Naiwny klasyfikator Bayesa

Maksym Telepchuk

19.03.2020

Naiwne klasyfikatory Bayesa to zbiór algorytmów klasyfikacji opartych na twierdzeniu Bayesa. Jest to rodzina algorytmów, które mają wspólną ideę, tj. każda para klasyfikowanych cech jest od siebie niezależna.

Każdy element ze zbioru danych jest przedstawiany jako wektor cech. Każdy taki wektor jest mapowany na dyskretną wartość (problem klasyfikacji). W danych ze światu rzeczywistego często istnieją zależności pomiędzy tymi cechami. W "Naiwnym Klasyfikatorze Bayesa" zakłada się, że, że każda cecha zapewnia niezależny oraz równy względem innych cech wkład w wynik klasyfikacji. Czyli żadna para cech nie jest zależna od siebie oraz dla każdej cechy przypisuje się taką samą wagę.

1 Twierdzenie Bayesa

Podejście naiwnego klasyfikatora Bayesa jest bazowane na twierdzeniu Bayesa. Twierdzenie Bayesa określa prawdopodobieństwo wystąpienia zdarzenia A pod warunkiem, że zaszło inne zdarzenie B. Twierdzenie Bayesa jest wyrażone matematycznie jako następujące równanie:

$$P(A|B) = \frac{P(B|A)}{P(A)},$$

przy czym $P(B) \neq 0$.

Niech element ze zbioru danych będzie zmienną wektora $X = (x_1, x_2, \ldots, x_n)$, a y będzie zmienną klasy, do której jest ten wektor przypisywany. Wtedy wzór Bayesa można zapisać następująco:

$$P(y|X) = \frac{P(X|y)}{P(X)}$$

Z naiwnego założenia o niezależności cech wzór sprowadza się do następującej postaci:

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y) \ldots P(x_n|y)}{P(x_1) \ldots P(x_n)}$$

Ponieważ wartość dzielnika jest stała dla danego wektora, można go pominąć przy klasyfikacji. Wtedy przewidywana klasa dla danego wektora będzie się równała

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y),$$

czyli jest wybierany taki y, który prowadzi do największej wartości powyższego wyrazu.

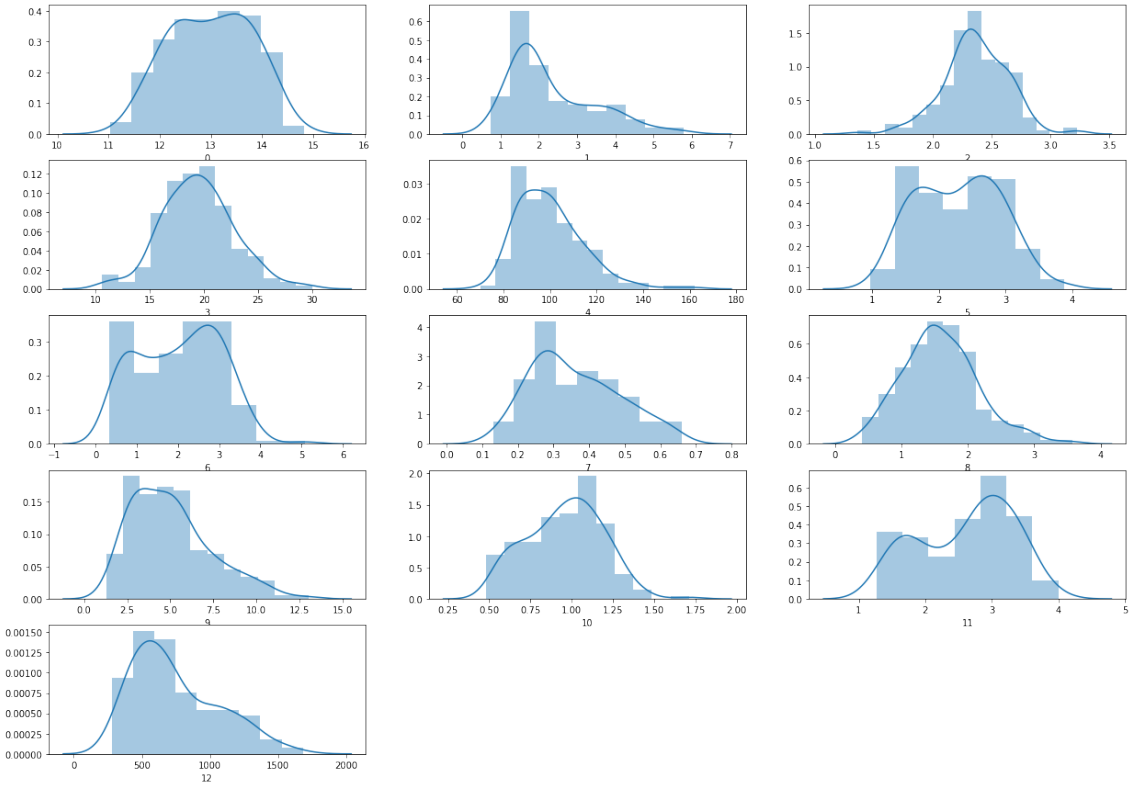
Zetem, przy wyliczonych prawdopodobieństwach wystąpienia wartości dla osobnych cech oraz klas na podstawie danego zbioru danych, klasyfikacja naiwnym klasyfikatorem Bayesa jest trywialna.

2 Zbiory danych

Dla analizy zostały wybrane następujące zbiory danych :

- glass - 6 typów szkła
- wine - skład chemiczny 3 rodzajów wina
- seeds - 3 typy nasienia

Loader dla zbioru wine jest dostępny w module **sklearn.datasets**. Loadery dla pozostałych zbiorów zostały zaimplementowane na podstawie plików **csv**. Poniżej są podane rozkłady cech dla każdego ze zbiorów.



Rysunek 1: Histogram cech wine

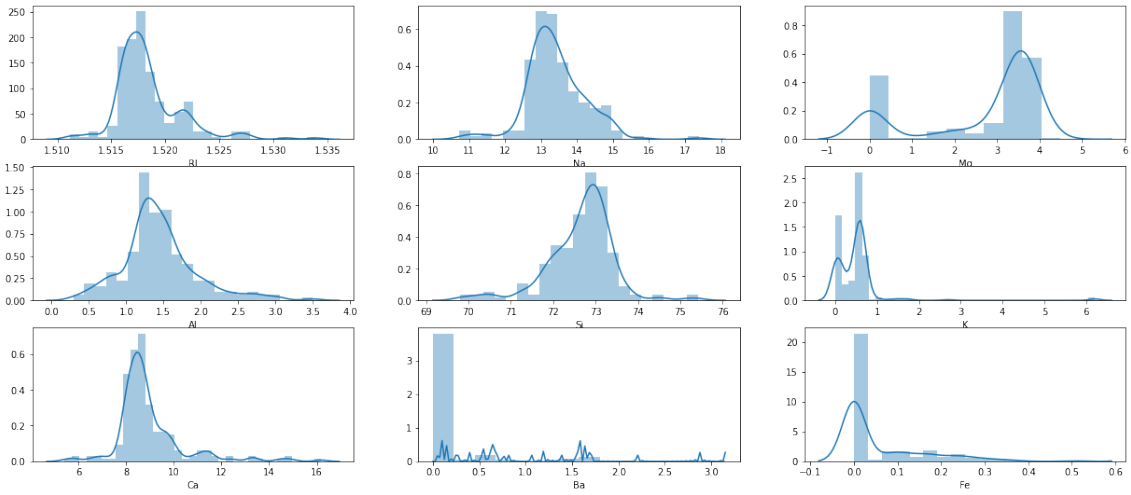
3 Implementacja klasyfikatora

Rozkłady cech z poprzedniego rozdziału wskazują na to, że warto spróbować użyć podejścia Gaussowskiego przy implementacji bayesowskiego klasyfikatora. W danym zadaniu m.in. została wykorzystana model klasyfikatora 'Gaussian Naive Bayes', w którym zakłada się, że ciągłe wartości związane z każdą cechą są rozkładane zgodnie z rozkładem Gaussa. Rozkład Gaussa jest również nazywany Rozkładem normalnym. Wykres tego rozkładu daje krzywą w kształcie dzwonu, która jest symetryczna względem średniej wartości cech.

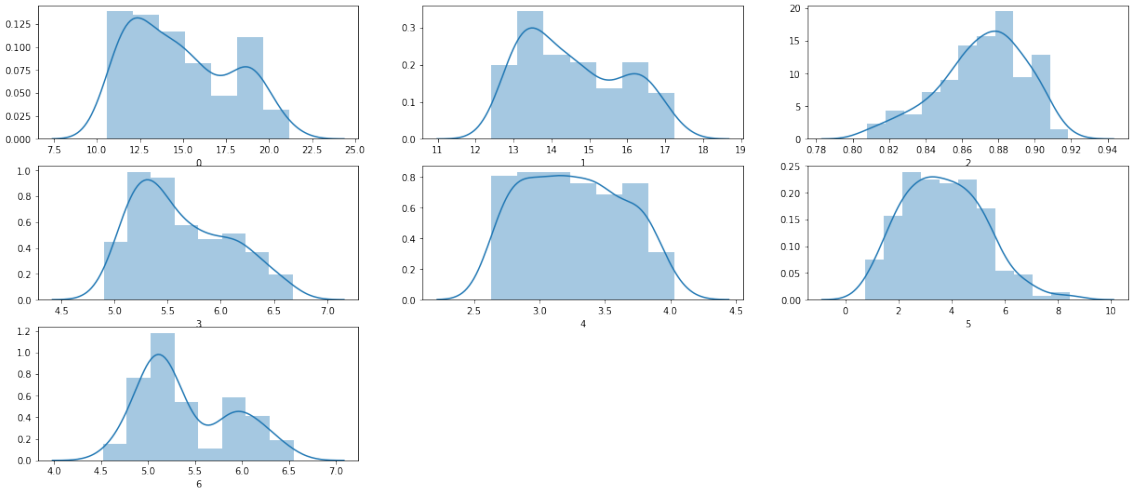
Zakłada się, że prawdopodobieństwo cech jest gaussowskie, stąd prawdopodobieństwo warunkowe podaje:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

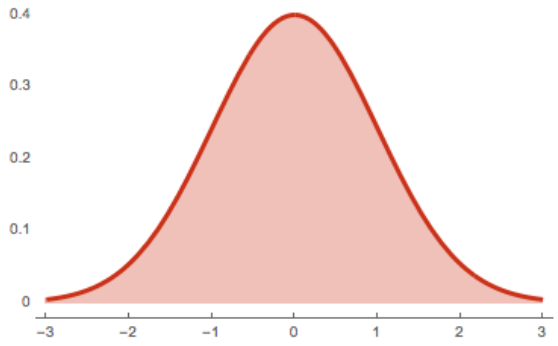
Biblioteka **sklearn** zawiera gotową implemetację, która pozwala na szacowanie σ i μ przy użyciu maksymalnego prawdopodobieństwa.



Rysunek 2: Histogram cech glass



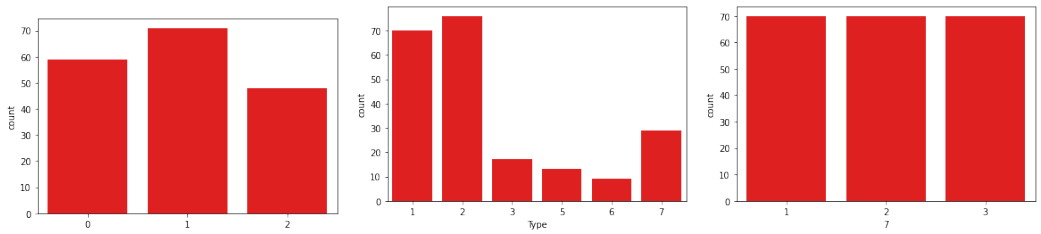
Rysunek 3: Histogram cech seeds



Rysunek 4: Rozkład Gaussa

4 Wyniki klasyfikatora

Na rozkładzie predykowanych klas widać, że w niektórych zbiorach dane są nie zbalansowane. Dla tego oprócz metryki **accuracy** została użyta również metryka **f1_score**.



(a) Wine (b) Glass (c) Seeds

Rysunek 5: Rozkłady predykowanych klas

Na poniższych tabelach zamieszczone średnia z 10 wyników losowego podziału klasyfikatora dla rozmiaru testowego zbioru 20 i 40 procent, oraz wyniki cross walidacji (średnia) dla danych zbiorów. Dla każdej z me

	F1_Score	Accuracy
wine	0.99	0.98
glass	0.37	0.28
seeds	0.94	0.94

Tabela 1: test size = 0.4

	F1_Score	Accuracy
wine	0.99	0.99
glass	0.35	0.25
seeds	0.93	0.92

Tabela 2: test size = 0.2

	F1_Score	Accuracy
wine	0.98	0.98
glass	0.46	0.44
seeds	0.90	0.90

Tabela 3: k fold = 5

	F1_Score	Accuracy
wine	0.97	0.97
glass	0.43	0.44
seeds	0.90	0.90

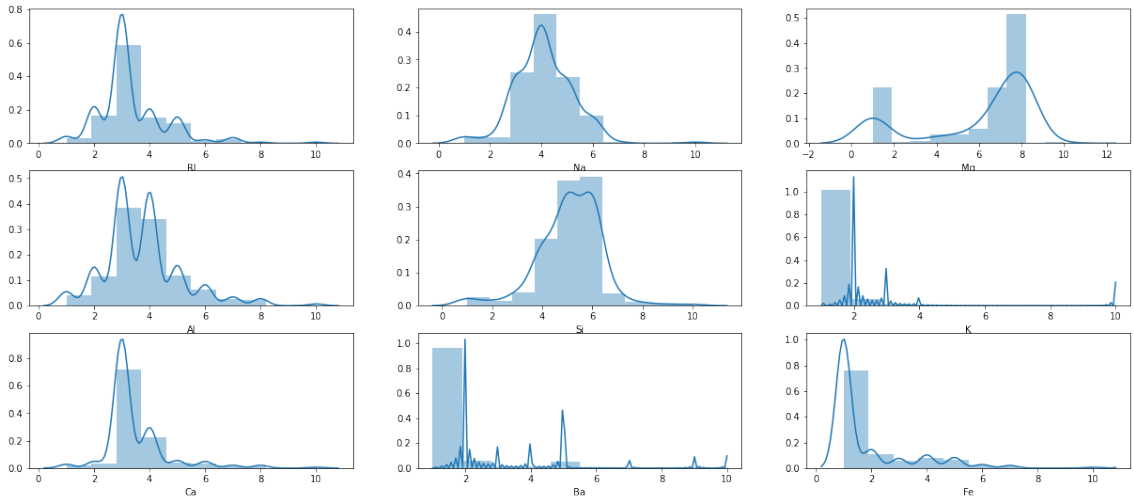
Tabela 4: k fold = 10

	Accuracy	F1_Score
wine	0.97	0.97
glass	0.44	0.47
seeds	0.90	0.90

Tabela 5: Stratyfikowana cross walidacja dla n_splits = 5

	Accuracy	F1_Score
wine	0.98	0.98
glass	0.48	0.44
seeds	0.90	0.90

Tabela 6: Stratyfikowana cross walidacja dla n_splits = 10



Rysunek 6: Rozkład cech po liniowej dyskretyzacji dla 10 binów.

5 Dyskretyzacja

Z wyników klasyfikatora można wywnioskować, że takie podejście nie sprawdza się przy testowaniu na zbiorze danych glass. W tym zbiorze danych spróbowano użyć dyskretyzacji. Wartości metryk podane w tabelach są to średnia wartości z wyniku 10 klasyfikacji przy przemieszanym zbiorze danych.

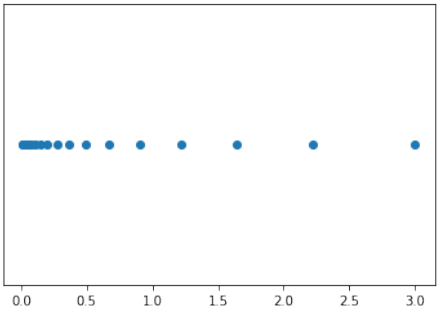
Pierwszym sposobem na radzenie sobie z ciągłym danymi jest równomierne podzielenie wartości (od maksymalnej do minimalnej wartości w zakresie cechy) na n równych części (binów). Każda wartość cechy zatem jest zaokrąglana do najbliższego takiego podziału. Wyniki liniowej dyskretyzacji w przypadku 40% zbioru testowego są podane w tabeli 7:

#bins	Accuracy(%)	F1_Score (%)
10	25.58	28.11
20	40.11	35.10
30	40.58	36.88

Tabela 7: Liniowa dyskretyzacja.

Cechy 'Ba' i 'Fe' są bardzo zdebalansowane. Większość wartość znajduje się w okolicach zera. Kolejnym typem dyskretyzacji, której zostało użyto jest dyskretyzacja geometryczna. Biny w tym przypadku są rozmieszczone równomiernie na skali logarytmicznej (rysunek 7). Ponieważ nie początkiem geometrycznej dyskretyzacji nie może być zero (log(0)), w tym przypadku zero zostało zastąpione wartością 0.001. Wyniki liniowej dyskretyzacji w przypadku 40% zbioru testowego są podane w tabeli 8:

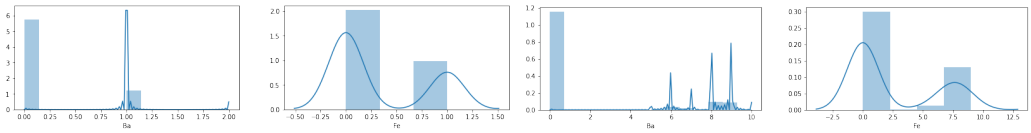
Mieszany sposób dyskretyzacji został wybrany jako trzeci. Na wykresach rozkładów cech dla tego zboiru widać, że założenie o gaussowskim rozkładzie nie działa dla cech Mg, Ba, K. Średnie wartości dla tych cech to 2.68, 0.17 i 0.49 odpowiednio. Warto spróbować użyć dyskretyzacji tych cech. Dla cechy Mg wartości cech zostały zaokrąglone do najbliższych wartości z listy [0,2.68,3.4,3.65,3.8,3.95,4] (0 i geometryczna dyskretyzacja dla 6 binów). Dla cechy Ba wartości



Rysunek 7: Rozkład binów przy geometrycznej dyskretyzacji

#bins	F1_Score (%)	Accuracy(%)
10	36.63	29.06
2	38.91	30.23

Tabela 8: Liniowa dyskretyzacja.



(a) Rozkład dyskretyzowanych cech po geometrycznej dyskretyzacji dla 2 binów.

(b) Rozkład dyskretyzowanych cech po geometrycznej dyskretyzacji dla 10 binów.

zostały zaokrąglone do najbliższych wartości z listy [0, 0.6, 0.73, 0.86, 0.99, 1.12, ... , 3.15] (0 i liniowa dla 20 binów). Dla cechy Mg wartości cech zostały zaokrąglone do najbliższych wartości z listy [0,0.03,0.49,0.6,0.8] (wartości najczęściej są skupione wokół tych wartości oraz zostały wybrane manualnie). W wyniku taka dyskretyzacja dała następujące wyniki

	F1_Score	Accuracy
glass (test size = 0.2)	52.99	46.51
glass (test size = 0.4)	42.5	45.34

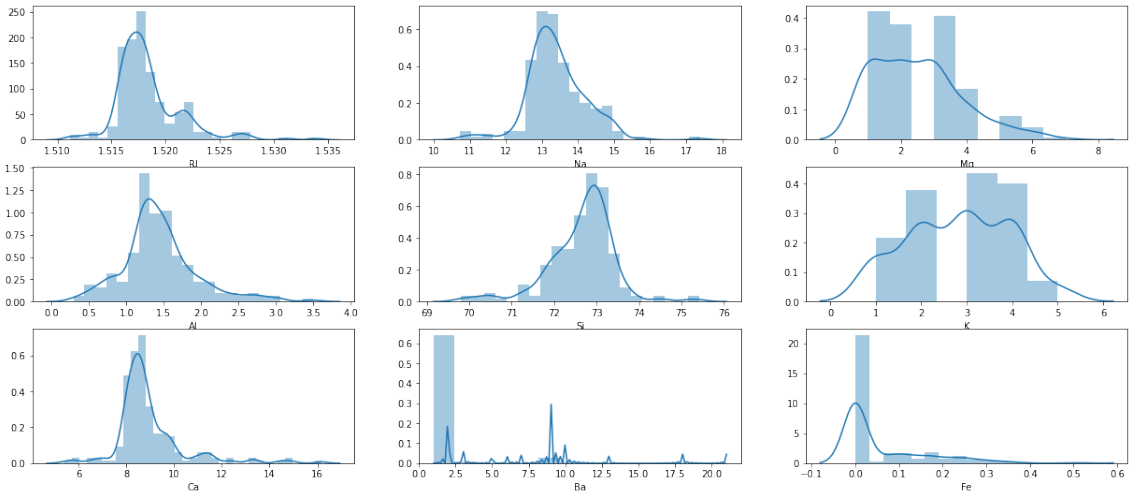
Tabela 9: Glass po dyskretyzacji

Czyli mieszana dyskretyzacja podwyższyła wyniki klasyfikacji prawie w dwa razy.

6 Wnioski

Pomimo zbyt uproszczonych założeń, wbrew pozorom naiwni klasyfikatorzy Bayesa działali całkiem dobrze w wielu rzeczywistych sytuacjach. Wymagają niewielkiej ilości danych treningowych do oszacowania niezbędnych parametrów. Naiwni uczniowie i klasyfikatorzy Bayesa mogą być niezwykle szybcy w porównaniu do bardziej wyrafinowanych metod. Oddzielenie klasowych rozkładów cech warunkowych oznacza, że każdy rozkład można niezależnie oszacować jako rozkład jednowymiarowy. To z kolei pomaga złagodzić problemy wynikające z przekleństwa wymiarowości (curse of dimensionality).

W przypadku danych ciągłych dyskretyzacja cech oraz założenie, że cechy pochodzą z rozkładu normalnego często polepszają wyniki końcowe.



Rysunek 9: Rozkład cech po mieszanej dyskretyzacji.