SI650 Homework 1

Chenyun Tao, uniq: cyuntao

1. a)

S	$P(B=1 \mid S)$	$P(U=1 \mid S)$	$P(E=1 \mid S)$	prior $P(S)$
1	0.71428	0.71428	0.42857	0.58334
0	0.4	0.6	0.4	0.41666

b) $P(S=1 \mid B=0, U=1, E=0)$

$$= \frac{P(S=1, B=0, U=1, E=0)}{P(B=0, U=1, E=0)}$$

$$= \frac{P(B=0, U=1, E=0 \mid S=1)P(S=1)}{P(B=0, U=1, E=0 \mid S=1)P(S=1) + P(B=0, U=1, E=0 \mid S=0)P(S=0)}$$

$$= \frac{P(B=0 \mid S=1)P(U=1 \mid S=1)P(E=0 \mid S=1)P(S=1)}{P(B=0 \mid S=1)P(U=1 \mid S=1)P(E=0 \mid S=1)P(S=1) + P(B=0 \mid S=0)P(U=1 \mid S=0)P(E=0 \mid S=0)P(S=0)}$$

$$= \frac{0.28572 \cdot 0.71428 \cdot 0.57143 \cdot 0.58334}{0.28572 \cdot 0.71428 \cdot 0.57143 \cdot 0.58334 + 0.6 \cdot 0.6 \cdot 0.6 \cdot 0.41666} = \boxed{0.43}$$

$P(S=0 \mid B=0, V=1, E=0)$

$= 1 - P(S=1 \mid B=0, V=1, E=0)$

$= 1 - 0.43$

$= \boxed{0.57} > P(S=1 \mid B=0, V=1, E=0)$

Since $P(S=1 \mid B=0, V=1, E=0)$

$< P(S=0 \mid B=0, V=1, E=0)$

I would conclude a is not a spam.

c) $P(S=1 \mid B=0, V=1, E=0) = 1$

$P(S=0 \mid B=0, V=1, E=0) = 0$

I get different values from values in problem B, probably because there is only 1 sample with $B=0$,

$V=1$, $E=0$. The sample size is too small.

d) No. For the prior $P(S)$, we must have $P(S=0)+P(S=1)=1$

e) Yes. I could flip the value E from 0 to 1 in an arbitrary entry in Table 1, where $S=0, E=0$. For example, I could change value E at the 4th row in table 1, as is shown below.

S	B	V	E
0	1	1	1
1	0	1	0
1	1	1	1
0	1	1	1

Then, $P(E=0 \mid S=0) = 0.4$

Thus, $P(S=0 \mid B=0, V=1, E=0)$

$$= \frac{0.6 \cdot 0.6 \cdot 0.4 \cdot 0.41666}{P(B=0, V=1, E=0)} = \frac{0.06}{P(B=0, V=1, E=0)}$$

Now, $P(S=1 \mid B=0, V=1, E=0) > P(S=0 \mid B=0, V=1, E=0)$, so the conclusion changes.

f) In general, $P(B, V, E \mid S)$

$= P(B \mid S) P(V \mid S, B) P(E \mid S, B, V)$, so the independence assumption could only hold if $P(V \mid S, B) = P(V \mid S)$ and $P(E \mid S, B, V) = P(E \mid S)$. However, in reality, usually the events S, B, V, E are correlated, so $P(V \mid S, B) = P(V \mid S)$ and $P(E \mid S, B, V) = P(E \mid S)$, and thus $P(B, V, E \mid S)$ $= P(B \mid S) P(V \mid S) P(E \mid S)$ don't necessarily hold.

3\. a) precision: $\frac{12}{20} = \boxed{0.6}$ recall: $\frac{12}{20} = \boxed{0.6}$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \boxed{0.6}$$

$$\text{MAP: } \frac{1}{20} \times \left(1 + \frac{2}{3} + \frac{3}{4} + \frac{4}{6} + \frac{5}{8} + \frac{6}{9} + \frac{7}{12} + \frac{8}{13} + \frac{9}{16} + \frac{10}{17} + \frac{11}{18} + \frac{12}{19}\right)$$

$$= \boxed{0.398}$$

b) $CG_{10} = 2 + 1 + 2 + 1 + 2 + 1 = \boxed{9}$

$$DCG_{10} = 2 + \frac{1}{\log_2 3} + \frac{2}{\log_2 4} + \frac{1}{\log_2 6} + \frac{2}{\log_2 8}$$

$$+ \frac{1}{\log_2 9} = \boxed{4.9999}$$

$$IDCG_{10} = 2 + \frac{2}{\log_2 2} + \frac{2}{\log_2 3} + \frac{2}{\log_2 4} +$$

$$\frac{2}{\log_2 5} + \frac{1}{\log_2 6} + \frac{1}{\log_2 7} + \frac{1}{\log_2 8} + \frac{1}{\log_2 9}$$

$$+ \frac{1}{\log_2 10} = 8.816$$

$$NDCG_{10} = \frac{DCG_{10}}{IDCG_{10}} = \boxed{0.5671}$$