Πιθανοκρατικό μοντέλο

Το μοντέλο MAP

Αλέξανδρος Γκιμπερίτης
Βασίλης Μπούργος
Δημήτρης Σουραβλιάς

Εισαγωγικές έννοιες

- Κάθε έγγραφο d της συλλογής παριστάνεται από το δυαδικό διάνυσμα $x = (x_1, x_2, \ldots, x_M)$
- $x_t = 1$, αν ο όρος t υπάρχει στο έγγραφο
  $x_t = 0$, αν ο όρος t δεν υπάρχει στο έγγραφο
- Με τον ίδιο τρόπο παριστάνονται και τα ερωτήματα
- Ιδανικό σύνολο R για ένα ερώτημα q είναι το σύνολο των συναφών εγγράφων που μπορούμε να έχουμε ως απάντηση

Ορισμοί

Για τους όρους που εμφανίζονται στο ερώτημα
q ορίζουμε:

- p_t: την πιθανότητα ένα έγγραφο που επιλέγεται από το ιδανικό σύνολο να περιέχει τον όρο x_t.
- u_t: την πιθανότητα ένα έγγραφο που επιλέγεται από το ΜΗ ιδανικό σύνολο να περιέχει τον όρο x_t.
- Στόχος: να εκτιμήσουμε θεωρητικά τις πιθανότητες p_t και u_t για συγκεκριμένη συλλογή και ερώτημα.

Συνάρτηση ομοιότητας

- Στόχος: Εύρεση συνάρτησης για τη διαβάθμιση των εγγράφων που είναι συναφή με το ερώτημα

  Retrieval status value

$$RSV_d = \log \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=q_t=1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

- Ιδέα: Χρήση μονότονης συνάρτησης, όπως ο δυαδικός λογάριθμος, για καλύτερη διαβάθμιση

Συνάρτηση ομοιότητας (συνέχεια)

- Ορίζουμε την ποσότητα c_t που δίνεται από τον παρακάτω τύπο:

$$c_t = \log\frac{p_t(1-u_t)}{u_t(1-p_t)} = \log\frac{p_t}{1-p_t} + \log\frac{1-u_t}{u_t}$$

- Αρκεί λοιπόν να υπολογίσουμε το c_t για τους κοινούς όρους εγγράφου και ερωτήματος

Πίνακας ομοιότητας

documents		relevant	non-relevant	total
term present	$x_t = 1$	s	df_t - s	df_t
term absent	$x_t = 0$	S – s	(N – df_t) – (S – s)	N - df_t
	total	S	N – S	N

όπου:

N: το πλήθος των εγγράφων της συλλογής.

df_t: το πλήθος των εγγράφων που περιέχουν τον όρο t.

S: το πλήθος των επιστρεφόμενων εγγράφων.

s: το πλήθος των επιστρεφόμενων εγγράφων που περιέχουν τον όρο t.

Υπολογισμός c_t

- Με βάση τον προηγούμενο πίνακα προκύπτει $p_t = s/S$ και $u_t = (df_t - s)/(N-S)$ άρα η σχέση που δίνει το c_t γίνεται

$$c_t = \log \frac{s/(S-s)}{(df_t - s)/((N - df_t) - (S - s))}$$

- Για να αποφύγουμε πιθανή διαίρεση με το μηδέν, προσθέτουμε ½ στις ποσότητες του αριθμητή και του παρονομαστή
- Τελικά έχουμε την παρακάτω σχέση:

$$c_t = \log \frac{(s+\frac{1}{2})/(S-s+\frac{1}{2})}{(df_t - s+\frac{1}{2})/(N - df_t - S + s+\frac{1}{2})}$$

MLE

- Όταν έχουμε πειράματα με έξοδο κατηγορίες τότε η πιθανότητα ενός γεγονότος ισούται με τη συχνότητα των εμφανίσεων προς το σύνολο των γεγονότων (Σχετική Συχνότητα Γεγονότος)
- Όταν χρησιμοποιούμε Σ.Σ.Γ. έχουμε MLE (Maximally Likelihood Estimate)
- Η τιμή αυτή κάνει τα παρατηρηθέντα γεγονότα πιο πιθανά με μέγιστη τιμή (maximally)
- Όταν έχουμε MLE τότε κάποια γεγονότα που εμφανίζονται συχνά έχουν μεγάλη τιμή ενώ άλλα που δεν εμφανίζονται καθόλου παίρνουν τιμή 0

Smoothing

- Το smoothing είναι μια τεχνική η οποία αυξάνει την πιθανότητα των ενδεχομένων που δεν έχουν παρατηρηθεί και μειώνει την πιθανότητα όσων έχουν παρατηρηθεί
- Ένας απλός τρόπος είναι να προσθέσουμε έναν αριθμό α σε κάθε άθροισμα που έχουμε παρατηρήσει
- Αυτό το εφαρμόζουμε επειδή ενδεχόμενα με μεγάλο πλήθος παρατηρήσεων παίρνουν υπερβολικά μεγάλες τιμές (MLE)

MAP

- Αρχικά υποθέτουμε ότι έχουμε ομοιόμορφη κατανομή και θέτουμε α=1/2 και στην συνέχεια ενημερώνουμε την τιμή αυτή με βάση αυτά που έχουμε παρατηρήσει
- Η τεχνική αυτή ονομάζεται MAP (maximum a posteriori estimation) επειδή λαμβάνει υπ' όψιν της για τον υπολογισμό των πιθανοτήτων τα αρχικά αποτελέσματα και αυτά που προκύπτουν από τα παρατηρηθέντα γεγονότα

Παραδείγματα εφαρμογής

- Έστω τα παρακάτω έγγραφα και επερωτήσεις:
  - d1: a, b
  - d2: a, b, a, b
  - d3: a, b, a, b, c
  - d4: a, b, c
  - d5: a, a, c

  q1: a, b
  q2: a
  q3: c
  q4: a, c

- Ακολουθούν δύο παραδείγματα εφαρμογής με διαφορετικά αποτελέσματα στην επιστροφή εγγράφων

- Αρχικά υποθέτουμε ότι επιστρέφονται μόνο τα έγγραφα που περιέχουν όλους τους όρους του ερωτήματος

Παραδείγματα εφαρμογής(συνέχεια)

- q1: a,b
- S= {d1,d2,d3,d4}
- Άρα
  - N = 5, $df_a = 5$, $df_b = 4$
  - S=4, $s_a = 4$, $s_b = 4$

Εφαρμόζοντας στον τύπο έχουμε:

$$c_a = \log \frac{\frac{4+1/2}{4-4+1/2}}{\frac{5-4+1/2}{5-5-4+4+1/2}} = \log \frac{9/2}{3/2} = \log 3$$

$$c_b = \log \frac{\frac{4+1/2}{4-4+1/2}}{\frac{4-4+1/2}{5-4-4+4+1/2}} = \log \frac{9}{1/3} = \log 27 = 3\log 3$$

Ομοίως για τα επόμενα ερωτήματα:

Ερώτημα	S
a	d1, d2, d3, d4, d5
c	d3, d4, d5
a ,c	d3, d4, d5

	S	s_a	s_c	c_a	c_c
q2	5	5	-	log11	-
q3	3	-	3	-	log35
q4	3	3	3	log(7/5)	log35

Επιπλέον

- N=5
- df_a=5
- df_c=3

- Τώρα υποθέτουμε ότι επιστρέφονται και σχετικά και μη σχετικά έγγραφα

Κάνοντας πράξεις έχουμε:

Ερώτημα	S
a ,b	d1, d2, d3
a	d2, d3, d4, d5
c	d1, d3, d5
a ,c	d2, d3, d5

	S	S_a	S_b	S_c	C_a	C_b	C_c
q1	3	3	3	-	log(7/5)	log(7/5)	-
q2	4	2	-	-	log3	-	-
q3	3	-	-	2	-	-	log(10/6)
q4	3	3	-	2	log(7/5)	-	log(10/6)

Επιπλέον

- N=5
- df_a=5, df_b=4, df_c=3

Ευχαριστούμε για την προσοχή σας

ΕΡΩΤΗΣΕΙΣ?