ROB laboratorium 2.

Marcin Hanas

Sprawdzenie danych zbioru uczącego i zbioru testowego

Sprawdzono dane przez wybieranie pary cech (z indeksem 2. oraz od 3. do 6.) i wyświetlenie dla nich wykresów, oraz znalezienie minimalnych oraz maksymalnych wartości dla każdej cechy. Na podstawie uzyskanych danych ustalono, że ze zbioru uczącego należy usunąć próbki:

indeks	klasa	cecha 2.	cecha 3.	cecha 4.	cecha 5.	cecha 6.	cecha 7.	cecha 8.
186	3	5.12127	25.82285	382.13906	380.85837	145296.77100	1934.12276	16.57025
642	3	642	3	0.125	0	0	0	0

Natomiast ze zbioru testującego usunięto próbki:

indeks	klasa	cecha 2.	cecha 3.	cecha 4.	cecha 5.	cecha 6.	cecha 7.	cecha 8.
186	3	0.14815	0.005487	0.0025403	0.00010161	-5.1623 e-08	-7.5267 e-06	-1.7798 e-16
642	3	0.14815	0.005487	0.0025403	0.00010161	-5.1623 e-08	-7.5267 e-06	-4.9372 e-16

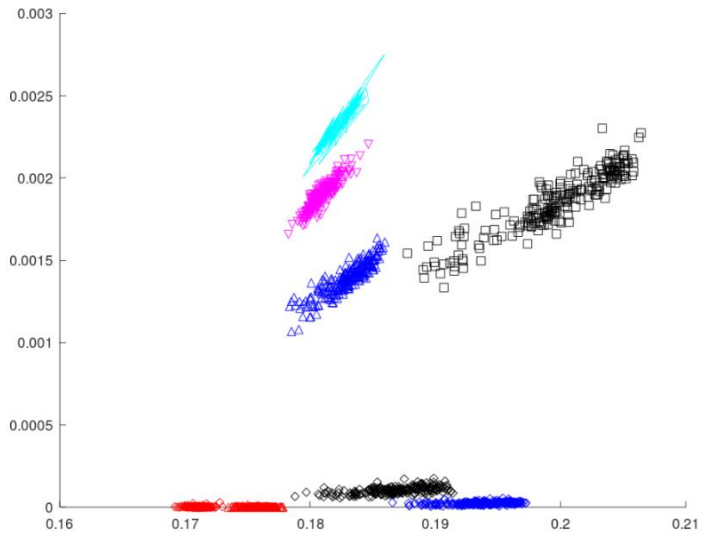
Porównano mediany oraz wartości odchylenia standardowego przed i po usunięciu dwóch próbek ze zbioru uczącego:

	cecha 2.	cecha 3.	cecha 4.	cecha 5.	cecha 6.	cecha 7.	cecha 8.
mediana	0.18679	0.014839	0.21045	0.20882	79.658	1.0604	0.0090846
odch. std.	0.18259	0.00014785	0.00017434	1.9996e-06	-8.9358e-11	1.3626e-10	-1.8427e-14

	cecha 2.	cecha 3.	cecha 4.	cecha 5.	cecha 6.	cecha 7.	cecha 8.
mediana	0.18412	0.00068258	0.00094867	1.4729e-05	-6.9796e-10	2.9713e-07	-9.4668e-11
odch. std.	0.18259	0.00014785	0.00017434	1.9996e-06	-9.0827e-11	1.3626e-10	-1.8742e-14

Wybór cech do klasyfikacji

Obserwując wykresy, do klasyfikacji wybrano zestaw cech o indeksach 2. I 4. – dla pary tych cech, klastery próbek danych klas były najbardziej odseparowane od siebie, co powinno doprowadzić do najmniejszego błędu klasyfikatora.



Niezależne cechy	Rozkład normalny wielowymiarowy	Okno parzena
0.02525	0.003842	0.02305

Jak widać na powyższych wynikach, najmniejszy błąd uzyskano dla metody wykorzystującej wielowymiarowy rozkład normalny – wymagającej znajomości parametrów rozkładu normalnego oraz niezakładającej niezależności cech. Drugą metodą okazało się okno parzena, które nie wymaga znajomości parametrów rozkładu i nie zakłada niezależności cech, a najgorszy wynik uzyskano w przypadku założenia niezależności cech.

Redukcja zbioru uczącego

Dokonano redukcji zbioru uczącego w sposób losowy (po 6 razy dla każdej wartości współczynnika redukcji), a następnie przeprowadzono klasyfikacje zbioru testowego. Poniżej przedstawiono uśrednione wyniki:

Współczynnik redukcji 0.1:

	Niezależne cechy	Rozkład normalny wielowymiarowy	Okno parzena
mediana	0.02808	0.006129	0.09431
odch. std.	0.003478	0.002653	0.01504

Współczynnik redukcji 0.25:

	Niezależne cechy	Rozkład normalny wielowymiarowy	Okno parzena
mediana	0.02781	0.006129	0.05653
odch. std.	0.00305913720689037	0.001530	0.008594

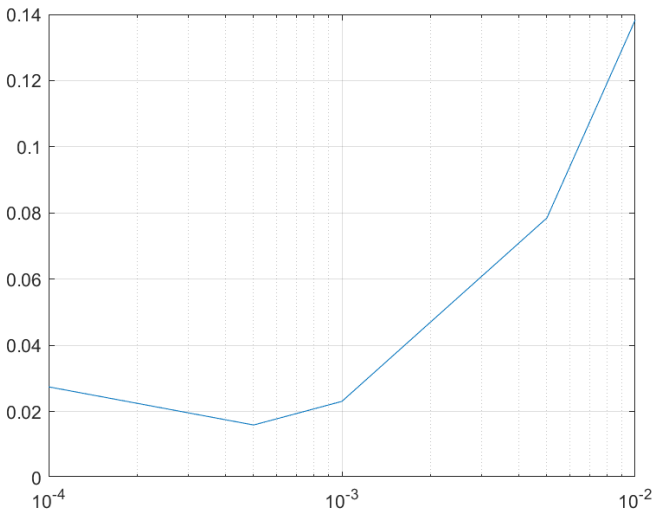
Współczynnik redukcji 0.5:

	Niezależne cechy	Rozkład normalny wielowymiarowy	Okno parzena
mediana	0.02616	0.004940	0.03531
odch. std.	0.001374	0.00069424	0.0042466

Jak widać na powyższych danych, losowe zmniejszanie zbioru uczącego powoduje wzrost liczby błędnie sklasyfikowanych próbek.

Zmiana szerokości okna parzena

Szerokość okna parzena	0.0001	0.0005	0.001	0.005	0.01
Mediana błędu	0.02744	0.01592	0.02305	0.07849	0.1383



Podczas badania działania klasyfikatora z różnymi wartościami szerokości okna parzena okazało się, że najmniejszy błąd uzyskano w przypadku okna o szerokości 0.0005. Niewiele większy błąd pojawił się w przypadku szerokości okna 0.0001 oraz 0.001 – te wartości również są właściwe do zastosowania w algorytmie.

Zmiana prawdopodobieństwa a priori

Dokonano zmiany prawdopodobieństwa a priori tak, aby dla maści czarnych było ono dwukrotnie większe, poprzez zmniejszenie o połowę liczby próbek w zbiorze testującym dla pozostałych klas. Wyniki symulacji dla 6 powtórzeń przedstawiono poniżej:

	Niezależne cechy	Rozkład normalny wielowymiarowy	Okno parzena
mediana	0.02060	0.003414	0.02804
odch. std.	0.001757	0.001438	0.0009995

Sprawdzono, jak wygląda macierz pomyłek klasyfikatora dla powyższych metod, gdy nie są wprowadzone zmiany prawdopodobieństwa:

Niezależne cechy	**220**	**0**	**0**	**8**	**0**	**0**	**0**	**0**	**0.03509**
	0	227	0	0	1	0	0	0	0.004386
	1	0	202	1	0	0	22	0	0.1140
	0	**0**	**0**	**228**	**0**	**0**	**0**	**0**	**0**
	0	**0**	**0**	**0**	**228**	**0**	**0**	**0**	**0**
	0	1	0	0	1	226	0	0	0.008772
	0	0	11	0	0	0	217	0	0.04825
	0	**0**	**0**	**0**	**0**	**0**	**0**	**228**	**0**
Rozkład normalny wielowymiarowy	**227**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0.004386**
	0	227	0	0	1	0	0	0	0.004386
	0	0	226	0	0	0	0	0	0.0087719
	0	**0**	**0**	**228**	**0**	**0**	**0**	**0**	**0**
	0	**0**	**0**	**0**	**228**	**0**	**0**	**0**	**0**
	0	1	0	0	2	225	0	0	0.01316
	0	0	2	0	0	0	226	0	0.008772
	0	**0**	**0**	**0**	**0**	**0**	**0**	**228**	**0**

Okno parzena	**228**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	0	228	0	0	0	0	0	0	0
	1	0	223	1	0	0	1	0	0.02193
	2	**0**	**0**	**225**	**0**	**1**	**0**	**0**	**0.01316**
	0	**0**	**0**	**0**	**218**	**0**	**0**	**10**	**0.04386**
	0	2	0	0	1	225	0	0	0.01316
	0	0	2	0	0	0	226	0	0.008772
	0	**0**	**0**	**0**	**21**	**0**	**0**	**207**	**0.09211**

Pogrubiono klasy maści czarnej, których prawdopodobieństwo zwiększono dwukrotnie. Jak widać, zmiana prawdopodobieństwa spowodowała polepszenie się jakości klasyfikacji dla dwóch pierwszych metod i pogorszenie dla trzeciej. Obserwując macierz pomyłek można zauważyć, że dla dwóch pierwszych metod zwiększono prawdopodobieństwa dla klas o małym wsp. błędu pomyłek. Metoda trzecia dla tych samych klas miała duże współczynniki błędu.

Klasyfikator 1-nn

Przetestowano działanie klasyfikatora 1-nn dla zbioru danych dla wybranych cech (2. i 4.). Mediany oraz odchylenia standardowe danych zbioru uczącego są następujące:

	Cecha 2.	Cecha 4.
mediana	0.1841	0.0009487
odch. std.	0.008845	0.0009513

Stosując na takich nieznormalizowanych danych klasyfikator 1-nn, uzyskano dla danych testujących wskaźnik jakości **0.0165**.

Jako, że obie cechy mają znacznie różniące się od siebie średnie, ustalono, że należy dokonać normalizacji danych – od danych z obu zbiorów uczącego i testującego odjęto medianę i podzielono je przez odchylenie standardowe z powyższej tabeli. W ten sposób dla wszystkich cech zbioru uczącego uzyskano taką samą medianę (równą 0) i odchylenie standardowe.

Po przeprowadzeniu normalizacji, wskaźnik jakości zmniejszył się i wyniósł **0.0049**.