ROB Klasyfikacja Bayesa. Paweł Paczuski (271082)

March 24, 2019

1 Sprawozdanie

1.1 Eliminacja wartości odstających

Ze zbioru uczącego usunięto próbki 642 (wartości skrajnie małe dla większości cech) oraz 186 (wartości skrajnie duże dla większości cech). Praktyczność usunięcia tych dwóch próbek widoczna jest skali wykresów generowanych przez plot2features. Przed eliminacją dane tworzyły dwa zgrupowania punktów: odstających i reszty. Po usunięciu wartości odstających skala wykresów pozwala zauważyć to, jak wyglądają różnice między klasami.

1.2 Błędy dla kombinacji

Za pomocą nchoosek(liczba_cech,2) wygenerowano wszystkie możliwe pary cech i policzono dla nich błędy trzech podejść do generowania parametrów dla klasyfikatora Bayesa:

kombinacja	pdfindep	pdfmulti	pdfparzen w=0.001
2 3	0.17818	0.17873	0.21765
2 4	0.02631	0.00493	0.02412
2 5	0.07565	0.05975	0.27083
2 6	0.13706	0.10197	0.27686
2 7	0.10252	0.09758	0.27686
2 8	0.22752	0.22862	0.27686
3 4	0.02138	0.02083	0.01699
3 5	0.13158	0.12774	0.23794
3 6	0.15570	0.14912	0.31469
3 7	0.16228	0.15406	0.31469
3 8	0.28125	0.28399	0.31469
4 5	0.15406	0.14583	0.23958
4 6	0.15844	0.15132	0.26864
4 7	0.23136	0.21217	0.26864
4 8	0.21272	0.21272	0.26864
5 6	0.37500	0.31305	0.47149
5 7	0.32511	0.26919	0.47149
5 8	0.38158	0.37829	0.47149
6 7	0.32675	0.28947	0.54331
6 8	0.40077	0.40241	0.57346
7 8	0.36842	0.36678	0.54331

Najlepiej klasyfikowana jest zatem kobinacja cech 3 i 4.

1.3 Klasyfikacja w oparciu o dwie wybrane cechy

Wybrano cechy 3 i 4 i dla nich policzono błędy klasyfikatora Bayesa. Dodatkowo, za pomocą funkcji toClient, dokonano projekcji wyników klasyfikatora na cztery etykiety dostarczone przez klienta (oryginalnie w danych było ich 8, co było spowodowane różnicą w sposobie otrzymania danych), co skutkowało zmniejszeniem błędu klasyfikacji dlatego, że błędne przestały być wyniki, którym wcześniej klasyfikator przypisywał równoważne dla klienta klasy.

Dla prawdopodobieństwa `apriori = 0.25`

cechy	pdfindep	pdfmulti	pdfparzen w=0.001
3 4	0.021382	0.020833	0.016996

używając etykiet klienta

cechy	pdfindep	pdfmulti	pdfparzen w=0.001
3 4	0.010417	0.009868	0.006031

1.4 Redukcja zbioru trenującego

Zbadano wpływ redukcji zbioru trenującego na wyniki klasyfikacji. Tabele zawierają uśrednione wyniki dla klasyfikacji z użyciem zredukowanego zbioru trenującego dla pięciu prób.

czesc	pdfindep	std	min	max
0.1000	0.0257	0.0049	0.0203	0.0323
0.2500	0.0205	0.0011	0.0192	0.0219
0.5000	0.0213	0.0008	0.0203	0.0225

czesc	pdfmulti	std	min	max
0.1000	0.0258	0.0046	0.0214	0.0329
0.2500	0.0202	0.0009	0.0192	0.0214
0.5000	0.0197	0.0013	0.0181	0.0214

czesc	pdfparzen w=0.001	std	min	max
0.1000	0.0402	0.0044	0.0351	0.0471
0.2500	0.0283	0.0022	0.0252	0.0312
0.5000	0.0203	0.0008	0.0197	0.0214

używając etykiet klienta

czesc	pdfindep	std	min	max
0.1000	0.0150	0.0057	0.0088	0.0241
0.2500	0.0102	0.0010	0.0088	0.0115
0.5000	0.0103	0.0010	0.0088	0.0110

czesc	pdfmulti	std	min	max
0.1000	0.0135	0.0053	0.0082	0.0225
0.2500	0.0089	0.0012	0.0071	0.0099
0.5000	0.0098	0.0011	0.0088	0.0115

czesc	pdfparzen w=0.001	std	min	max
0.1000	0.0249	0.0020	0.0219	0.0214
0.2500	0.0167	0.0024	0.0137	0.0214
0.5000	0.0091	0.0012	0.0071	0.0214

Im większy zbior trenujący, tym mniejszy błąd klasyfikacji. Największa wrażliwość na zwiększanie rozmiaru zbioru trenującego obserwowana jest dla metody z użyciem okna Parzena – wartość odchylenia standardowego używanego do oszacowania pdf uzależniona jest bezpośrednio od liczby próbek.

1.5 Różne szerokości okna Parzena

parzen width	error
0.000100	0.014254
0.000500	0.016996
0.001000	0.020285
0.005000	0.044408
0.010000	0.043311

używając etykiet klienta

parzen width	error
0.000100	0.004386
0.000500	0.005482
0.001000	0.008772
0.005000	0.031250
0.010000	0.030154

Im mniejsza szerokość okna Parzena, tym mniejszy błąd klasyfikacji.

1.6 Dwukrotnie większe prawdopodobieństwo apriori dla maści czarnych

Uśrednione dla pięciu prób wyniki klasyfikacji uzyskane przy użyciu zredukownego zbioru testowego.

pdfindep	std	min	max
0.0156	0.0034	0.0102	0.0183

pdfmulti	std	min	max
0.0167	0.0024	0.0132	0.0190

pdfparzen w=0.001	std	min	max
0.0186	0.0032	0.0139	0.0227

etykiety klienta:

pdfindep	std	min	max
0.0076	0.0020	0.0044	0.0095

pdfmulti	std	min	max
0.0082	0.0015	0.0066	0.0095

pdfparzen w=0.001	std	min	max
0.0107	0.0010	0.0095	0.0117

Zmiana pradobodobieństwa (oraz redukcja części zbioru testowego) miała pozytywny wpływ na wyniki klasyfikacji, co pozwala stwierdzić, że użyte prawdopodobieństwo lepiej oddaje naturę klasyfikowanych danych.

Macierz pomyłek klasyfikatora dla surowych danych testowych

228	0	0	0	0	0	0	0
0	226	0	0	0	1	0	1
12	0	207	0	0	0	9	0
1	0	2	225	0	0	0	0
0	0	0	0	227	0	0	1
0	5	0	0	0	223	0	0
1	0	5	0	0	0	222	0
0	1	0	0	0	0	0	227

Macierz pomyłek klasyfikatora po wspomnianych modyfikacjach:

```
228     0     0     0     0     0     0     0
  0   113     0     0     0     1     0     0
  4     0   105     1     0     0     4     0
  1     0     1   226     0     0     0     0
  0     0     0     0   226     0     0     2
  0     2     0     0     0   112     0     0
  1     0     2     0     0     0   111     0
  0     1     0     0     0     0     0   227
```

Zmiana prawdopodobieństwa apriori sprawiła, że więcej oryginalne błędne przypadków klasyfikacji zostało sklasyfikowanych poprawnie, np. wcześniej znacznie więcej obiektów klasy 3 było klasyfikowanych jako 1, po zmianie prawdopobieństw (oraz liczności zbioru testowego w obrębie maści czarnych) zmniejszył się udział takich przypadków w ogólnym błędzie klasyfikatora, poprawiając ogólnie jakość jego klasyfikacji.

1.7 Normalizacja danych

Odchylenie standardowe dla cech 3 i 4: `0.00092`, `0.00095`

Odchylenie standardowe w poszczególnych klasach cech 3 i 4

klasa	std cecha 3	std cecha 4
1	0.000063	0.000186
2	0.000162	0.000003
3	0.000022	0.000085
4	0.000012	0.000098
5	0.000023	0.000021
6	0.000255	0.000002
7	0.000010	0.000112
8	0.000126	0.000009

błędy klasyfikacji dla znormalizowanych cech 3 i 4

cechy	pdfindep	pdfmulti	pdfparzen w=0.001
3 4	0.021382	0.020833	0.020285

błędy klasyfikacji dla znormalizowanych cech 3 i 4 korzystając z etykiet klienta

cechy	pdfindep	pdfmulti	pdfparzen w=0.001
3 4	0.010417	0.009868	0.004386

Bardzo małe wartości odchylenia standardowego w klasach cech 3 i 4 sprawiają, że normalizacja danych nie jest specjalnie potrzebna, o czym świadczy brak różnicy w wynikach klasyfikacji dla przyjętej liczby cyfr znaczących.

1.8 1NN vs Bayes

Błąd klasyfikatora 1NN wyniósł: `ercf_1nn = 0.018092` Gdy zastosujemy etykiety klienta: `ercf_1nn_client = 0.004385`

Klasyfikator uzyskje 1NN porównywalne wyniki do klasyfikatora Bayesa zakładającego `apriori = 0.25`, jednak przy przejściu do etykiet klienta widać zauważalną różnicę błędów na korzyść 1NN, co sugeruje, że problematyczne dla 1NN punkty pochodzą z równoważnych dla klienta klas oryginalnie traktowanych jako różne.

Normalizacja danych dla 1NN pogorszyła minimalnie wyniki klasyfikacji `ercf_1nn_norm = 0.021382` (stały się identyczne jak dla klasyfikatora Bayesa), używając etykiet klienta: `ercf_1nn_client_norm = 0.004386` (różnica dopiero na ostatniej cyfrze znaczącej).