier\_introduction.php