1 Information Theory

1.1 Entropy and Probability Distributions

a) The marginal distribution of X =

$$\{(\frac{1}{8}+\frac{1}{16}+\frac{1}{16}+\frac{1}{4}),(\frac{1}{16}+\frac{1}{8}+\frac{1}{16}+0),(\frac{1}{32}+\frac{1}{32}+\frac{1}{16}+0),(\frac{1}{32}+\frac{1}{32}+\frac{1}{16}+0)\}$$

$$= \{\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\}$$

The marginal distribution of Y =

$$\{(\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\frac{1}{32}),(\frac{1}{16}+\frac{1}{8}+\frac{1}{32}+\frac{1}{32}),(\frac{1}{16}+\frac{1}{16}+\frac{1}{16}+\frac{1}{16}),(\frac{1}{4}+0+0+0)\}$$

$$= \{\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\}$$

b) The Entropy of X, H(X) =

$$= -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{4}log_2\frac{1}{4} - \frac{1}{8}log_2\frac{1}{8} - \frac{1}{8}log_2\frac{1}{8}$$

$$= .5 + .5 + .375 + .375$$

$$= 1.75\, bits$$

The Entropy of Y, H(Y) =

$$= 4 * (-\frac{1}{4}log_2\frac{1}{4})$$

$$= 4 * .5$$

$$= 2\, bits$$

c)

$$H(X|Y) = \sum_{i=1}^{4} p(Y=i)H(X|Y=i)$$

$$= \frac{1}{4}H(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}) + \frac{1}{4}H(\frac{1}{4},\frac{1}{2},\frac{1}{8},\frac{1}{8}) + \frac{1}{4}H(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}) + \frac{1}{4}H(1,0,0,0)$$

$$= \frac{1}{4} * \frac{7}{4} + \frac{1}{4} * \frac{7}{4} + \frac{1}{4} * 2 + \frac{1}{4} * 0$$

$$= 1.375\, bits$$

$$H(Y|X) = H(X|Y) - H(X) + H(Y)$$

$$= 1.375 - 1.75 + 2$$

$$= 1.625\, bits$$

$$H(X,Y) = H(Y|X) + H(X)$$
$$= 1.625 + 1.75$$
$$= 3.375\, bits$$

d) The mutual information I(X;Y)

$$I(X;Y) = H(X) - H(X|Y) \tag{1}$$
$$= 1.75 - 1.375$$
$$= 0.375\, bits$$

If we calculate I(X;Y) as follows:

$$I(X;Y) = H(Y) - H(Y|X) \tag{2}$$
$$= 2 - 1.625$$
$$= 0.375\, bits$$

So, (1) and (2) are generating the same value which authenticates the validity of the symmetry property of the mutual information.

2 N -gram Language Models

2.1 Language Model Training

a) see ngram_LM.py

b)

unigram	count
the	734066
of	360504
to	330708
and	326187
in	250687
a	228170
that	154963
is	14967
s	121474
for	104848
it	85982
as	80903
be	74177
with	73328
on	69736

bigram	count
of the	82750
in the	65399
to the	32188
and the	27717
on the	20190
it is	19828
to be	18774
that the	15456
for the	17383
the us	16979
the world	16123
with the	15745
by the	14295
of a	13571
at the	13429