WSI-24L-G104

Adam Kwaśnik 05.06.2024

Zadanie 7.

Cel zadania

Celem zadania jest zaimplementowanie przez Państwa Naiwnego Klasyfikatora Bayesowskiego, a w szczególności jego konkretnej odmiany, jaką jest Gaussowski Naiwny Bayes. Będą Państwo analizować i klasyfikować te same dane, które pojawiły się w zadaniu nr 4 - Dataset Wine.

Naiwny klasyfikator bayesowski to rodzaj klasyfikatora opartego na twierdzeniu Bayesa. Działa on na podstawie założenia, że cechy są niezależne, co często określa się jako “naiwne” założenie. Pomimo tego, że cechy rzadko są rzeczywiście niezależne, klasyfikatory te osiągają zadowalającą skuteczność w wielu problemach analitycznych.

Twierdzenie Bayesa pozwala obliczyć prawdopodobieństwo przynależności przykłady do konkretnej klasy na podstawie obserwowanych cech

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(B \mid A)\,\mathsf{P}(A)}{\mathsf{P}(B)}$$

- P(A|B) - prawdopodobieństwo wystąpienia zdarzenia A pod warunkiem zdarzenia B
- P(B|A) - prawdopodobieństwo wystąpienia zdarzenia B pod warunkiem zdarzenia A
- P(A) - a priori prawdopodobieństwo zdarzenia A
- P(B) - całkowite prawdopodobieństwo zaobserwowania zdarzenia B

Gaussowski Naiwny Bayes to odmiana stosowana do danych o rozkładzie ciągłym. Zakłada się, że wartości cech dla każdej klasy są zgodne z rozkładem normalnym.

W przypadku Gaussowskiego Naiwnego Bayesa prawdopodobieństwo warunkowe cechy x_i dla klasy y jest obliczane na podstawie rozkładu normalnego:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

- $P(x_i \mid y)$ prawdopodobieństwo warunkowe, że cecha x_i przyjmuje określoną wartość, biorąc pod uwagę, że przykład należy do klasy y.
- x_i wartość cechy i dla konkretnego przykładu x
- y klasa, do której należy dany przykład
- μ_y średnia wartość cechy x_i w klasie y
- σ_y^2 wariancja cechy x_i w klasie y

Państwa zadaniem jest implementacja klasyfikatora Gaussowskiego Naiwnego Bayesa oraz wykorzystanie zbioru treningowego do obliczenia parametrów algorytmu - średniej i wariancji. Następnie należy przetestować algorytm na zbiorze testowym.

Proszę zwrócić uwagę, że dataset zawiera wiele cech i kilka klas, dlatego będą musieli obliczyć państwo wiele różnych wariancji i średnich. Gdy obliczą Państwo te wartości możliwe będzie:

1. Obliczenie prawdopodobieństwa warunkowego dla wybranej cechy i klasy.
2. A następnie obliczenie łącznego prawdopodobieństwa cech z danego przykładu dla każdej klasy korzystając z prawdopodobieństw warunkowych.

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Pierwszy wzór przedstawia sposób obliczenia prawdopodobieństwa klasy y dla cech x1,...,xn

P(y) określa prawdopodobieństwo wystąpienia klasy y. Oblicza się to jako stosunek liczby wystąpień danej klasy do całkowitej liczby przykładów w zbiorze.

Drugi wzór przedstawia sposób wyboru klasy, do której należy przykład. Należy wybrać klasę z najwyższym posteriori prawdopodobieństwem jako przewidywaną klasę dla danego przykładu.

W celu uzyskania bardziej wiarygodnej oceny modelu, proszę zastosować kroswalidację korzystając z biblioteki scikit-learn.

Sprawozdanie

W sprawozdaniu należy zawrzeć wyniki skuteczności klasyfikatora dla metryk: accuracy, precision, recall, F1.

Proszę o porównanie wyników zaimplementowanego Naiwnego Bayessa z najlepszymi wynikami algorytmów wykorzystanych w zadaniu nr 4.