WSI

Zadanie 7

Autor:
Michał Paradowski

10.06.2024 (V1.0)

Contents

1 Zadanie

1.1 Cel zadania

Celem zadania jest zaimplementowanie Gaussowskiego Naiwnego Klasyfikatora Bayesowskiego oraz przeprowadzenie analizy i klasyfikacji danych z użyciem tego algorytmu. Dane pochodzą z zestawu danych "Wine", użytego wcześniej w zadaniu nr 4.

1.2 Opis algorytmu

Naiwny klasyfikator bayesowski to rodzaj klasyfikatora opartego na twierdzeniu Bayesa. Zakłada on, 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.

1.2.1 Twierdzenie Bayesa

Twierdzenie Bayesa pozwala obliczyć prawdopodobieństwo przynależności przykładu do konkretnej klasy na podstawie zaobserwowanych cech:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

gdzie:

- $P(A|B)$ - 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

1.2.2 Gaussowski Naiwny Bayes

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

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_{y,i}^2}} \exp\left(-\frac{(x_i - \mu_{y,i})^2}{2\sigma_{y,i}^2}\right)$$

gdzie:

- x_i - wartość cechy i dla konkretnego przykładu x
- y - klasa, do której należy dany przykład
- $\mu_{y,i}$ - średnia wartość cechy x_i w klasie y
- $\sigma^2_{y,i}$ - wariancja cechy x_i w klasie y

1.3 Implementacja

1.3.1 Obliczanie parametrów

Parametry algorytmu, tj. średnie i wariancje dla każdej cechy w poszczególnych klasach, zostały obliczone na podstawie zbioru treningowego.

Parametry modelu

Gaussowski Naiwny Bayes oblicza dwa podstawowe parametry dla każdej cechy j w klasie c:

- Średnia $\mu_{c,j}$
- Wariancja $\sigma^2_{c,j}$

Obliczanie parametrów

Średnia

Średnia dla cechy j w klasie c jest obliczana jako:

$$\mu_{c,j} = \frac{1}{N_c} \sum_{i=1}^{N_c} x_{i,j}$$

gdzie N_c jest liczbą próbek w klasie c, a $x_{i,j}$ jest wartością cechy j w próbce i należącej do klasy c.

Wariancja

Wariancja dla cechy j w klasie c jest obliczana jako:

$$\sigma^2_{c,j} = \frac{1}{N_c} \sum_{i=1}^{N_c} (x_{i,j} - \mu_{c,j})^2$$

1.3.1 Klasyfikacja

Klasyfikacja nowych przykładów odbywa się na podstawie obliczania prawdopodobieństwa przynależności do każdej klasy i wybierania klasy z najwyższym prawdopodobieństwem a posteriori:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

1.4 Eksperymenty i wyniki

1.4.1 Zbiór danych

Zbiór danych użyty do eksperymentów pochodzi z tego samego źródła co w zadaniu 4 i zawiera spis win podzielonych na 3 klasy ze względu na 13 cech. Tak więc po wytrenowaniu, dla każdej klasy GNKB musi wyznaczyć prawdopodobieństwo przynależności wina do tej klasy a następnie wybrać tą klasę dla której ów prawdopodobieństwo było najwyższe.

1.4.2 Metryki oceny

W celu oceny skuteczności klasyfikatora wykorzystano następujące metryki:

- Dokładność (accuracy)
- Precyzja (precision)
- Czułość (recall)
- Wskaźnik F1 (F1-score)

1.4.3 Wyniki

Wyniki algorytmu dla 3 różnych wartości random_seed. Za każdym razem przedstawiono uśrednione wyniki otrzymane po kroswalidacji. Dokładoność, Precyzja, Czułość i F1-score zostały uśrednione dla wszystkich 3 klas z pomocą biblioteki skicit-learn.

random_state	Accuracy	Precision	Recall	F1-score
10	0.9497	0.9528	0.9497	0.9495
10	0.9497	0.9528	0.9497	0.9495
20	0.9497	0.9528	0.9497	0.9495

Table 1.1: Wyniki klasyfikacji dla poszczególnych klas. random_state = 0

Widać że wyniki są bardzo dobre a losowość ma na nie pomijalny wpływ.

1.5 Porównanie z innymi algorytmami

Porównanie wyników zaimplementowanego Gaussowskiego Naiwnego Bayesa z najlepszymi wynikami algorytmów wykorzystanych w zadaniu nr 4. Najlepsze wyniki otrzmano dla algorytmu SVM z jądrem 'linear', parametrem $C = 3$ oraz max_iterations $= 10^3$.

Accuracy	Precision	Recall	F1-score
0.9458	0.9536	0.9458	0.9463

Tutaj wyniki zostały uśrednione dla 3 różnych wartości random_seed.

Widać, że w każdym kryterium Gaussowski Naiwny Bayes dawał niemal identyczne wyniki do algorytmy SVM. Jego zaletą jest fakt iż nie trzeba dobierać parametrów w celu uzyskania najlepszych wyników co jest czasochłonne kiedy stosujemy SVM. Dodatkowo algorytm Bayesa wytrenował się i uzyskał wyniki w jedyne 2.199 sekundy w przeciwieństwie do algorytmu SVM który potrzebował 3.348 sekund. W tym przypadku więc zdecydowanie opłaca się zdecydować na algorytm Bayesa.

List of Figures