Rafał Galczak - ROB Lab. 2

Rafał Galczak

Marzec 2020

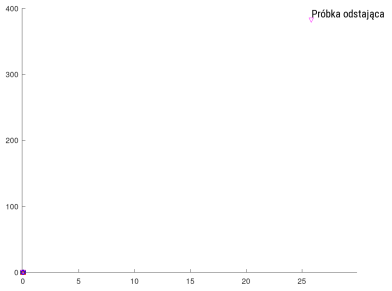
Analiza danych

Dla danych uczących sprawdzono średnią, medianę i odchylenie standardowe. Można zobaczyć, że dla niektórych atrybutów mediana i średnia różnią się o kilka rzędów wielkości:

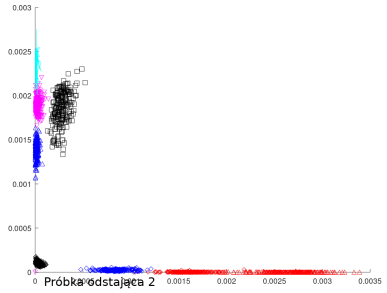
Średnie:	`2.0882e-01 7.9658e+01 1.0604e+00 9.0846e-03`
Mediany:	`1.9996e-06 -8.9358e-11 1.3626e-10 -1.8427e-14`

Usunięcie danych odstających

Na wykresie 1a widać jedną, mocno odstającą próbkę (indeks próbki ze zbioru uczącego: 186) która została usunięta.



(a) Próbka odstająca na wykresie dla atrybutów 3 i 4

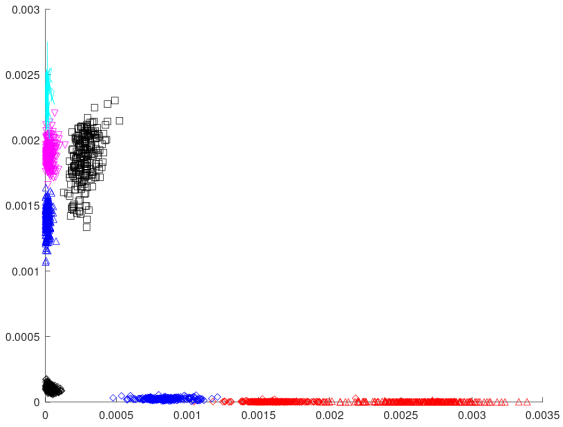


(b) Druga próbka odstająca dla tych samych atrybutów

Po usunięciu próbki widoczna była jeszcze jedna próbka odstająca blisko punku (0, 0), którą widać na wykresie 1b. Ta próbka również została usunięta (indeks próbki w zbiorze uczącym: 641).

Wybór atrybutów klasyfikacji

Do klasyfikacji użyto atrybutów 2 i 3.



Rysunek 2: Atrybuty 2 i 3

Jak widać na wykresie 2 trzeci atrybut (oś Y) nie radzi sobie z segregacją klas 2, 6, 8, klas 1, 3 i klas 1,4. Natomiast drugi atrybut (oś X) słabo dzieli klasy 3, 4 i 5. Wybrane atrybuty nie mają wspólnych grup klas, których nie umieją podzielić.

Wyniki

Wyniki dla klasyfikatora Bayesa

`pdf_indep`	`pdf_multi`	`pdf_parzen` (h=0.001)
`0.021382`	`0.020833`	`0.016996`

Tablica 1: Odsetek błędnie sklasyfikowanych przypadków

Wyniki dla różnych wielkości zbioru uczącego

	średnia	odchylenie standardowe
10% zbioru uczącego		
pdf_idep	0.0255482	0.0020732
pdf_multi	0.0246711	0.0021584
pdf_parzen	0.0408991	0.0069629
25% zbioru uczącego		
pdf_idep	0.0208333	0.0019383
pdf_multi	0.0209430	0.0015213
pdf_parzen	0.0281798	0.0021091
50% zbioru uczącego		
pdf_idep	0.0213816	0.0010257
pdf_multi	0.0205044	0.0010687
pdf_parzen	0.0200658	0.0018833

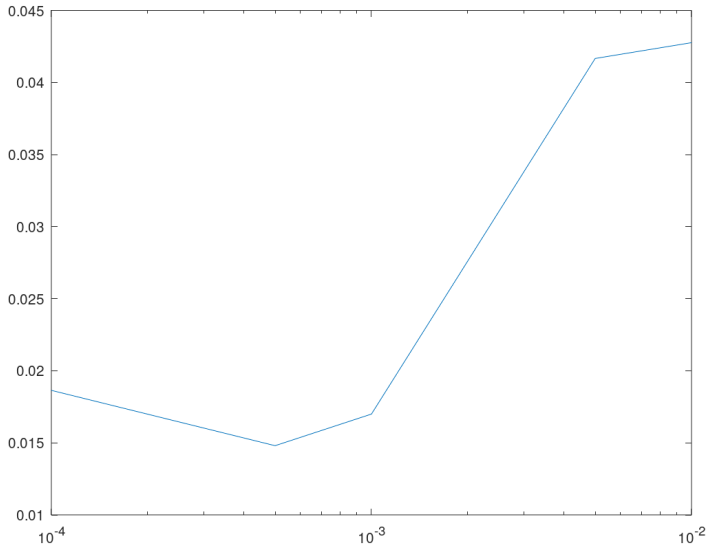
Tablica 2: Odsetek błędnie sklasyfikowanych przypadków dla różnych części zbioru uczącego

Na podstawie tabeli 2 widać, że im więcej próbek w zbiorze uczącym tym mniejszy odsetek błędnych klasyfikacji (wyjątek stanowi funkcja dla prawdopodobieństw niezależnych, tutaj wyniki dla 25% zbioru uczącego są lepsze niż dla 50% zbioru uczącego, ale wyniki dla 50% zbioru uczącego ciągle pozostają lepsze niż dla 10% zbioru uczącego). Odchylenie standardowe bez wyjątku maleje wraz ze wzrostem liczby próbek w zbiorze uczącym.

Wyniki dla różnych szerokości okna parzena

h_1	średni odsetek błędów
0.0001	0.01864035
0.0005	0.01480263
0.001	0.01699561
0.005	0.04166667
0.01	0.04276316

Tablica 3: Odsetek błędów dla różnych szerokości okna parzena



Rysunek 3: To samo na skali logarytmicznej

Z danych z tabeli 3 jak i wykresu 3 widać, że im większa wartość okna parzena tym większy błąd klasyfikatora.

Wyniki dla zmodyfikowanych rozkładów prawdopodobieństwa w zbiorze testowym

Prawdopodobieństwa występowania przykładów w zbiorze testowym zostały zmodyfikowane zgodnie z treścią zadania (`[0.165 0.085 0.085 0.165 0.165 0.085 0.085 0.165]`).

pdf_indep	pdf_multi	pdf_parzen
0.017398	0.016520	0.014035

Tablica 4: Średni odsetek błędów dla zmodyfikowanych prawdopodobieństw

W porównaniu do wyników z klasyfikacji o niezmodyfikowanych prawdopodobieństwach odsetek błędnych klasyfikacji zmniejszył się dla wszystkich klasyfikatorów.

Analiza macierzy pomyłek klasyfikatora pokazuje, że klasy 4, 5 i 8 mają bardzo mało pomyłek (od 0 do 2). Są to klas, których prawdopodobieństwo wystąpienia zwiększyliśmy (z pominięciem klasy 1). Reszta klas tymczasem ma tych pomyłek więcej (3-10). Ponieważ zwiększyliśmy prawdopodobieństwo wystąpienia próbek klas, które są klasyfikowane z bardzo małym błędem, a zmniejszyliśmy prawdopodobieństwo wystąpienia próbek klas, których klasyfikacja ma wyższy współczynnik błędu, to naturalnym jest, że ocena klasyfikatora wzrośnie (a tym samym spadnie odsetek błędnie klasyfikowanych próbek).

Porównanie z klasyfikatorem najbliższego sąsiedztwa

Na podstawie odchylenia standardowego cech 2 i 3 (tabela 5) stwierdzono, że normalizacja cech nie jest konieczna (odchylenia standardowe obu cech różnią się nieznacznie ok. 3.1%).

Cecha	Odchylenie standardowe
2	0.00092062
3	0.00095129

Tablica 5: Odchylenia standardowe cech 2 i 3

Odsetek błędnie sklasyfikowanych próbek dla klasyfikatora 1-NN wyniósł 0.018092. Oznacza to, że poradził sobie lepiej niż klasyfikator Bayesa dla funkcji rozkładu `pdf_indep` i `pdf_multi`, nie jest jednak lepszy niż klasyfikator Bayesa przybliżający funkcję rozkładu prawdopodobieństwa za pomocą metody Parzena.