

WSI - Modele bayesowskie

Marcin Szymczak 314910

1 Opis badanej metody

Naiwny klasyfikator Bayesa to probabilistyczny algorytm klasyfikacji oparty na twierdzeniu Bayesa, które można zapisać w postaci:

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \quad (1)$$

Gdzie:

- $P(Y|X)$: Posteriori - prawdopodobieństwo klasy Y przy danych cechach X .
- $P(X|Y)$: Prawdopodobieństwo zaobserwowania cech X przy założeniu klasy Y .
- $P(Y)$: Prior - Prawdopodobieństwo a priori klasy Y .
- $P(X)$: Prawdopodobieństwo wystąpienia cech X (w klasyfikacji jest pomijane, ponieważ jest stałe dla wszystkich klas).

Pseudokod:

1. Dla każdej klasy Y :

- Obliczane jest prior $P(Y)$.
- Dla każdej cechy X_i , obliczane jest $P(X_i|Y)$ zakładając rozkład Gaussa:

$$P(X_i|Y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \quad (2)$$

2. Dla nowej próbki X :

- Obliczane jest $P(Y|X)$ dla każdej klasy Y .
- Wybierana jest klasa z największym $P(Y|X)$.

2 Planowane eksperymenty numeryczne

Zbadanie dokładności na data set Iris. Sprawdzenie wpływu podziału danych na set treningowy i testowy.

3 Warunki eksperymentów numerycznych

- Podział danych 60%-40%
- Podział danych 70%-30%
- Podział danych 80%-20%
- Każdy podział testowany po 10 razy.

4 Otrzymane Wyniki

Podział danych	Accuracy (%)
60%-40%	96.67
70%-30%	97.78
80%-20%	100.0

Tabela 1: Wpływ podziału danych na dokładność modelu.

5 Wnioski

- Zbiór danych Iris jest dość małym zbiorem danych przez co mamy bardzo dużą dokładność.
- Dla tego zbioru danych model jest bardzo stabilny i podział nie wpływa znacząco na wysoką jakość modelu.
- Żeby rzeczywiście sprawdzić skuteczność modelu, należało by go przetestować na większym zbiorze danych.