Zadanie 7. Modele Bayesowskie

Julia Jodczyk

11 lutego 2022

1 Treść zadania

Zaimplementuj naiwny klasyfikator Bayesa oraz zbadaj działanie algorytmu w zastosowaniu do zbioru danych Iris Data Set. Pamiętaj, aby podzielić zbiór danych na zbiór trenujący oraz uczący.

2 Wstęp

Naiwny klasyfikator Bayesa jest klasyfikatorem statystycznym bazującym na twierdzeniu Bayesa. Nadaje się szczególnie do rozwiązywania problemów z wieloma wymiarami na wejściu. Klasyfikator opiera się na założeniu, że atrybuty na wejściu są od siebie niezależne.

Klasyfikator rozwiązuje problemy w następujący sposób: mając dany zbiór atrybutów $X = (x_1, x_2, ..., x_n)$ oraz zbiór klas $\mathbb{Y}$ przypisujemy X taką klasę $y \in \mathbb{Y}$, dla której

$$P(Y = y|X = (x_1, x_2, ..., x_n)) \tag{1}$$

osiąga wartość maksymalną.

Do znalezienia tej wartości $P(Y = y|X)$ wykorzystujemy twierdzenie Bayesa:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{2}$$

Ponieważ wartość $P(X)$ dla wszystkich rozpatrywanych wartości y jest taka sama, to można pominąć ją w obliczeniach i szukać jedynie $argmax_y(P(X|Y)P(Y))$.

Wartość $P(X|Y)$ można łatwo obliczyć korzystając z założenia o niezależności atrybutów wejściowych. Wtedy

$$P(X|Y) = \Pi P(X = x_k|Y) \tag{3}$$

Istnieją trzy typy Naiwnego klasyfikatorów Bayesa:

- Bernoulliego - stosowany, gdy atrybuty wejściowe mają wartości binarne,

- Kategoryczny - dla przypadków, w których atrybuty na wejściu mają wartości dyskretne,
- Gaussowski - wykorzystywany, gdy atrybuty wejściowe mają wartości rzeczywiste.

W przypadku zbioru danych Iris Data Set musimy zastosować klasyfikator Gaussowski. Zakładamy więc, że atrybuty na wejściu mają rozkład Gaussa, czyli poszczególne prawdopodobieństwa $P(X = x_k|Y)$ liczymy ze wzoru

$$P(X_i = x_k|Y) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(x-m)}{2\sigma^2}\} \tag{4}$$

gdzie σ^2 to wariancja, a m to średnia wartość oczekiwana.

3 Opis implementacji

W klasie `naive_bayes_classifier` znajduje się właściwa logika algorytmu. Na początku na podstawie zbioru treningowego obliczane są średnia wartość oczekiwana, wariancja oraz prawdopodobieństwo wstępne ($P(Y)$). Następnie dla każdego X_i ze zbioru testowego wyliczane jest prawdopodobieństwo a posteriori ($P(Y|X_i)$). Na koniec wybierane jest takie $y \in Y$, dla którego prawdopodobieństwo a posteriori było największe.

Zbiór danych został podzielony na treningowy i testowy w proporcji 0,8:0,2.

4 Skuteczność rozwiązania

Skuteczność różniła się w zależności od wartości `random_state` w funkcji `train_test_split` z biblioteki `sklearn` dzielącej zbiór danych na treningowy i testowy. Większość wartości znajdywała się w zakresie od 93% do 100%.

5 Wnioski

- Naiwny klasyfikator Bayesa ma bardzo krótki czas działania, oraz jest prosty do zaimplementowania.
- Przez brak parametrów nie jest trudny w stosowaniu.
- Mimo prostoty często działa lepiej od innych algorytmów klasyfikujących.
- Klasyfikator nie potrzebuje zbyt wielu danych by osiągać dużą skuteczność - przy małym zbiorze danych (750 próbek) skuteczność była bardzo wysoka.
- Klasyfikator będzie działał gorzej, jeśli atrybuty wejściowe będą od siebie mocno zależne.