Klasyfikacja Optymalna Bayesa

ROB laboratorium nr 2

Marcin Skrzypkowski

1. **Sprawdzanie Danych**

Aby wybrać próbki odstające wykorzystano funkcję min, która zwróciła indeksy próbek mające najmniejsze wartości w każdej z etykiet. Następnie przetestowane zostały wszystkie próbki: dla każdej z nich porównana była wartość wszystkich cech z sąsiadującymi z nią próbkami, jeśli różnica była niepokojąco duża, próbka została usunięta. W ten sposób stwierdzono, że próbki nr 186 i 642 należy usunąć.

2. **Klasyfikator Optymalny Bayesa**

W celu znalezienia cech potrzebnych do zbudowania klasyfikatora optymalnego Bayesa wykorzystano funkcję **plot2features**. Poszukiwano pary cech, które na wykresie przedstawiają **8** wyraźnie rozdzielonych grup próbek. Ostatecznie wybrano cechę **1**. oraz **3**. Poniżej znajduje się wykres z reprezentacją próbek na podstawie powyższych cech.

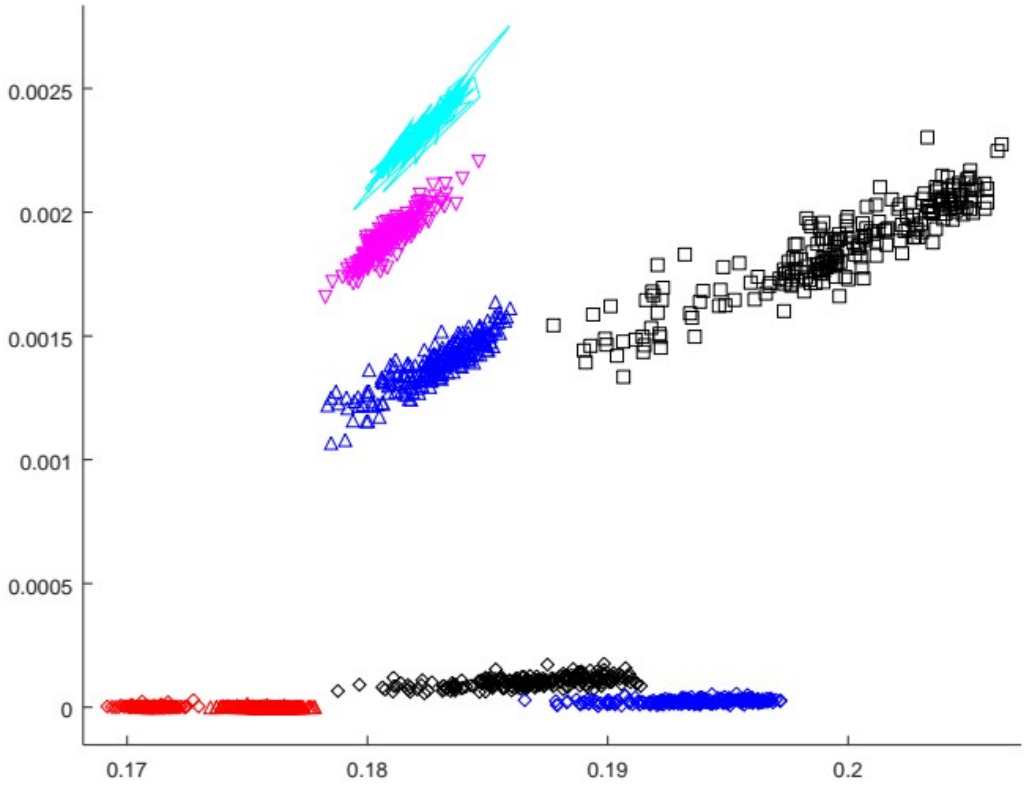

*Illustration 1: wykres **plot2features** cech 1. oraz 3. zbioru uczącego*

Następnie na podstawie wybranych cech zbudowano klasyfikator optymalny Bayesa. Zaimplementowano trzy funkcje liczące gęstość prawdopodobieństwa rozkładów:

- przy założeniu, że cechy są niezależne
- przy założeniu, że mamy do czynienia z rozkładem wielowymiarowym
- z wykorzystaniem okna Parzena

Poniżej znajdują się wyniki klasyfikacji optymalnej Bayesa dla wszystkich trzech wariantów. Przyjęto prawdopodobieństwa apriori jako równe **0.25** (**0.125** z powodu podziału na 8 klas – każda klasa została podzielona na pół). Jako cechy wybrano te dwie które dały najlepsze rozdzielenie próbek na wykresie **plot2features**, a więc cechę **1.** oraz **3.**

metoda	*niezależna*	*wielowymiarowa*	*Parzen*
Współczynnik błędu	*0.0252193*	*0.0038377*	*0.0241228*

Na podstawie powyższego zestawienia można by stwierdzić, że podejście wielowymiarowe daje najlepsze wyniki klasyfikacji. Dodatkowo okno Parzena i podejście cech niezależnych zwracają podobne współczynniki, więc warto zastanowić się nad manipulacją wielkości okna w ostatniej metodzie. Wyniki eksperymentów zostały przedstawione w punkcie **4.**

3. Wpływ doboru próbek w zbiorze testowym na wyniki klasyfikacji

W celu sprawdzenia zachowania wyników klasyfikacji na powyższą zmianę przeprowadzono klasyfikację optymalną Bayesa dla **0.1, 0.25** oraz **0.5** zbioru uczącego. Ponieważ podzbiory muszą zostać wylosowane, dla każdej z tych wielkości należało przeprowadzić eksperymenty kilkukrotnie a następnie wyprowadzić średnią dla każdego z klasyfikatorów, dla wszystkich trzech funkcji prawdopodobieństwa. Poniżej znajdują się wyniki przeprowadzonych testów - średnie współczynników błędu dla każdego podzbioru oraz każdej metody:
Każdy podzbiór został przetestowany 5 razy.

Wielkość podzbioru	niezależna	wielowymiarowa	Parzen
0.1	0.029167	0.010088	0.089912
0.25	0.025658	0.0057018	0.058882
0.5	0.031579	0.0044956	0.038706

Poniżej znajdują się wartości odchyleń standardowych dla opisanych powyżej przeprowadzonych eksperymentów:

Wielkość podzbioru	niezależna	wielowymiarowa	Parzen
0.1	0.0022066	0.0062941	0.012061
0.25	0.0026407	0.0013761	0.0046617
0.5	0.0082383	0.0014191	0.0047257

Zarówno metoda wielowymiarowa i Parzena są podatne na zmiany w wielkości zbioru uczącego. Metoda niezależna nie pokazuje tak odmiennych wyników, nawet przy 10% zestawu uczącego współczynnik błędu niewiele się zmienił.

4. Wpływ szerokości okna Parzena na na klasyfikację

W celu sprawdzenia wielkości okna Parzena na jakość klasyfikacji wybrano 5 różnych szerokości okien i wyznaczono współczynnik błędu dla każdej z tych wielkości. Poniżej przedstawiono wyniki eksperymentów.

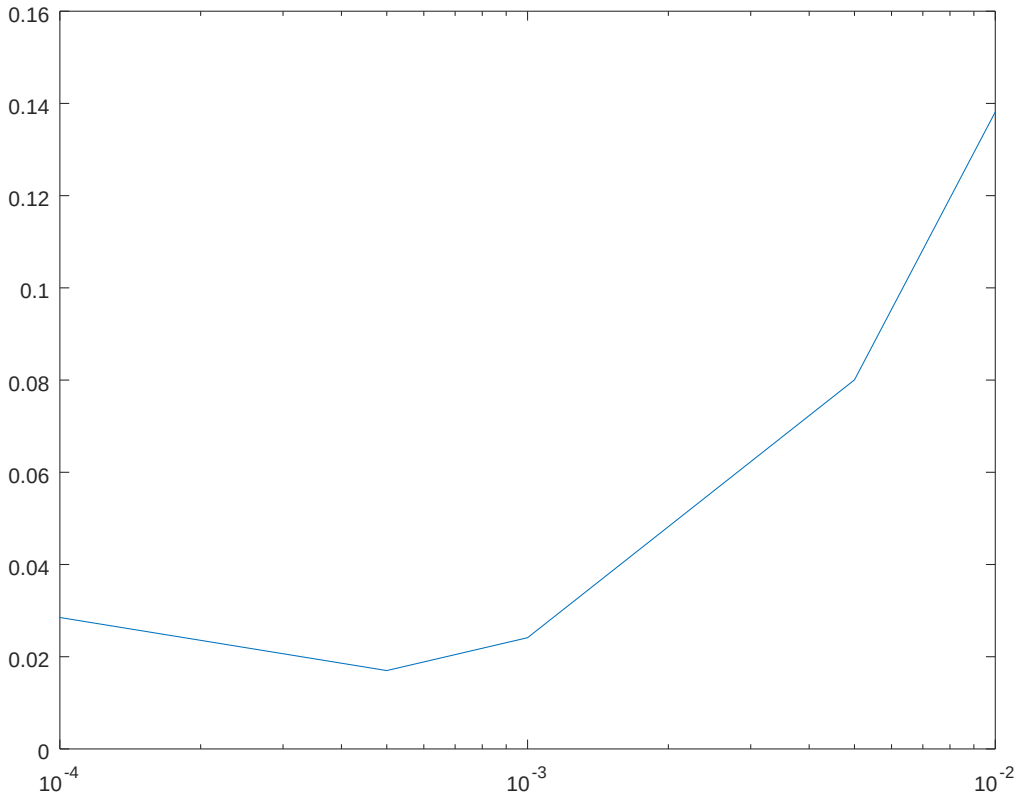

Tabela wartości, która psłużyła to utworzenia powyższego wykresu:

Szerokość h_1	0.0001	0.0005	0.001	0.005	0.01
Współczynnik błędu	0.028509	0.016996	0.024123	0.080044	0.13816

Na podstawie wykresu i tabeli można wywnioskować, że zmniejszanie okna Parzena poprawia wyniki klasyfikacji, ale tylko do pewnego stopnia. Zbyt małe okno również negatywnie wpływa na wyniki. Zbyt duże okna mogą spowodować złą klasyfikację próbek, natomiast zbyt małe mogą sprawić, ocena klasy będzie zbyt rygorystyczna.

5. Testy dla innych prawdopodobieństw *a priori*

W celu sprawdzenia wpływu zmian prawdopodobieństw powtórzone zostały eksperymenty z punktu 3. Poniżej podano wyniki dla każdej z trzech metod obliczania gęstości prawdopodobieństwa.

metoda	*niezależna*	*wielowymiarowa*	*Parzen*
Współczynnik błędu	*0.021199*	*0.0033626*	*0.02807*

Wyniki są zaskakująco zbliżone do eksperymentów przed zmianą prawdopodobieństw, można by przypuszczać, że wprowadzenie nowych wartości *a priori* zwiększyłoby błędy klasyfikacji, jeśli wiemy że wszystkich próbek jest po równo w każdej klasie. Możliwe więc, że niektóre próbki zostały źle zakwalifikowane.

6. Klasyfikator 1-NN

W celu zdecydowania, czy dane należy znormalizować, sprawdzone zostały odchylenia standardowe cech wybranych wcześniej w punkcie 2. Sprawdzono też błąd zwracany przez klasyfikator 1-NN przed i po normalizacji.

W celu normalizacji wykorzystano następujący wzór:

$$\frac{(x_i - x_{mean})}{\sigma}$$

Gdzie σ jest odchyleniem standardowym a x_{mean} średnią wartością zbioru. Poniżej podano wyniki testów:

	Różnica odchyleń standardowych cechy 1. i 3.
Przed normalizacją	0.0089521
Po normalizacji	8.8818e-16

	Błąd klasyfikatora 1-NN
Przed normalizacją	0.017544
Po normalizacji	0.0049342

Można stwierdzić, że warto przeprowadzić normalizację w powyższym przypadku, zmniejszyło to błąd o jeden rząd wielkości. Przed jej wykonaniem klasyfikator 1-NN dał podobne rezultaty co klasyfikator Bayesa z niezależnymi cechami oraz wykorzystaniem metody Parzena, natomiast po normalizacji błąd jest tego samego rzędu co klasyfikator optymalny Bayesa z metodą wielowymiarową.