Klasyfikacja bayesa

Szymon Bugaj

1 czerwca 2015

Streszczenie

Sprawozdanie z laboratorium poświęconego klasyfikatorom Bayesa z przedmiotu ROB (Rozpoznawanie obrazów). Zaimplementowane zostały trzy typy klasyfikatorów Bayesa:

- z założeniem o niezależności cech, gdzie każda ma rozkład normalny,
- z założeniem o wielowymiarowym rozkładzie normalnym cech,
- z aproksymacją rozkłada za pomocą okna Parzena

Spis treści

1 Usunięcie wartości odstających ze zbioru danych

Nieusunięcie wartości odstających znacząco wpływa na otrzymane wyniki klasyfikacji.

cechy	błąd naiwny	błąd wielowym	błąd parzen
(2, 3)	20.61%	20.50%	14.36%
(2, 4)	19.68%	4.93%	1.48%
(2, 5)	21.44%	24.18%	17.54%
(2, 6)	28.51%	28.45%	20.12%
(2, 7)	15.73%	16.23%	20.12%
(2, 8)	24.78%	24.73%	20.12%
(3, 4)	15.02%	3.29%	0.33%
(3, 5)	14.86%	14.20%	23.41%
(3, 6)	18.09%	16.34%	24.95%
(3, 7)	15.52%	15.13%	24.95%
(3, 8)	25.22%	25.11%	24.95%
(4, 5)	30.54%	19.24%	11.73%
(4, 6)	33.33%	26.43%	24.23%
(4, 7)	26.43%	24.89%	24.23%
(4, 8)	29.33%	29.22%	24.23%
(5, 6)	34.21%	32.13%	46.00%
(5, 7)	33.99%	27.14%	46.00%
(5, 8)	30.92%	31.41%	46.00%
(6, 7)	30.92%	25.99%	53.07%
(6, 8)	34.87%	35.03%	56.03%
(7, 8)	29.71%	29.82%	53.07%

Tabela 1: Błąd na zbiorze testowym, dla trzech rodzaji klasyfikatorów dla wszystkich możliwych par cech. $h1 = 0.0007, apriori = [0.25, 0.25, 0.25, 0.25]$

```
%Clean data
train(186,:) = [];
train(640,:) = [];
```

2 Wybór podzbioru cech do klasyfikacji

Wykonałem testy dla wszystkich podzbiorów 2 elementowych wybranych ze zbioru cech dla każdego typu klasyfikatora Bayesa. Wyniki znajdują się w tabeli 1. Najlepsze wyniki uzyskałem dla cech 3 i 4.

wielkość zbioru	błąd naiwny	błąd wielowym	błąd parzen
1.00	15.02%	3.29%	0.33%
0.75	15.08%	3.12%	0.44%
0.50	14.91%	3.02%	1.21%
0.25	14.75%	1.81%	1.59%

Tabela 2: Błąd dla trzech rodzaji klasyfikatorów uczonych na podzbiorach zbioru uczącego. $h1 = 0.0007, apriori = [0.25, 0.25, 0.25, 0.25]$

3 Wielkość zbioru uczącego

Zbiór uczący zmniejszany był zgodnie z algorytmem (po redukcji przykładów różnych klas było dokładnie tyle samo):

```
%Reduce train set
train_t = [];
for i = 1:4
    bool_cl = train(:,1)==i;

    part = train(bool_cl,:);
    size(part);
    part = part(1:size(part,1).*t(ireduce),:);

    train_t = [train_t; part];
end
```

Wyniki zaprezentowane są w tabeli 2. Nie jest dla mnie zrozumiałe dlaczego błąd dla klasyfikatora z wielowymiarowym rozkładem normalnym maleje, gdy zbiór uczący zmniejsza się - osiągając dwa razy lepszy wynik dla tylko ćwiartki zbioru uczącego.

W każdym razie wyniki są zadowalające nawet tylko dla jednej czwartej zbioru uczącego.

4 Dobór parametru h1 dla klasyfikatora z oknem parzena

Wyniki znajdują się w tabelach 3 i 4. Wybrano 3 cechy: 3,4 i 5. Najlepszy błąd klasyfikacji otrzymano dla $h1 = 0.0007 + / - 0.0001$. Prawdopodobieństwa apriori były równe sobie i wykorzystano cały zbiór uczący.

h1	błąd parzen
0.0010	0.33%
0.0110	15.95%
0.0210	24.89%
0.0310	38.32%
0.0410	38.38%
0.0510	38.43%
0.0610	38.38%
0.0710	38.38%
0.0810	38.38%
0.0910	38.38%

Tabela 3: Błąd dla klasyfikatora z oknem parzena, dla różnych wielkości parametru h1.

h1	błąd parzen
0.0005	0.33%
0.0006	0.27%
0.0007	0.27%
0.0008	0.27%
0.0009	0.33%
0.0010	0.33%
0.0011	0.33%
0.0012	0.38%

Tabela 4: Błąd dla klasyfikatora z oknem parzena, dla różnych wielkości parametru h1.

apriori	błąd naiwny	błąd wielowym	błąd parzen
$[0.25, 0.25, 0.25, 0.25]$	15.02%	3.29%	0.33%
$[0.17, 0.33, 0.33, 0.17]$	10.67%	4.39%	0.73%

Tabela 5: Wyniki dla klasyfikatorów przy różnych prawdopodieństwach apriori

5 Zmiana prawdopodobieństw apriori oraz odpowiedni wybór podzbioru z oryginalnego zbioru uczącego

Przeprowadziłem testy dla prawdopodobieństw apriori = [0.17, 0.33, 0.33, 0.17] oraz po odpowiedniej zmianie zbioru uczącego i testowego (by prawdopodieństwa apriori zgadzały się z częstością występowania elementów w zbiorach).

```
%Reduce train set
%prepare dataset with given apriori
apriori = [0.5;1;1;0.5];
train_t = [];
for i = 1:size(apriori, 1)
    bool_cl = train(:,1)==i;

    part = train(bool_cl,:);
    size(part);
    part = part(1:size(part,1)*apriori(i),:);

    train_t = [train_t; part];
end
test_t = [];
for i = 1:size(apriori, 1)
    bool_cl = test(:,1)==i;

    part = test(bool_cl,:);
    part = part(1:size(part,1)*apriori(i),:);

    test_t = [test_t; part];
end
```

Rezultaty nie uległy zmianie

6 Porównanie wyników do klasyfikatora 1-NN

DLa klasyfikatora 1-NN uzyskałem błąd o wartości 0.49%. Najlepszym wynikiem uzyskanym przy użyciu klasyfikatora Bayesa jest 0.27%.

7 Podsumowanie

Podczas klasyfikacji bayesowskiej należy pamiętać o:

- jeżeli wybierany jest podzbiór cech, wybór jest istotny,
- przy klasyfikatorze z oknem parzena wybór h1 jest bardzo istotny.