Chen Zhang czhang49

CS 598 CXZ Fall 2016 HW1

① a) Because different words have different weights to the meaning of query. e.g. in a search for "news about presidential campaign", about should be given less weights compared with "presidential".

b) $IDF(w) = \log\left(\frac{M+1}{K}\right)$, K being number of docs containing w, M being total number of docs.

If word w is contained in d, $M' = M+1$, $K' = K+1$, since $M > K \Rightarrow IDF(w)$ will decrease.

if word w is not contained in d, $M' = M+1$, $K' = K$, $\Rightarrow$ IDF(w) will increase

c) i) $\text{precision} = \frac{\text{relavant doc \#}}{\text{retrived total doc \#}} = \frac{5}{10} = \frac{1}{2}$

ii) $\text{Recall} = \frac{\text{relavant doc \#}}{\text{total relevant \#}} = \frac{5}{16}$

iii) $F_1 = \frac{2PR}{P+R} = \frac{2 \times \frac{1}{2} \times \frac{5}{16}}{\frac{5}{16} + \frac{1}{2}} = \frac{\frac{5}{16}}{\frac{13}{16}} = \frac{5}{13}$

iv) $\text{Average Precision} = \frac{\frac{1}{1} + \frac{2}{2} + \frac{3}{4} + \frac{4}{5} + \frac{5}{8}}{10} = \frac{\frac{167}{40}}{10} = \frac{167}{400}$

d) i) Cumulative gain = $1+1+0+1+1+0+0 = 4$

ii)

Discounted Cumulative gain ($\log = \log_2$)

$$= 1 + \frac{1}{\log 2} + \frac{1}{\log 4} + \frac{1}{\log 5} = 1+1+\frac{1}{2}+\frac{1}{2.3} = 2.93$$

Normalized Discounted Cumulative gain

$$= \frac{DCG@7}{\text{ideal } DCG@7} = \frac{2.93}{1+\frac{1}{\log 2}+\frac{1}{\log 3}+\cdots\frac{1}{\log 7}}$$

$$= \frac{2.93}{4.3} = 0.681$$

② a)

V	$P(A=1 \mid V)$	$P(K=1 \mid V)$	$P(L=1 \mid V)$	prior $P(V)$
0	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{2}$
1	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{2}{6}$	$\frac{1}{2}$

b) $P(V=1 \mid A=0, K=1, L=0)$

$$= \frac{P(V=1, A=0, K=1, L=0)}{P(A=0, K=1, L=0)}$$

$$= \frac{P(V=1)\,P(A=0, K=1, L=0 \mid V=1)}{P(A=0, K=1, L=0)}$$

$$= \frac{P(V=1)\,P(A=0 \mid V=1)\,P(K=1 \mid V=1)\,P(L=0 \mid V=1)}{P(A=0, K=1, L=0)}$$

$$= \frac{\frac{1}{2}\left(1-\frac{4}{6}\right)\frac{5}{6}\left(1-\frac{2}{6}\right)}{P(A=0, K=1, L=0)} = \frac{\frac{5}{54}}{P(A=0, K=1, L=0)}$$

Similarly $P(V=0 \mid A=0, K=1, L=0) = \frac{P(V=0)\,P(A=0 \mid V=0)\,P(K=1 \mid V=0)\cdot P(L=0 \mid V=0)}{P(A=0, K=1, L=0)}$

$$= \frac{\frac{1}{9}}{P(A=0, K=1, L=0)}$$

$$\frac{P(V=1 \mid A=0, K=1, L=0)}{P(V=0 \mid A=0, K=1, L=0)} = \frac{5}{54} \cdot 9 = \frac{45}{54} < 1 \Rightarrow \underline{\text{cannot}} \text{ say if has virus}$$

c) Directly from Table,

$P(V=1 \mid A=0, K=1, L=0) = P(V=0 \mid A=0, K=1, L=0) = \frac{1}{2}$,

not the same value as in B. Because in the table the observed data points are too few.

d) No. We must satisfy $P(V=0) + P(V=1) = 1$ in the prior(V) column. The rest of entries we can fill in number from $0 \sim 1$.

e) $$\frac{P(V=1 \mid A=0, K=1, L=0)}{P(V=0 \mid A=0, K=1, L=0)} = \frac{P(V=1)P(A=0|V=1)P(K=1|V=1)P(L=0|V=1)}{P(V=0)P(A=0|V=0)P(K=1|V=0)P(L=0|V=0)}$$

$$= \frac{\frac{5}{6}\cdot(1-\frac{2}{6})}{\frac{3}{6}\cdot(1-\frac{2}{6})} \cdot \frac{P(A=0|V=1)}{P(A=0|V=0)} = \frac{5}{3}\frac{P(A=0|V=1)}{P(A=0|V=0)}$$

as long as we make $\frac{P(A=0|V=1)}{P(A=0|V=0)} > \frac{3}{5}$ by change the tag of A, we make $\frac{P(A=0|V=1)}{P(A=0|V=0)} > 1$ =>

(e.g. make A=0 for V=1 & A=1 for V=0)

the email has virus => changed conclusion of b)

f) For $p(A, L, K \mid V=0)$ there are 8 combinations of A, L, K, so need to provide 7 values since the summation is 1.
Similarly for $p(A, L, K \mid V=1)$, still need 7 values, so need 14 values in total.

g) Because in reality, if a message contains virus and there's attachment, then the sender is mostly known to the receiver, otherwise no one would open attchment sent from unknown people. So the folks who create the virus message wouldn't do that.
There're other logical reasons in reality that breaks conditional independence as well.

a)

③ since $P(X=x) = \frac{u^x e^{-u}}{x!}$, $u>0$,

$$\Rightarrow p(x_1 \cdots x_n | u) = \prod_{i=1}^{n} \frac{u^{x_i} e^{-u}}{x_i!}$$

$$\Rightarrow \log p(x_1 \cdots x_n | u) = \sum_{i=1}^{n} \log\left(\frac{u^{x_i} e^{-u}}{x_i!}\right)$$

$$= \sum_{i=1}^{n} \log u^{x_i} + \log e^{-u} - \log x_i!$$

$$= \sum_{i=1}^{n} x_i \log u - u \log e - \log x_i!$$

$$\frac{\partial \log p(x_1 \cdots x_n | u)}{\partial u} = \sum_{i=1}^{n} \left(\frac{x_i}{u} - \log e\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{x_i}{u} = \sum_{i=1}^{n} \log e \Rightarrow \frac{1}{u} \sum_{i=1}^{n} x_i = n \log e$$

$$\Rightarrow u = \frac{\sum_{i=1}^{n} x_i}{n \log e} = \frac{\bar{x}}{\log e} = \bar{x} \quad (\log = \ln)$$

b) now $p(x_1 \cdots x_n | u) \cdot p(u)$

$$= \lambda e^{-\lambda u} \cdot \prod_{i=1}^{n} \frac{u^{x_i} e^{-u}}{x_i!}$$

$$\log P(x_1 \cdots x_n | u) P(u)$$

$$= \log \lambda - \lambda u + \sum_{i=1}^{n} x_i \log u - u \log e - \log x_i !$$

$$\frac{\partial \log P(x_1 \cdots x_n | u) P(u)}{\partial u} = -\lambda + \sum_{i=1}^{n} \left(\frac{x_i}{u} - 1 \right)$$

set $\frac{\partial \log P(x_1 \cdots x_n | u) P(u)}{\partial u} = 0$, we have

$$-\lambda + \frac{1}{u} \sum_{i=1}^{n} x_i - n = 0 \Rightarrow u = \frac{\sum_{i=1}^{n} x_i}{n + \lambda}$$