lunedì 30 novembre 2020 11:30

$$C = \{c_1, c_2, \dots, c_n\}$$

Th. BAYES

$$P(c_i \mid d) = \frac{P(d \mid c_i) P(c_i)}{P(d)}$$

Per assegnare la classe al doc. d devo calcolare

$$\arg\max_{c_i} P(c_i \mid d) =$$

$$\arg\max_{c_i} \frac{P(d \mid c_i) P(c_i)}{P(d)}$$ — poiché $P(d)$ è costante

$$\arg\max_{c_i} P(d \mid c_i) P(c_i)$$

$$P(c_i) = \frac{\#\text{ doc. di classe } c_i \text{ nel training set}}{\#\text{ totale doc nel training set}}$$

$$P(d \mid c_i) = P(t_1 \wedge t_2 \wedge \dots \wedge t_k \mid c_i) =$$ assumo l'indipendenza dei termini

$$P(d \mid c_i) = P(t_1 \mid c_i) P(t_2 \mid c_i) \dots P(t_k \mid c_i) = \prod_{j=1}^{k} P(t_j \mid c_i)$$

$$P(t_k \mid c_i) = \frac{\#\text{ volte in cui } t_k \text{ compare nei doc di classe } c_i \text{ nel training set}}{\#\text{ tot di termini presenti in tutti i doc della classe } c_i}$$

$$P(t_k \mid c_i) = \frac{\#\text{ volte in cui } t_k \text{ è nei doc di classe } c_i \text{ nel Tr. set} + 1}{\#\text{ tot termini nei doc di classe } c_i + |V|}$$

CORREZIONE DI LAPLACE

1) Siano dati l'insieme delle categorie C = {c1, c2} e una collezione di 1000 documenti definiti sul vocabolario V = {T1,T2,T3,T4,T5}.

a) Costruire un classificatore *bayesiano* per C, addestrandolo sul seguente training set TR:

TR = {<D1,c1>, <D2,c1>, <D3,c2>, <D4,c2>}

dove per ogni documento si riporta di seguito l'elenco delle parole in esso presenti, con le relative occorrenze:

D1={T1:1, T2:2, T3:3} D2={T4:1}

D3={T1:2, T2:5} D4={T3:4, T4:2}

NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 6)

b) Determinare la classe di appartenenza del seguente documento d={T2:2,T5:2}

(PUNTI 2)

Per il teorema di Bayes

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)}$$

Per classif d devo calcolare $\arg\max_{c_i} P(c_i|d) \Rightarrow$

$$\arg\max_{c_i} \frac{P(d|c_i)P(c_i)}{P(d)}$$

Poiché $P(d)$ è costante, posso tranquillo nel calcolo del max e quindi mi basta calcolare $\arg\max_{c_i} P(d|c_i)P(c_i)$

$$P(c_i) = \frac{\#\text{ doc di classe } c_i \text{ nel training set}}{|\text{training set}|}$$

$$P(c_1) = \frac{2}{4} \qquad P(c_2) = \frac{2}{4}$$

$$P(d|c_i) = P(t_1 \wedge t_2 \wedge \dots \wedge t_k | c_i) \quad \text{assunz. indip.} \quad = P(t_1|c_i)P(t_2|c_i)\dots P(t_k|c_i)$$

$$= \prod_{j=1}^{k} P(t_j|c_i)$$

$$P(T_1|c_1) = \frac{1+1}{7+5} = 2/12$$

$$P(T_2|c_1) = (2+1)/12 = 3/12$$

$$P(T_1|c_2) = (2+1)/(13+5) = 3/18$$

$$P(T_2|c_2) = (5+1)/18 = 6/18$$

$P(T_2 | c_1) = (2+1)/12 = 3/12$

$P(T_3 | c_1) = (3+1)/12 = 4/12$

$P(T_4 | c_1) = (1+1)/12 = 2/12$

$P(T_5 | c_1) = (0+1)/12 = 1/12$

$P(T_2 | c_2) = (5+1)/18 = 6/18$

$P(T_3 | c_2) = (4+1)/18 = 5/18$

$P(T_4 | c_2) = (2+1)/18 = 3/18$

$P(T_5 | c_2) = (0+1)/18 = 1/18$

$$P(T_k | c_i) = \frac{\#\text{volte in cui } T_k \text{ compare nei doc di classe } c_i + 1}{\#\text{tot termini nei doc di classe } c_i + |V|}$$

Classifico $d = \langle T_2 : 2, T_5 : 2 \rangle = \langle T_2, T_2, T_5, T_5 \rangle$

$$P(c_1 | d) = \frac{P(c_1) P(d | c_1)}{P(d)} = P(c_1) P(T_2 | c_1) P(T_2 | c_1) P(T_5 | c_1)^2 = \frac{2}{4} \frac{3}{12} \frac{3}{12} \frac{1}{12} \frac{1}{12} = \alpha$$

$$P(c_2 | d) = P(c_2) P(T_2 | c_2) P(T_2 | c_2) P(T_5 | c_2) P(T_5 | c_2) = \frac{2}{4} \frac{6}{18} \frac{6}{18} \frac{1}{18} \frac{1}{18} = \beta$$

$\alpha > \beta \Rightarrow d \in C_1$

$\alpha < \beta \Rightarrow d \in C_2$

1) Siano dati l'insieme delle categorie C={c1, c2} e una collezione di 150 documenti definiti sul vocabolario V = {T1, T2, T3, T4, T5}. Costruire un classificatore *k-NN (k=3)* per C, addestrandolo sul seguente training set TR = {<D1, c1>, <D2, c2>, <D3, c1>, <D4, c2>} dove per ogni documento si riporta di seguito l'elenco delle parole con le relative occorrenze:

	T1	T2	T3	T4	T5
D1	3	3	0	4	0
D2	1	0	2	0	1
D3	0	1	0	2	1
D4	0	2	0	0	4

D 2 0 0 0 2

Determinare inoltre la classe di appartenenza del seguente documento d={T1:2, T5:2}

Nota Bene: rappresentare i documenti utilizzando le occorrenze dei termini e utilizzare la *similarità del prodotto interno*. (PUNTI 6)

Il classificatore k-NN è il TRAINING SET

Per classificare d devo calcolare

$SIM(\vec{d}, \vec{D}_{x=1,2,3,4})$ e prendere la categoria di MAGGIORANZA dei primi 3 risultati più SIMILI

$SIM(\vec{d}, \vec{D_1}) = 2\times3 + 0\times3 + 0\times0 + 0\times4 + 2\times0 = 6$

$SIM(\vec{d}, \vec{D_2}) = 2\times1 + 2\times1 = 4$

$SIM(\vec{d}, \vec{D_3}) = 2\times1 = 2$

$SIM(\vec{d}, \vec{D_4}) = 2\times4 = 8$

RANKING

D4 → c_2

D1 → c_1

D2 → c_2

D3

2 su 3 sono di cat. c_2

$d \in c_2$