Alicja Kubiszyn, 318685

Naiwny Klasyfikator Bayesa

1. Opis algorytmu

Algorytm składa się z dwóch części – trenującej i uczącej. W części trenującej najpierw obliczane jest prawdopodobieństwo a-priori wystąpienia danej klasy. Jest ono obliczane jako suma wystąpień danej klasy w zbiorze trenującym podzielona przez moc tego zbioru. Następnie obliczane jest prawdopodobieństwo, że obiekt X należy do danej klasy, znając wartość atrybutu. Ponieważ wartości atrybutów klas nie są dyskretne, najpierw algorytm próbuje zamienić je na wartości dyskretne. Robi to poprzez posortowanie całego zbioru trenującego, gdzie kluczem są kolejne atrybuty, i podzielenie go na grupy, w których wartości danego atrybutu mieszczą się w określonych, stałych przedziałach. Następnie zliczane są wystąpienia danej klasy i dzielone są przez moc całego podzbioru. Prawdopodobieństwa te są wyliczane dla każdego atrybutu.
Następnie algorytm przechodzi do testowania. Dla każdego obiektu ze zbioru X_train przypisuje się prawdopodobieństwa, że dany obiekt należy do danej klasy oraz mnożone są przez prawdopodobieństwa atrybutów.

2. Opis planowanych eksperymentów

Napisany przeze mnie algorytm zależy tylko od jednego parametru: stosunku wielkości zbioru trenującego do testującego. W przeprowadzonych doświadczeniach zostanie sprawdzony wpływ tego parametru na jakość uzyskanych wyników.

3. Analiza wyników

1. Parametr = 0.2

```
Percentage of failures:  15.833333333333332
Percentage of failures:  11.666666666666666
Percentage of failures:  16.666666666666664
Percentage of failures:  15.0
Percentage of failures:  5.0
Mean:  12.833333333333332
```

Średni procent pomyłek wynosi 12,8%. Jednak w jednym przypadku procent ten wynosi tylko 5 %.

2. Parametr = 0.1

```
Percentage of failures:  28.14814814814815
Percentage of failures:  28.888888888888886
Percentage of failures:  11.851851851851853
Percentage of failures:  14.074074074074074
Percentage of failures:  19.25925925925926
Mean:  20.444444444444446
```

Średnia się pogorszyła – wynosi teraz 20, 44%. Jednak w dwóch przypadkach wynik ten wynosił aż 28%.

3. Parametr = 0.05

```
Percentage of failures:  37.06293706293706
Percentage of failures:  27.972027972027973
Percentage of failures:  35.66433566433567
Percentage of failures:  41.25874125874126
Percentage of failures:  44.75524475524475
Mean:  37.34265734265735
```

Im mniejszy parametr, czyli im mniejszy zbiór trenujący tym gorsze wyniki. Najgorszy wynik algorytm uzyskał w piątej próbie – blisko połowa wyników była nieprawidłowa. Średnia to 37.34%

4. Parametr = 04.

```
Percentage of failures:  10.0
Percentage of failures:  7.777777777777778
Percentage of failures:  5.555555555555555
Percentage of failures:  11.11111111111111
Percentage of failures:  10.0
Mean:  8.88888888888889
```

W tym wypadku jakość algorytmu znacząco się polepszyła – w prawie wszystkich próbach procent pomyłek jest mniejszy od 10 %.

5. Parametr = 0.5

```
Percentage of failures:  10.666666666666668
Percentage of failures:  4.0
Percentage of failures:  2.666666666666667
Percentage of failures:  2.666666666666667
Percentage of failures:  6.666666666666667
Mean:  5.333333333333334
```

Jeśli zbiór trenujący i testujący zostanie podzielony na pół, średnia wyników wyniesie tylko 5% - w dwóch przypadkach będzie to jedynie 2, 67 %.

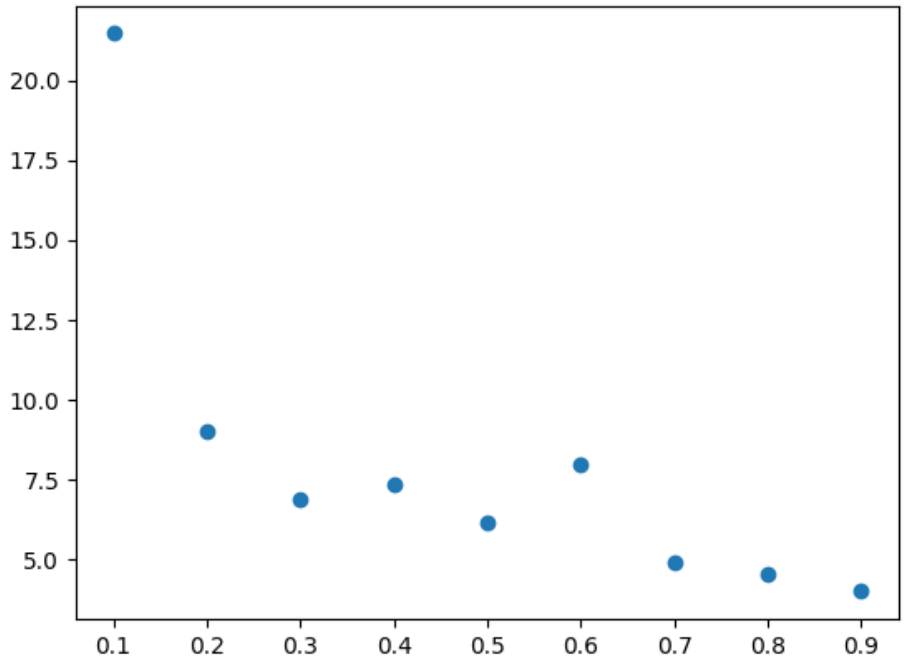
6. Parametr = 0.6

```
Percentage of failures:  8.333333333333332
Percentage of failures:  5.0
Percentage of failures:  5.0
Percentage of failures:  6.666666666666667
Percentage of failures:  10.0
Mean:  7.0
```

Ciekawy wynik wyszedł w tym przypadku – średnia wyników jest większa niż w przypadku parametru równego 0.5.

7. Wykres

Wykres przedstawia zależność stosunku wielkości zbioru trenującego do uczącego i średniej procentu pomyłek.



Im większy ten stosunek tym mniejsza ilość pomyłek

4. Wnioski

Im większy jest zbiór trenujący tym lepsze są otrzymywane wyniki. Jednak nie jest opłacalne dzielić cały zbiór w ten sposób, ponieważ celem powinno być jak najwięcej dobrze przewidzianych obiektów – lepiej dobrze przewidzieć 110 na 120 obiektów niż 31 na 32. Jak widać na wykresie, od około x = 0.2 można zaobserwować zbliżone wyniki, więc nie ma sensu dobierać większy parametr niż 0.2 lub 0.3.