Naiwny Bayes

1 Twierdzenie Bayesa

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

gdzie:

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

2 Klasyfikator naiwny Bayesowski

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

$$P(c_y|x) = \frac{P(x|c_y)P(c_y)}{\sum_{c_i \in C} P(x|c_i)P(c_i)}.$$

Obliczenie prawdopodobieństwa dla danych wielowymiarowych wymaga modelowania łącznego rozkładu prawdopodobieństwa:

$$P(c_y|x_1, \ldots, x_d) = \frac{P(x_1, \ldots, x_d|c_y)P(c_y)}{\sum_{c_i \in C} P(x_1, \ldots, x_d|c_i)P(c_i)}.$$

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

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

Wybranie klasy z najwyższym prawdopodobieństwem wymaga obliczenia w równaniu tylko licznika (mianownik będzie identyczny dla każdej z klas):

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

2.1 Wygładzanie

Kiedy nie ma przykładów z daną wartością atrybutu, prawdopodobieństwo danej klasy zostałoby wyzerowane:

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

W takich wypadkach stosujemy wygładzanie:

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

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
overcast	cool	high	false

Mini-projekt: 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 należy stosować wygładzanie.

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