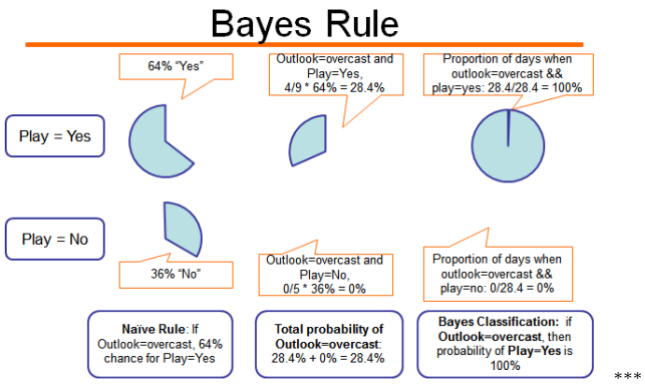
Ćwiczenie 2

Naiwny Klasyfikator Bayesa dla danych symbolicznych (Naive Bayes Classifier)



***

Zadanie do wykonania

1) Utwórz na pulpicie katalog w formacie Imię_nazwisko, w którym umieść wszystkie pliki związane z ćwiczeniem.

2) Przeczytaj teorię związaną z Naiwnym Klasyfikatorem Bayesa - w razie problemów ze zrozumieniem, przeanalizuj przykłady na kartce.

3) Wygeneruj system treningowy (australian_TRN.txt) oraz testowy (australian_TST.txt) za pomocą programu data_splitter.exe.

4) Zaimplementuj w dowolnym języku programowania (preferowany C++) klasyfikator Bayesa. Wynikiem działania programu powinny być dwa pliki:
a) dec_bayes.txt - zawierający podjęte decyzje dla obiektów systemu australian_TST.txt na podstawie obiektów systemu australian_TRN.txt
b) acc_bayes.txt - zawierający dwie wartości: Accuracy Globalne klasyfikacji i Accuracy Zbalansowane klasyfikacji.

5) W razie problemów z programowaniem, zapoznaj się z programem demonstracyjnym C++, na stronie http://wmii.uwm.edu.pl/∼artem w zakładce Dydaktyka-/Systemy Sztucznej Inteligencji. Program umieść w swoim katalogu na pulpicie.

Naiwny Klasyfikator Bayesa - część teoretyczna

Klasyfikator z rodziny metod probabilistycznych. Wymaga założenia, że atrybuty systemu decyzyjnego są niezależne od siebie, takie założenie często jest niezgodne z sytuacją rzeczywistą, stąd klasyfikator nazywamy naiwnym. Pomimo opisanego założenia klasyfikator działa w wielu przypadkach zaskakująco dobrze. Istnieje teoria,

że tego typu klasyfikacją może kierować się nasz umysł - patrz artykuł:
http://reverendbayes.wordpress.com/2008/05/29/bayesian-theory-in-new-scientist/.

Załóżmy, że mamy dane systemy decyzyjne: treningowy (U_{trn}, A, d) oraz testowy (U_{tst}, A, d), gdzie U jest zbiorem obiektów, $A = \{a_1, a_2, ..., a_n\}$ zbiorem atrybutów warunkowych, $d \in D = \{d_1, d_2, ..., d_k\}$ atrybutem decyzyjnym.

Klasyfikacja obiektu testowego $v \in U_{tst}$ opisanego jako $a_1(v), a_2(v), ..., a_n(v)$ polega na obliczeniu dla każdej z klas decyzyjnych parametru:

$$P(d = d_i | b_1 = a_1(v), b_2 = a_2(v), ..., b_n = a_n(v))$$

gdzie w postaci Twierdzenia Bayesa mamy,

$$\frac{P(b_1 = a_1(v), b_2 = a_2(v), ..., b_n = a_n(v) | d = d_i) * P(d = d_i)}{P(b_1 = a_1(v), b_2 = a_2(v), ..., b_n = a_n(v))}$$

Mianownik możemy pominąć ponieważ jest stały dla wszystkich klas decyzyjnych. Przy założeniu że atrybuty są niezależne licznik możemy obliczyć jako

$$P(b_1 = a_1(v), b_2 = a_2(v), ..., b_n = a_n(v) | d = d_i) * P(d = d_i) = P(d = d_i) * \prod_{m=1}^{n} P(b_m = a_m(v) | d = d_i)$$

W praktyce możemy zastosować oszacowanie częściowe,

$$P(b_m = a_m(v) | d = d_i) = \frac{\text{liczba wystąpien wartosci } b_m = a_m(v) \text{ w klasie } d_i}{\text{liczność klasy treningowej } d_i}$$

Każda klasa głosuje wartością parametru:

$$Param_{d=d_i} = P(d = d_i) * \prod_{m=1}^{n} P(b_m = a_m(v) | d = d_i)$$

W tym wariancie, w przypadku gdy dla klasy d_i uzyskamy wartość $P(b_m = a_m(v)|d = d_i) = 0$, wyszukujemy w pozostałych klasach najmniejszej niezerowej wartości $P(b_m = a_m(v)|d = d_j)$ i przypisujemy ją do $P(b_m = a_m(v)|d = d_i)$ po odpowiednim zmniejszeniu (np dzieląc przez 2). Jeżeli klas z zerowym wystąpieniem $b_m = a_m(v)$ jest więcej, każdej z nich przypisujemy tę zmniejszoną wartość. Innym sposobem na radzenie sobie z problemem zerowych wartości $P(b_m = a_m(v)|d = d_i)$ jest faworyzowanie pozostałych klas zawierających wartość $b_m = a_m(v)$, następuje zwiększanie liczników o jeden podczas obliczania prawdopodobieństwa. W przypadku gdy żadna z klas nie zawiera wartości $b_m = a_m(v)$, prawdopodobieństwa zerowe są pomijane w iloczynie.

Aby zapobiec problemowi małych bliskich zeru liczb, możemy poszczególne prawdopodobieństwa logarytmować. W praktycznym zastosowaniu, dopuszcza się użycie sumy prawdopodobieństw, (ponieważ mnożenie przez małe wartości może doprowadzić do zmniejszenia parametru klasy do zera, przy ograniczonej dokładności) w naszym wariancie algorytmu każda klasa decyzyjna będzie głosowała za pomocą parametru:

$$Param_{d=d_i} = P(d = d_i) * \sum_{m=1}^{n} P(b_m = a_m(v)|d = d_i)$$

W przypadku problemu z zerową licznością deskryptora $b_m = a_m(v)$, stosujemy faworyzowanie klas, podobnie jak w wariancie z iloczynem prawdopodobieństw.

Gdy podczas klasyfikacji parametry klas są jednakowe, konflikt rozwiązujemy nadając obiektowi testowemu losową decyzję.

W przypadku gdy atrybuty są ciągłe, oraz zakładając że mają rozkład normalny, poszczególne prawdopodobieństwa $P(b_m = a_m(v)|d = d_i)$ szacujemy za pomocą funkcji Gaussa.

$$f(x) = \frac{1}{\sqrt{(2 * \pi * \sigma_c^2)}} * e^{\frac{-(x-\mu_c)^2}{2*\sigma_c^2}}$$

Do jej obliczenia potrzebujemy średnie z klas oraz wariancje wartości w klasach.

$$\mu_c = \frac{\sum_{i=1}^{licznosc\ klasy\ c} a(v_i)}{licznosc\ klasy\ c}$$

$$\sigma_c^2 = \frac{1}{licznosc\ klasy\ c} * \sum_{i=1}^{licznosc\ klasy\ c} (a(v_i) - \mu_c)^2$$

Przykład działania Naiwnego Klasyfikatora Bayesa

Wczytujemy system testowy (problemy do rozwiązania z ukrytymi decyzjami eksperta) postaci,

Tabela 1: System Testowy (X, A, c)

	a_1	a_2	a_3	a_4	c
x_1	2	4	2	1	4
x_2	1	2	1	1	2
x_3	9	7	10	7	4
x_4	4	4	10	10	2

oraz system treningowy (bazę wiedzy służącą do rozwiązywania problemów)
Widzimy, że $P(c = 2) = \frac{3}{6} = \frac{1}{2}$, $P(c = 4) = \frac{1}{2}$.

Rozpoczynamy od klasyfikacji obiektu testowego
x_1 2 4 2 1 4

Wyliczamy $Param_{c=2} = P(c = 2) * \sum_{i=1}^{4} P(a_i = v_i|c = 2)$,
$P(a_1 = 2|c = 2) = \frac{1}{3}$
$P(a_2 = 4|c = 2) = \frac{0}{3}$ nie da się obsłużyć tej wartości, żadna z klas nie zawiera $a_2 = 4$
$P(a_3 = 2|c = 2) = \frac{1}{3}$
$P(a_4 = 1|c = 2) = \frac{3}{3}$

Tabela 2: System Treningowy (Y, A, c)

	a_1	a_2	a_3	a_4	c
y_1	1	3	1	1	2
y_2	10	3	2	1	2
y_3	2	3	1	1	2
y_4	10	9	7	1	4
y_5	3	5	2	2	4
y_6	2	3	1	1	4

Ostatecznie $Param_{c=2} = \frac{1}{2} * (\frac{1}{3} + \frac{0}{3} + \frac{1}{3} + \frac{3}{3}) = \frac{5}{6}$;

oraz $Param_{c=4} = P(c = 4) * \sum_{i=1}^{4} P(a_i = v_i|c = 4)$,
$P(a_1 = 2|c = 4) = \frac{1}{3}$
$P(a_2 = 4|c = 4) = \frac{0}{3}$ nie da się obsłużyć tej wartości, żadna z klas nie zawiera $a_2 = 4$
$P(a_3 = 2|c = 4) = \frac{1}{3}$
$P(a_4 = 1|c = 4) = \frac{2}{3}$
Ostatecznie $Param_{c=4} = \frac{1}{2} * (\frac{1}{3} + \frac{0}{3} + \frac{1}{3} + \frac{2}{3}) = \frac{2}{3}$;

$Param_{c=2} > Param_{c=4}$, obiekt x_1 dostaje decyzję 2, ta decyzja nie jest zgodna z ukrytą decyzją eksperta stąd obiekt jest błędnie sklasyfikowany,

Podczas klasyfikacji pierwszego obiektu napotykamy liczność zero dla $a2 = 4$, której nie możemy obsłużyć, ponieważ żadna z istniejących klas nie zawiera wartości $a2 = 4$.

Przejdźmy do klasyfikacji obiektu testowego
x_2 1 2 1 1 4

Wyliczamy $Param_{c=2} = P(c = 2) * \sum_{i=1}^{4} P(a_i = v_i|c = 2)$,
$P(a_1 = 1|c = 2) = \frac{1}{3}$ zwiększamy licznik o 1, ponieważ $P(a_1|c = 4) = 0$, $P(a_1 = 1|c = 2) = \frac{2}{3}$
$P(a_2 = 2|c = 2) = \frac{0}{3}$ nie da się obsłużyć tej wartości, żadna z klas nie zawiera $a_2 = 2$
$P(a_3 = 1|c = 2) = \frac{1}{3}$
$P(a_4 = 1|c = 2) = \frac{3}{3}$
Ostatecznie $Param_{c=2} = \frac{1}{2} * (\frac{2}{3} + \frac{0}{3} + \frac{1}{3} + \frac{3}{3}) = 1$;

oraz $Param_{c=4} = P(c = 4) * \sum_{i=1}^{4} P(a_i = v_i|c = 4)$,
$P(a_1 = 1|c = 4) = \frac{0}{3}$ w tej sytuacji zwiększam o 1 licznik $P(a_1 = 1|c = 2)$, faworyzując klasy zawierające przynajmniej jedną wartość $a_1 = 1$
$P(a_2 = 2|c = 4) = \frac{0}{3}$ nie da się obsłużyć tej wartości, żadna z klas nie zawiera $a_2 = 2$
$P(a_3 = 1|c = 4) = \frac{1}{3}$
$P(a_4 = 1|c = 4) = \frac{2}{3}$
Ostatecznie $Param_{c=2} = \frac{1}{2} * (\frac{0}{3} + \frac{0}{3} + \frac{1}{3} + \frac{2}{3}) = \frac{1}{2}$;

$Param_{c=2} > Param_{c=4}$, obiekt x_2 dostaje decyzję 2, ta decyzja jest zgodna z ukrytą decyzją eksperta stąd obiekt jest poprawnie sklasyfikowany,

Przejdźmy do klasyfikacji obiektu testowego
x_3 9 7 10 7 4

Wyliczamy $Param_{c=2} = P(c=2) * \sum_{i=1}^{4} P(a_i = v_i | c=2)$,
$P(a_1 = 9|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_2 = 7|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_3 = 10|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_4 = 7|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
Ostatecznie $Param_{c=2} = \frac{1}{2} * (\frac{0}{3} + \frac{0}{3} + \frac{0}{3} + \frac{0}{3}) = 0$;

oraz $Param_{c=4} = P(c=4) * \sum_{i=1}^{4} P(a_i = v_i | c=4)$,
$P(a_1 = 9|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_2 = 7|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_3 = 10|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_4 = 7|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
Ostatecznie $Param_{c=2} = \frac{1}{2} * (\frac{0}{3} + \frac{0}{3} + \frac{0}{3} + \frac{0}{3}) = 0$;

$Param_{c=2} == Param_{c=4}$, obiekt x_3 dostaje decyzję losową $los(2,4) = 4$, ta decyzja jest zgodna z ukrytą decyzją eksperta stąd obiekt jest poprawnie sklasyfikowany,

Przejdźmy do klasyfikacji obiektu testowego
x_4 4 4 10 10 4

Wyliczamy $Param_{c=2} = P(c=2) * \sum_{i=1}^{4} P(a_i = v_i | c=2)$,
$P(a_1 = 4|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_2 = 4|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_3 = 10|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_4 = 10|c=2) = \frac{0}{3}$ nie da się obsłużyć tej wartości
Ostatecznie $Param_{c=2} = \frac{1}{2} * (\frac{0}{3} + \frac{0}{3} + \frac{0}{3} + \frac{0}{3}) = 0$;

oraz $Param_{c=4} = P(c=4) * \sum_{i=1}^{4} P(a_i = v_i | c=4)$,
$P(a_1 = 4|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_2 = 4|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_3 = 10|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
$P(a_4 = 10|c=4) = \frac{0}{3}$ nie da się obsłużyć tej wartości
Ostatecznie $Param_{c=2} = \frac{1}{2} * (\frac{0}{3} + \frac{0}{3} + \frac{0}{3} + \frac{0}{3}) = 0$;

$Param_{c=2} == Param_{c=4}$, obiekt x_4 dostaje decyzję losową $los(2,4) = 4$, ta decyzja nie jest zgodna z ukrytą decyzją eksperta stąd obiekt jest poprawnie sklasyfikowany,

Ostatecznie wyliczamy parametry:

$$Gobal\_Accuracy = \frac{\text{liczba obiektów tst poprawnie sklasyfikowanych w całym systemie}}{\text{liczba obiektów sklasyfikowanych w całym systemie}}$$

$$Balanced\_Accuracy = \frac{\sum_{i=1}^{liczba\ klas} \frac{\text{liczba obiektów tst poprawnie sklasyfikowanych w klasie } c_i}{\text{liczba obiektów sklasyfikowanych w klasie } c_i}}{liczba\ klas}$$

$$Gobal\_Accuracy = \frac{2}{4} = \frac{1}{2}$$

$$Balanced\_Accuracy = \frac{\frac{1}{2} + \frac{1}{2}}{2} = \frac{1}{2}$$

Obiekt tst	Ukryta decyzja eksperta	Decyzja naszego kasyfikatora
x_1	4	2
x_2	2	2
x_3	4	4
x_4	2	4

***

the source of image from the first page: http://www.simafore.com/