

Wskazówki do ćwiczenia 7.

WSI - ćwiczenia 2022Z

Jakub Łyskawa Stanisław Pawlak

January 18, 2023

Prawdopodobieństwo klasy C dla próbki $\mathbf{x}$

$$P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$$

Prawdopodobieństwo klasy C dla próbki $\mathbf{x}$

$$P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$$

$P(\mathbf{x})$ jest dla każdej klasy takie samo,
więc klasa najbardziej prawdopodobna to

$$\operatorname{argmax}_C P(C)P(\mathbf{x}|C)$$

Naiwny klasyfikator Bayesa

Jeżeli się założy, że cechy są od siebie niezależne, to:

$$P(\mathbf{x}|C) = P(x_1|C)P(x_2|C)\dots$$

Więc szukana klasa:

$$\operatorname{argmax}_C P(C) \prod_i P(x_i|C)$$

Skąd wziąć prawdopodobieństwa?

$$\operatorname{argmax}_C P(C) \prod_i P(x_i|C)$$

Skąd wziąć prawdopodobieństwa?

- $P(C)$ - jaka część próbek w zbiorze trenującym jest z klasy C

Skąd wziąć prawdopodobieństwa?

$$\operatorname{argmax}_C P(C) \prod_i P(x_i|C)$$

Skąd wziąć prawdopodobieństwa?

- $P(C)$ - jaka część próbek w zbiorze trenującym jest z klasy C
- $P(x_i|C)$ - na podstawie rozkładu i -tego atrybutu dla klasy C
 - Jeżeli atrybut jest dyskretny - jaka część elementów z klasy C ma taką wartość?

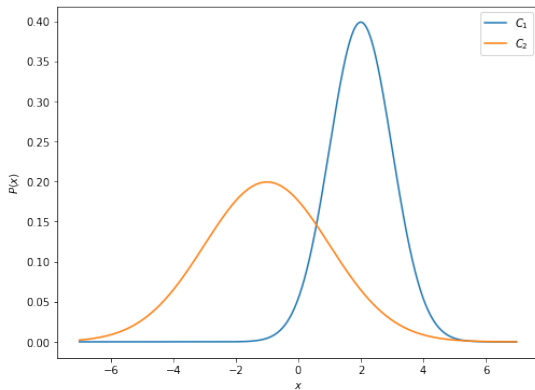
Skąd wziąć prawdopodobieństwa?

$$\operatorname{argmax}_C P(C) \prod_i P(x_i|C)$$

Skąd wziąć prawdopodobieństwa?

- $P(C)$ - jaka część próbek w zbiorze trenującym jest z klasy C
- $P(x_i|C)$ - na podstawie rozkładu i -tego atrybutu dla klasy C
 - Jeżeli atrybut jest dyskretny - jaka część elementów z klasy C ma taką wartość?
 - Jeżeli atrybut jest ciągły - trzeba założyć jakiś rozkład (zwykle zakłada się normalny)

Rozkład normalny



$x = 2$ wskazuje bardziej na C_1

$x = -4$ wskazuje bardziej na C_2

k-krotna walidacja krzyżowa

Val
Train
Train
Train

Train
Val
Train
Train

Train
Train
Val
Train

Train
Train
Train
Val

Jakość zmierzona na zbiorze walidacyjnym może zależeć od tego, jak dane zostaną podzielone.

Alternatywą jest k-krotna walidacja krzyżowa.

Zbiór danych jest dzielony na k fragmentów. Modele trenuje się na $k-1$ fragmentach a jakość się mierzy na pozostałym. Ostateczny wynik jest uśredniony z k iteracji, gdzie kolejne z k fragmentów stają się nowym zbiorem walidacyjnym.

Wynik jest miarą jakości modelu wytrenowanego na całym zbiorze.