

Lekcja 8: Klasyfikator Bayesa

S. Hoa Nguyen

1 Materiał

a) **Problem:** Niech X oznacza przykład, którego klasa nie jest znana. Każdy przykład jest reprezentowany w postaci n -wymiarowego wektora $X = (x_1, x_2, \dots, x_n)$. Zaklasyfikować X do właściwej klasy.

b) **Algorytm klasyfikujący**

(a) dla każdej klasy C_1, C_2, \dots, C_d wyznacz prawdopodobieństwo $P(C_i|X)$

(b) klasyfikuj X do klasy C dla której wartość $P(C|X)$ jest największa.

c) **Twierdzenie Bayesa**

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

gdzie:

$P(C)$ - Prawdopodobieństwo a-priori występowania klasy C .

$P(X|C)$ - Prawdopodobieństwo występowania a-posteriori X w klasie C .

$P(X)$ - Prawdopodobieństwo występowania X

d) **Naiwny klasyfikator Bayesa**

(a) założenie: wszystkie atrybuty są **niezależne**

(b) wniosek: $P(X|C) = \prod_{i=1}^n P(x_i|C)$

(c)

$$P(x_i|C) = \frac{|\{x : a_i(x) = x_i\} \cap C|}{|C|}$$

e) **Klasyfikator Bayesa w praktyce:**

$$P(x_i|C) = \frac{|\{x : a_i(x) = x_i\} \cap C| + 1}{|C| + |A_i|}$$

f) **Klasyfikator Bayesa z kosztami pomyłek:**

- (a) dla każdej klasy C_1, C_2, \dots, C_d wyznacz $Cost(X, C_i) = \sum_{i \neq j} \rho(C_j, C_i) P(C_j|X)$
- (b) klasyfikuj X do klasy C dla której wartość $Cost(X, C)$ jest najmniejsza.

2 Zadania podstawowe

Zadanie 1. (1 pkt) Zastosując naiwny klasyfikator Bayesa na podstawie zbioru treningowego dla dziedziny modeli samochodów przedstawionej w tablicy *car_examples.ex1* klasyfikuj następujące przykłady:

x	klasa	cena	osiagi	niezawodnosc
1	<i>malý</i>	<i>umiarkowana</i>	<i>–</i>	<i>mala</i>
2	<i>miejski</i>	<i>–</i>	<i>dobre</i>	<i>duza</i>
3	<i>kompakt</i>	<i>umiarkowana</i>	<i>dobre</i>	<i>duza</i>
4	<i>duzy</i>	<i>umiarkowana</i>	<i>slabe</i>	<i>przecietna</i>

- Wyznaczyć $P(\text{akceptacja} = 0)$ i $P(\text{akceptacja} = 1)$
- Dla każdego przykładu testowego X , wyznacz $P(C|X)$
- Wyznacz klasę, dla której wartość $P(C|X)$ jest największa.

Zadanie 2. Powtórz zadanie 1 uwzględniając koszty pomyłek: $\rho(1, 0) = 0,7$ i $\rho(0, 1) = 0,3$

Zadanie 3. Porównywać jakości klasyfikacji sieci neuronowej jednokierunkowej i klasyfikatora Bayesa (zaimplementowane w systemie Weka) dla problemu klasyfikacji klientów banku. Dane o klientach są zapisane w pliku *australian.arff*.

Zadanie 4. (Klasyfikacja modeli samochodów (2 pkt)) Zaimplementować naiwny klasyfikator Bayesa i sprawdzić poprawność działania na zbiorze zawierającym opisy samochodów *cars.data*. Decyzją jest stopień akceptacji klientów (*unacc, acc, good, very good*)