Wyszukiwanie informacji w sieci, Lista 5 **21.05.2007**

1. [1] (11) Przy rangowaniu dokumentów dwumianową metodą probabilistyczną (rozdz. 11) dokument x i zapytanie q są reprezentowane przez wektory o długości równej rozmiarowi słownika, gdzie $x_i = 1/q_i = 1$, gdy i-ty term ze słownika występuje w dokumencie/zapytaniu. Ranga dokumentu x określana jest na podstawie wartości

$$RSV = \sum_{i:x_i=q_i=1} \log \frac{p_i(1-u_i)}{u_i(1-p_i)}$$

gdzie $p_i = P(x_i = 1|R, q)$ i $u_i = P(x_i = 1|NR, q)$. (Tutaj R/NR oznacza dokumenty istotne/nieistotne; czyli p_i określa częstość i-tego termu wśród dokumentów istotnych dla zapytania, a u_i wśród nieistotnych dla zapytania.)

Gdybyśmy mieli wiedzę o tym, które dokumenty są istotne, czyli znalibyśmy tabelkę

	dok. istotne	dok. nieistotne	razem
$x_i = 1$	s	$n - s$	n
$x_i = 0$	$S - s$	$(N - n) - (S - s)$	$N - n$
Razem	S	$N - S$	N

to wartości p_i i u_i moglibyśmy aproksymować jako: $p_i = s/S$, $u_i = (n - s)/(N - S)$.

Zadanie

(a) Porównaj wpływ termów o następujących własnościach na wartość rangi : termy występujące często w całej kolekcji dokumentów; termy występujące często tylko wśród dokumentów istotnych; termy występujące często tylko wśród dokumentów nieistotnych.

(b) Używając powyższych oznaczeń i przyjmując, że dokumenty istotne/nieistotne odpowiadają dokumentom z kategorii c i poza kategorią c, sformułuj klasyfikację dokumentów metodą Bayesa, w modelu dwu-mianowym (p. rodział 13, tab. 13.2).

(c) Porównujemy omówioną w punkcie (a) metodę rangowania ze sformułowaną w podpunkcie (b) metodą klasyfikacji. Potraktujmy obie metody jako metody klasyfikacji dokumentów. Odpowiedz na pytania z podpunktu (a) dla klasyfikatora Bayesa z punktu (b). Na tej podstawie scharakteryzuj dokumenty, dla których ocena obu algorytmów będzie odmienna.

2. [0.5] (12) Rozważmy zapytanie $q = t_1t_2t_3t_4$ oraz 10 dokumentów $d_1, \ldots, d_{10}$ z informacją, które z nich są istotne dla zapytania (oraz informacje o liczbie wystąpień poszczególnych termów w dokumentach):

	t_1	t_2	t_3	t_4	istotny
d_1	2	1	0	0	0
d_2	1	0	0	0	0
d_3	1	0	1	1	1
d_4	1	2	1	1	1
d_5	0	2	0	1	1
d_6	0	1	1	0	0
d_7	1	1	0	0	0
d_8	0	1	2	1	1
d_9	1	1	0	0	0
d_{10}	1	0	1	0	0

Dla każdego z dokumentów, wyznacz wartość jego rankingu według metody probabilistycznej z rozdziału 12 (pamiętaj o „wygładzeniu" prawdopodobieństw zerowych; podaj jaką metodę wygładzania stosujesz).

3. [1] (13) Jedna z modyfikacji „naiwnego klasyfikatora Bayesa" polega na dodaniu nowego parametru t. Wykorzystujemy go w ten sposób, że formuła wyboru kategorii dla dokumentu $x = x_1 x_2 \ldots x_n$,

$$c = \text{argmax}_{c_j \in C} P(c_j) \prod_{i=1}^{n} P(x_i|c_j),$$

gdzie C to zbiór kategorii, przybiera postać

$$c = \text{argmax}_{c_j \in C} P(c_j)^t \prod_{i=1}^{n} P(x_i|c_j).$$

A zatem zmieniamy wpływ czynnika $P(c_j)$ na wynik przy pomocy parametru t.

   (a) Spróbuj uzasadnić czy i dlaczego taka modyfikacja zazwyczaj poprawia wyniki działania klasyfikatora.

   (b) Załóżmy, że działanie pewnego klasyfikatora Bayesa na zbiorze testowym opisuje następująca tabela

	a	b	c	d
a	50	26	20	12
b	3	8	2	0
c	0	0	4	0
d	0	1	0	1

   gdzie a, b, c i d to różne kategorie.

   Wartość na przecięciu wiersza i i kolumny j w powyższej tabeli oznacza liczbę dokumentów z kategorii i, które zostały zakwalifikowane jako dokumenty z kategorii j.

   Jakich zmian w klasyfikacji dokumentów należy się spodziewać dla różnych wartości t w zmodyfikowanej metodzie Bayesa?

4. [0.5] (13.3) W procesie treningu naiwnego klasyfikatora Bayes'a (rys. 13.3 w podręczniku), prawdopodobieństwo kategorii $P(c_j)$ przybliża się poprzez frakcję dokumentów w klasie

c_j w całej kolekcji dokumentów. Alternatywną miarą mogłaby być proporcja sumarycznego rozmiaru (liczba termów) dokumentów z klasy c_j do sumarycznego rozmiaru całej kolekcji dokumentów treningowych.

Uzasadnij dlaczego nie wybiera się tej drugiej miary.

5. [1] (13.12) We wzorach wyznaczających $X^2(W, C)$ dla termu W i kategorii C, przyjęliśmy oznaczenia:

   - $O_{1,1}$ to zaobserwowana (w danych treningowych) liczba dokumentów z kategorii C, w których występuje term W;
   - $O_{1,0}$ to liczba dokumentów z kategorii C, w których nie występuje term W;
   - $O_{0,1}, O_{0,0}$ – analogicznie (dla dokumentów nie należących do kategorii C);
   - $P(W) = (O_{1,1} + O_{0,1})/N$ to zaobserwowane prawdopodobieństwo występowania termu W w danych treningowych;
   - $P(C) = (O_{1,1} + O_{1,0})/N$ to prawdopodobieństwo występowania kategorii C;
   - $E_{i,j} = N \cdot P_i(C) \cdot P_j(W)$ to oczekiwana liczba dokumentów z kategorii C (gdy $i = 1$) lub poza C (gdy $i = 0$) zawierających term W (dla $j = 1$) lub nie zawierających W (dla $j = 0$); przez „oczekiwaną liczbę" rozumiemy tutaj liczbę, którą uzyskalibyśmy gdyby przynależność do C i występowanie W były niezależne.

   Powyżej przyjmujemy, że N to liczba dokumentów w zestawie treningowym, $P_1(X) = P(X)$ oraz $P_0(X) = 1 - P(X)$.

   Polecenie. Uzasadnij, że przy powyższych oznaczeniach zachodzi

   $$|O_{1,1} - E_{1,1}| = |O_{1,0} - E_{1,0}| = |O_{0,1} - E_{0,1}| = |O_{0,0} - E_{0,0}|.$$

6. [0.5] (13.7) Jakie wartości osiągają współczynniki $I(W, C)$ i $X^2(W, C)$ dla termu W całkowicie niezależnego od kategorii C (W występuje z tym samym prawdopodobieństwem w C jak i poza C) oraz termu W całkowicie zależnego od kategorii C (W występuje z prawdopobieństwem 1 w C oraz z prawdopodobieństwem 0 poza C; lub odwrotnie).

7. [0.5] (13.13) Miary chi-kwadrat ($X^2(W, C)$) i mutual information ($I(W, C)$) nie rozróżniają korelacji pozytywnej i negatywnej. W praktyce zdecydowanie bardziej przydatne są cechy o korelacji pozytywnej. Zaproponuj sposób wyeliminowania cech o korelacji negatywnej.

8. [1] (13.14) Zadanie polega na klasyfikacji słów, klasyfikator ma odróżniać słowa angielskie od nie-angielskich. (A zatem odpowiednikami dokumentów są tutaj słowa, a termów litery.) W danych losowych pojawiają się następujące słowa z poniższym rozkładem prawdopodobieństwa:

słowo	angielskie?	prawdopodobieństwo
ozb	0	4/9
uzu	0	4/9
zoo	1	1/18
bun	1	1/18

(a) Oblicz prawdopodobieństwa $P(c)$ i $P(w|c)$ dla kategorii c angielskie/inne i liter w ze zbioru b, n, o, u, z. (Tak jak to się liczy w naiwnym klasyfikatorze Bayes'a.) Użyj „wygładzania" polegającego na tym, że zerowe prawdopodobieństwa symboli są przybliżane jako 0.01. (W efekcie uzyskujemy $P(A) + P(\neg A) > 1$, ale nie martwimy się tym.)

(b) Jak sklasyfikowane zostanie słowo *zoo*?

(c) Chcemy sklasyfikować słowo *zoo* przyjmując, że rozkłady odpowiadające różnym pozycjom litery w słowie są różne. Policz w tym celu potrzebne wartości $P(W, i|c)$ oznaczające prawdopodobieństwo litery W na pozycji i wśród słów kategorii c. Wyznacz kategorię słowa *zoo*.

Uwaga: prawdopodobieństwo $a/18$ możesz traktować jak a wystąpień danego słowa w zbiorze treningowym.

9. [2] (14) Zweryfikuj prawdziwość następujących zdań:

(a) Liczba liniowych separatorów między dwoma klasami jest nieskończona lub równa zero.

(b) Centroid p znormalizowanych wektorów n-wymiarowych jest również znormalizowany, czyli $\| y \| = \sum_{i=1}^{n} y_i^2 = 1$ dla

$$y = (\sum_{x \in X} x)/|X|$$

gdzie $\sum_{j=1}^{n} x_{ij}^2 = 1$ dla każdego $i \in [1, |X|]$.

(c) Jeśli zbiór elementów pewnej kategorii jest liniowo separowalny, to klasyfikator Bayesa znajdzie dla niego liniowy separator.

(d) Jeśli zbiór elementów pewnej kategorii jest liniowo separowalny, to klasyfikator Rocchio znajdzie dla niego liniowy separator.

10. [1] Używając poznanych metod klasyfikacji, zaproponuj sposób klasyfikacji i rangowania dokumentów (ze względu na zapytanie), które pozwolą wyszukiwać dokumenty podobne do zapytania, w tym również dokumenty w innych językach niż język zapytania. Proces „uczenia" wyszukiwarki może być kosztowny, ale realizacja zapytań powinna być możliwa w sensownym czasie.