Laboratorium Rozpoznawania Obrazów – Ćwiczenie #2
Klasyfikacja optymalna Bayesa

Piotr Mikołajczyk

2. Proszę sprawdzić dane, a szczególnie zbiór uczący. Wartości odstające w tym zbiorze mogą mieć opłakane skutki dla jakości klasyfikacji (ja znalazłem dwie wartości odstające). Informacja o tym, które próbki zostały usunięte, musi znaleźć się w sprawozdaniu.

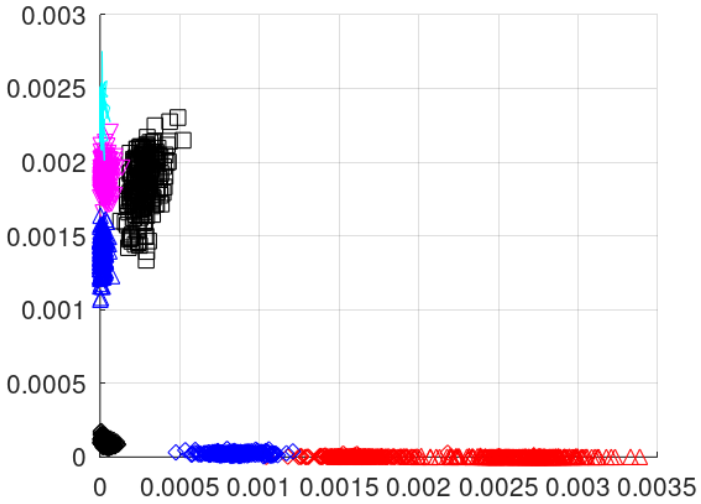
Wartości odstające -> 642 186 – wykryte za pomocą funkcji min oraz max.

3. Proszę wybrać dwie cechy i zbudować dla nich klasyfikator optymalny Bayesa, wyznaczając gęstość prawdopodobieństwa. zgodnie z punktami 1-3. Prawdopodobieństwa a priori przyjąć równe 0.125 (jeśli są równe, nie trzeba ich podawać jawnie). To czy traficie Państwo w cechy najlepsze, nie jest szczególnie istotne, ale warto przy wstępnej analizie danych zwrócić uwagę na „potencjał" klasyfikacyjny poszczególnych cech i wybrać dwie najbardziej obiecujące.

Szerokość okna parzena:0.001

Tabela 1 – Błąd klasyfikacji w procentach dla poszczególnych funkcji gęstości prawdopodobieństwa dla wybranej przestrzeni cech (2,4) oraz (3,4).

Cechy / Rozkład PDF	Indep	Multi	Parzen
(2,4)	2.6316e-02	4.9342e-03	2.4123e-02
(3,4)	2.1382e-02	2.0833e-02	1.6996e-02



Rys. 1 – Wybrane Pary cech (3,4)

Ze względu na mniejszy błąd klasyfikatora dla zestawu 3,4 – wybrano ten zestaw cech. Inne zestawy dawały znacznie większy błąd.

4. Proszę sprawdzić, jaki wpływ na klasyfikację zbioru testowego, ma dobór próbek w zbiorze uczącym (np. wzięcie 1/10, ¼, ½ i całego zbioru uczącego). ***Uwaga: stosowną część próbek ze zbioru uczącego należy wylosować niezależnie dla poszczególnych klas; ponieważ wprowadzamy element losowy, to eksperyment trzeba powtórzyć (minimum 5 razy) i podać uśrednione wyniki (prócz tego warto obejrzeć wartości minimalne, maksymalne i odchylenie standardowe)***
Tu powinniście Państwo zaimplementować funkcję `reduce`*, która zostawia stosowną część poszczególnych klas. W tym punkcie redukcja dotyczy jedynie zbioru uczącego.*

Szerokość okna parzena:0.001

Tabela 2 – Błąd klasyfikacji w procentach dla poszczególnych funkcji gęstości prawdopodobieństwa dla różnej części zbioru uczącego.

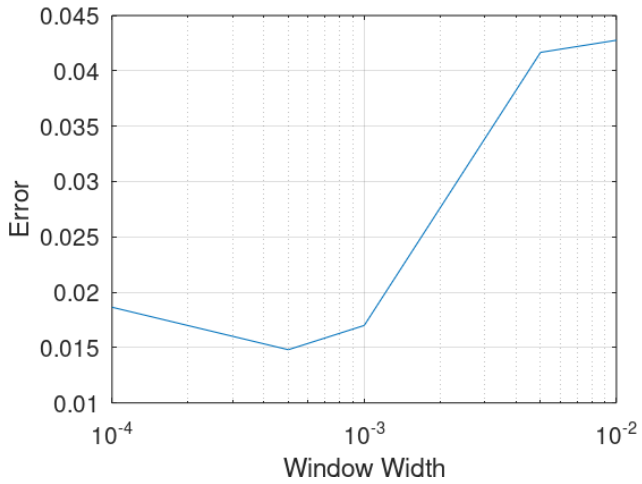
Procent zbioru uczącego / Rozkład PDF	Indep	Multi	Parzen
10	2.6864e-02	2.7047e-02	3.8816e-02
25	2.3099e-02	2.3355e-02	2.7887e-02
50	2.1601e-02	2.0468e-02	2.0066e-02
100	2.1382e-02	2.0504e-02	1.7032e-02

Zgodnie z poleceniem przeprowadzono eksperyment 15 razy dla różnej wielkości zbioru uczącego po czym uśredniono wyniki. Wraz z większą częścią procentową próbek losowanych dla poszczególnych klas ze zbioru, błędy klasyfikacji zmniejszają się.

5. Proszę sprawdzić, jaki wpływ na klasyfikację zbioru testowego, ma dobór parametru (to oczywiście tylko w przypadku klasyfikatora z oknem Parzena).

Tabela 3 – Błąd klasyfikatora w zależności od szerokości okna Parzena.

Szerokość okna	Błąd [%]
1.0000e-04	1.8640e-02
5.0000e-04	1.4803e-02
1.0000e-03	1.6996e-02
5.0000e-03	4.1667e-02
1.0000e-02	4.2763e-02



Rys.2 – Zależność błędu klasyfikatora od szerokości okna Parzena

Wraz ze wzrostem szerokości okna parzena na początku spada nam błąd klasyfikacji a następnie coraz bardziej rośnie – dla konkretnego zestawu danych można znaleźć w ten sposób optymalną szerokość okna dla uzyskania najmniejszego błędu.

6. Jak zmienią się wyniki klasyfikacji jeśli prawdopodobieństwo a priori będzie dwukrotnie większe dla maści czarnych (0.165, 0.085, 0.085, 0.165, 0.165, 0.085, 0.085, 0.165)? Stosowną redukcję wykonujecie tutaj ***tylko*** *na zbiorze testowym. Uwzględnijcie też uwagę z punktu czwartego.*

Próbę wyjaśnienia wyników może wspomóc obejrzenie macierzy pomyłek klasyfikatora – macie Państwo gotową funkcję confMx, która tworzy taką macierz.

Szerokość okna parzena dla wszystkich przypadków:0.001

Na początku stworzono macierze pomyłek dla prawdopodobieństw apriori oraz adekwatnych części zbioru testowego.

a)

apriori = [0.165 0.085 0.085 0.165 0.165 0.085 0.085 0.165];
parts = [1.0 0.5 0.5 1.0 1.0 0.5 0.5 1.0];

Tabela 4 – Macierz pomyłek dla przewagi kart maści czarnej.

[1,1] Indep	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	228	0	0	0	0	0	0	0
	2	0	226	0	0	0	1	0	1
	3	16	0	201	0	0	0	11	0
	4	2	0	3	223	0	0	0	0
	5	0	0	0	0	228	0	0	0
	6	0	5	0	0	0	223	0	0
	7	1	0	5	0	0	0	222	0
	8	0	3	0	0	0	0	0	225

Tabela 5 – Macierz pomyłek dla przewagi kart maści czarnej.

[2,1] Multi	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	228	0	0	0	0	0	0	0
	2	0	226	0	0	0	1	0	1
	3	15	0	202	0	0	0	11	0
	4	4	0	1	223	0	0	0	0
	5	0	0	0	0	228	0	0	0
	6	0	4	0	0	0	224	0	0
	7	0	0	5	0	0	0	223	0
	8	0	3	0	0	0	0	0	225

Tabela 6– Macierz pomyłek dla przewagi kart maści czarnej.

[3,1] Parzen	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	226	0	0	2	0	0	0	0
	2	0	224	0	0	0	1	0	3
	3	8	0	206	3	0	0	11	0
	4	0	0	0	228	0	0	0	0
	5	0	0	0	0	228	0	0	0
	6	0	2	0	0	0	226	0	0
	7	0	0	5	0	0	0	223	0
	8	0	2	0	0	0	0	0	226

b)
apriori = [0.225 0.025 0.025 0.225 0.225 0.025 0.025 0.225];
parts = [1.0 0.1 0.1 1.0 1.0 0.1 0.1 1.0];

Tabela 7 – Macierz pomyłek dla znacznej przewagi kart maści czarnej.

[1,1] Indep	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	228	0	0	0	0	0	0	0
	2	0	224	0	0	0	2	0	2
	3	19	0	198	0	0	0	11	0
	4	2	0	0	226	0	0	0	0
	5	0	0	0	0	227	0	0	1
	6	0	3	0	0	0	225	0	0
	7	1	0	5	0	0	0	222	0
	8	0	0	0	0	0	0	0	228

Tabela 8 – Macierz pomyłek dla znacznej przewagi kart maści czarnej.

[2,1] Multi	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	228	0	0	0	0	0	0	0
	2	0	224	0	0	0	2	0	2
	3	17	0	200	0	0	0	11	0
	4	2	0	0	226	0	0	0	0
	5	0	0	0	0	227	0	0	1
	6	0	3	0	0	0	225	0	0
	7	0	0	7	0	0	0	221	0
	8	0	0	0	0	0	0	0	228

Tabela 9 – Macierz pomyłek dla znacznej przewagi kart maści czarnej.

[3,1] Parzen	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	226	0	0	2	0	0	0	0
	2	0	224	0	0	0	2	0	2

	3	16	0	199	1	0	0	12	0
	4	0	0	0	228	0	0	0	0
	5	0	0	0	0	228	0	0	0
	6	0	2	0	0	0	226	0	0
	7	0	0	7	0	0	0	221	0
	8	0	1	0	0	0	0	0	227

c)
apriori = [0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125];
parts = [1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0];

Tabela 10 – Macierz pomyłek dla równej ilości kart maści czarnej i czerwonej.

[1,1] Indep	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	228	0	0	0	0	0	0	0
	2	0	227	0	0	0	0	0	1
	3	14	0	201	0	0	0	13	0
	4	1	0	1	226	0	0	0	0
	5	0	0	0	0	228	0	0	0
	6	0	7	0	0	0	221	0	0
	7	0	0	5	0	0	0	223	0
	8	0	0	0	0	0	0	0	228

Tabela 11 – Macierz pomyłek dla równej ilości kart maści czarnej i czerwonej.

[2,1] Multi	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	228	0	0	0	0	0	0	0
	2	0	227	0	0	0	0	0	1
	3	13	0	202	0	0	0	13	0
	4	1	0	1	226	0	0	0	0
	5	0	0	0	0	228	0	0	0
	6	0	6	0	0	0	222	0	0
	7	0	0	10	0	0	0	218	0
	8	0	0	0	0	0	0	0	228

Tabela 12 – Macierz pomyłek dla równej ilości kart maści czarnej i czerwonej.

[3,1] Parzen	*Etykiety wybrane przez klasyfikator*								
Prawdziwe Etykiety		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
	1	225	0	0	3	0	0	0	0
	2	0	228	0	0	0	0	0	0
	3	0	0	212	1	0	0	15	0
	4	0	0	0	228	0	0	0	0
	5	0	0	0	0	228	0	0	0
	6	0	5	0	0	0	223	0	0
	7	0	0	9	0	0	0	219	0

	8	0	1	0	0	0	0	0	227

Tabela x – Błąd klasyfikacji w procentach dla poszczególnych funkcji gęstości prawdopodobieństwa dla różnych wariantów kombinacji prawdopodobieństwa apriori oraz części zbioru

Przypadek/ PDF	Indep	Multi	Parzen
a)	1.9408e-02	1.8202e-02	1.6228e-02
b)	2.0395e-02	1.9846e-02	2.1820e-02
c)	1.8640e-02	1.9737e-02	1.5570e-02

Zgodnie z poleceniem przeprowadzono eksperyment minimum 5 razy dla różnej wielkości zbioru uczącego. Z przeprowadzonego eksperymentu wynika że najmniejszy błąd klasyfikacji dla okna Parzena jest dla zbiorów o pełnych. Najgorszy przypadek dla wszystkich rozkładów jest dla przypadku b) dla znacznej przewagi maści czarnych kart. Okazuje się że przypadek jednowymiarowy jest gorszy dla zestawu a) od c) natomiast wielowymiarowy c) od a).

7. Jak mają się wyniki klasyfikatorów Bayesa, do klasyfikatora 1-NN z pierwszego ćwiczenia? (Chodzi oczywiście o to, żeby uruchomić klasyfikator 1-NN na danych kart. Przy okazji musicie Państwo rozstrzygnąć, czy dane kart należy dla tego klasyfikatora normalizować, czy nie.) Ponieważ mamy tu przyzwoitej wielkości zbiór uczący i zbiór testowy, należy sklasyfikować zbiór testowy funkcją `cls1nn` *i obliczyć współczynnik błędu.*

Normalizacja nie jest potrzebna. Rząd wielkości wartości dla klasyfikatora ten sam.

Współczynnik Błędu = 1.8640e-02