CS510: Advanced Topics in Information Retrival **Fall 2017**

Assignment 2

Zhuo Li(zhuol2) *Handed In: September 13, 2017*

1. Entropy

   a. The minimum value of $H(W)$ happens when only one word has probability 1 and all others get 0:
   $$H(W)_{min} = 0$$
   The maximum value of $H(W)$ happens when each word has exactly the same probability:
   $$\begin{aligned} H(W)_{max} &= \sum_{i=1}^{N} -p(W = w_i) \log_2 p(W = w_i) \\ &= -N \frac{1}{N} \log_2 \frac{1}{N} \\ &= \log_2 N \end{aligned}$$

   b. The article reach minimum $H(W)$:
   $$\{w_1\}$$
   The article reach maximum $H(W)$:
   $$\{w_1\ w_2\ w_3\ w_4\ w_5\ w_6\}$$

   c. When A_1 and A_2 has exact one word respectively and they are different, the maximum $H(W)$ for A_3 can be reached:
   $$H(W) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$
   These articles can be:
   $$\begin{aligned} &A_1 : \{w_1\} \\ &A_2 : \{w_2\} \\ &A_3 : \{w_1\ w_2\} \end{aligned}$$

2. Conditional Entropy and Mutual Information

   a. Because $p(x|x) = 1$, then
   $$\begin{aligned} H(X|X) &= \sum_{x \in \Omega_x} p(x) H(X|X = x) \\ &= -\sum_{x \in \Omega_x} p(x) \sum_{x \in \Omega_x} p(x|x) \log_2 p(x|x) \\ &= 0 \end{aligned}$$

b. Because X and Y are independent

$$p(x, y) = p(x)p(y)$$

Then

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x)p(y)} \\ &= 0 \end{aligned}$$

3. Mutual Information of Words

a. We can have

$$p(X_A = 0, X_B = 1) = \frac{N_B - N_{AB}}{N}$$
$$p(X_A = 0, X_B = 0) = \frac{N - N_A - N_B + N_{AB}}{N}$$

b. From Table 1 we have

$$\begin{aligned} p(X_{computer} = 1) &= \frac{1390}{26394} \\ p(X_{computer} = 0) &= \frac{25004}{26394} \end{aligned} \tag{1}$$

$$\begin{aligned} p(X_{program} = 1) &= \frac{2370}{26394} \\ p(X_{program} = 0) &= \frac{24024}{26394} \end{aligned} \tag{2}$$

$$\begin{aligned} p(X_{baseball} = 1) &= \frac{2144}{26394} \\ p(X_{baseball} = 0) &= \frac{24250}{26394} \end{aligned} \tag{3}$$

Then we can calculate using (1)(2)(3)

$$
\begin{aligned}
& I(X_{computer}; X_{program}) \\
&= p(x_{computer}=0, x_{program}=0) \log \frac{p(x_{computer}=0, x_{program}=0)}{p(x_{computer}=0)p(x_{program}=0)} \\
&+ p(x_{computer}=0, x_{program}=1) \log \frac{p(x_{computer}=0, x_{program}=1)}{p(x_{computer}=0)p(x_{program}=1)} \\
&+ p(x_{computer}=1, x_{program}=0) \log \frac{p(x_{computer}=1, x_{program}=0)}{p(x_{computer}=1)p(x_{program}=0)} \\
&+ p(x_{computer}=1, x_{program}=1) \log \frac{p(x_{computer}=1, x_{program}=1)}{p(x_{computer}=1)p(x_{program}=1)} \\
&= \frac{22983}{26394} \log \frac{22983/26394}{25004/26394 \times 24024/26394} \\
&+ \frac{2021}{26394} \log \frac{2021/26394}{25004/26394 \times 2370/26394} \\
&+ \frac{1041}{26394} \log \frac{1041/26394}{1390/26394 \times 24024/26394} \\
&+ \frac{349}{26394} \log \frac{349/26394}{1390/26394 \times 2370/26394} \\
&= 0.00921
\end{aligned}
$$

$$
\begin{aligned}
& I(X_{computer}; X_{baseball}) \\
&= p(x_{computer}=0, x_{baseball}=0) \log \frac{p(x_{computer}=0, x_{baseball}=0)}{p(x_{computer}=0)p(x_{baseball}=0)} \\
&+ p(x_{computer}=0, x_{baseball}=1) \log \frac{p(x_{computer}=0, x_{baseball}=1)}{p(x_{computer}=0)p(x_{baseball}=1)} \\
&+ p(x_{computer}=1, x_{baseball}=0) \log \frac{p(x_{computer}=1, x_{baseball}=0)}{p(x_{computer}=1)p(x_{baseball}=0)} \\
&+ p(x_{computer}=1, x_{baseball}=1) \log \frac{p(x_{computer}=1, x_{baseball}=1)}{p(x_{computer}=1)p(x_{baseball}=1)} \\
&= \frac{22883}{26394} \log \frac{22883/26394}{25004/26394 \times 24250/26394} \\
&+ \frac{2121}{26394} \log \frac{2121/26394}{25004/26394 \times 2144/26394} \\
&+ \frac{1367}{26394} \log \frac{1367/26394}{1390/26394 \times 24250/26394} \\
&+ \frac{23}{26394} \log \frac{23/26394}{1390/26394 \times 2144/26394} \\
&= 0.00320
\end{aligned}
$$

c. We can certainly find that $I(X_{computer}; X_{program}) > I(X_{computer}; X_{baseball})$. This conforms with my intuition because the higher the mutual information, the closer the two words co-relate with each other. **Computer** and **program** are definitely more related than **computer** and **baseball**.

4. Kullback-Leibler Divergence (KL Divergence)

   a. We have

      1) $D(p||q) \geq 0$
      2) $D(p||q) = 0 \; iff \; p = q$, i.e. the two distributions are the same.

   b. Assume

$$\begin{aligned} p(X=0) &= 1/2 \\ p(X=1) &= 1/2 \end{aligned} \tag{1}$$

$$\begin{aligned} q(X=0) &= 1/3 \\ q(X=1) &= 2/3 \end{aligned} \tag{2}$$

   Then we have:

$$\begin{aligned} D(p||q) &= H(p,q) - H(p) \\ &= p(X=1)\log\frac{p(X=1)}{q(X=1)} + p(X=0)\log\frac{p(X=0)}{q(X=0)} \\ &= \frac{1}{2}\log\frac{1/2}{2/3} + \frac{1}{2}\log\frac{1/2}{1/3} \\ &= \log_2 3 - 2/3 = 0.918 \end{aligned}$$

$$\begin{aligned} D(q||p) &= H(p,q) - H(q) \\ &= q(X=1)\log\frac{q(X=1)}{p(X=1)} + q(X=0)\log\frac{q(X=0)}{p(X=0)} \\ &= \frac{2}{3}\log\frac{2/3}{1/2} + \frac{1}{3}\log\frac{1/3}{1/2} \\ &= 5/3 - \log_2 3 = 0.082 \end{aligned}$$

   So $D(p||q) \neq D(q||p)$.

   c. If an event has 0 probability in q, the denominator of our formula will be 0 which is incorrect. KL divergence is defined when they have only nonzero entries, so we should avoid to have this case. However, if we have to run with 0 entry, we can smooth the distributions in some way, for instance with a Bayesian prior, or (similarly) taking the convex combination of the observation with some valid (nonzero) distribution.