Naiwny Bayes

1 Twierdzenie Bayesa

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

gdzie:

- $P(A|B)$ to prawdopodobieństow warunkowe zdarzenia A, jeżeli dane jest B;
- $P(B|A)$ to prawdopodobieństow warunkowe zdarzenia B, jeżeli dane jest A;
- $P(A)$ i $P(B)$ to prawdopodobieństwa *a priori* danego zdarzenia.

2 Klasyfikator naiwny Bayesowski

Klasyfikowanie polega na wybraniu klasy o największym prawdopodobieństwie zgodnie z twierdzeniem Bayesa:

$$P(c_y|x) = \frac{P(x|c_y)P(c_y)}{\sum\limits_{c_i \in C} P(x|c_i)P(c_i)}. \tag{2}$$

Obliczamy prawdopodobieństwo dla każdej klasy i wybieramy najwyższe. Sumę z mianownika możemy pominąć, ponieważ jest taka sama dla każdej klasy i nie ma wpływu na wynik porównania.

W praktyce, obliczając prawdopodobieństwo, zwykle musimy wziąć pod uwagę więcej niż jeden atrybut:

$$P(c_y|x_1, \ldots, x_d) = \frac{P(x_1, \ldots, x_d|c_y)P(c_y)}{\sum\limits_{c_i \in C} P(x_1, \ldots, x_d|c_i)P(c_i)}. \tag{3}$$

Przy założeniu **wzajemnej niezależności atrybutów** (z tego naiwnego założenia bierze się nazwa klasyfikatora), możemy zapisać:

$$P(x_1, \ldots, x_d|c_y) = \prod_{i=1}^{d} P(x_i|c_y). \tag{4}$$

Mianownik w Równaniu 3 będzie równy dla każdej z klas. Ponieważ klasyfikacja nie wymaga dokładnej znajomości prawdopodobieństwa każdej z klas, a jedynie wybrania tej z najwyższym prawdopodobieństwem, wystarczy porównanie wartości licznika.

Reasumując:

$$P(c_y|x_1, \ldots, x_d) \sim P(c_y) \prod_{i=1}^{d} P(x_i|c_y) \tag{5}$$

2.1 Wygładzanie

Może się zdarzyć, że jeden z elementów iloczynu w równaniu 4 będzie równy 0 – kiedy nie ma przykładów z daną wartością atrybutu:

$$P(x_i|c_y) = \frac{x_i}{N} = 0. \tag{6}$$

W takich wypadkach stosujemy wygładzanie:

$$P(x_i|c_y) = \frac{x_i + 1}{N + d}, \tag{7}$$

gdzie d to liczba możliwych wartości atrybutu.

Zadania

Zadanie 1.

Korzystając ze zbioru treningowego w pliku `Playgolf.xlsx`, klasyfikuj następujące przykłady klasyfikatorem naiwnym Bayesowskim:

outlook	**temp**	**humidity**	**windy**
sunny	cool	high	true
overcast	mild	normal	false
rainy	mild	normal	false

Mini-projekt (opcjonalny): Naiwny Bayes

Celem jest zaklasyfikowanie grzybów ze zbioru `agaricus-lepiota.data` (źródło) jako trujące (poisonous - klasa `p`) lub jadalne (edible - klasa `e`) przy użyciu klasyfikatora Naive Bayes.

Zaimplementuj klasyfikator i testuj na zbiorze `agaricus-lepiota.test.data`. Atrybut decyzyjny znajduje się w **pierwszej** kolumnie.

W wypadku prawdopodobieństwa równego 0, stosujemy wygładzanie.

Program powinien wypisać dokładność(accuracy), precyzję, pełność oraz F-miarę.