Joint probabilities and entropy

We have been looking at the problem of coding a sequence of length n with symbols/events drawn from some alphabet of size N. For any fixed n and N, there are N^n such sequences.

$$\begin{array}{ccccccc}
A_1 & A_1 & A_1 & A_1 & \cdots & A_1 & A_1 \\
A_1 & A_1 & A_1 & A_1 & \cdots & A_1 & A_2 \\
\vdots & & & \vdots & & & \vdots \\
A_1 & A_1 & A_1 & A_1 & \cdots & A_1 & A_N \\
A_1 & A_1 & A_1 & A_1 & \cdots & A_2 & A_1 \\
\vdots & & & \vdots & & & \vdots \\
A_7 & A_3 & A_8 & A_3 & \cdots & A_N & A_2 \\
\vdots & & & \vdots & & & \vdots \\
A_N & A_N & A_N & A_N & \cdots & A_N & A_N
\end{array}$$

$$\leftarrow \quad n \text{ symbols} \quad \rightarrow$$

Let's think of the set of N^n sequences mentioned as a new alphabet. We define a probability function $p(X_1, X_2, \ldots, X_n)$ on this alphabet. To keep the notation simple, we assume the alphabet is the set $\{1, 2, \ldots, N-1, N\}$, so that we are now thinking of:

$$\begin{array}{ccccccc}
1 & 1 & 1 & 1 & \cdots & 1 & 1 \\
1 & 1 & 1 & 1 & \cdots & 1 & 2 \\
\vdots & & & \vdots & & & \vdots \\
1 & 1 & 1 & 1 & \cdots & 1 & N \\
1 & 1 & 1 & 1 & \cdots & 2 & 1 \\
\vdots & & & \vdots & & & \vdots \\
7 & 3 & 8 & 3 & \cdots & N & 2 \\
\vdots & & & \vdots & & & \vdots \\
N & N & N & N & \cdots & N & N
\end{array}$$

Let $(i_1, i_2, \ldots, i_n)$ denote a possible sequence of length n. It has a probability

$$p(\ X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n).$$

Since $p()$ is a probability function, we have

$$\sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_n=1}^{N} p(\ X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \quad = \quad 1$$

We also also define the entropy:

$$H \ = \ \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_n=1}^{N} p(\ X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \log \frac{1}{p(\ X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)}$$

What is the optimal prefix code for this (huge) alphabet? To answer this question, we could create a Huffman code using the function $p(X_1, \ldots, X_n)$. The average code length of this alphabet would be within 1 bit of the entropy (recall lecture 4).

Unfortunately, creating such a Huffman code is not practical. First, the number of sequences N^n is huge for large N and large n. Representing such a probability function (so that we could create a Huffman code) would take alot of space and time. Second, how would we choose the function $p(X_1, \ldots, X_n)$? Take the case of English text. How could we specify the probability of all possible sequences of English text of length n?

The way around this problem is to approximate $p(X_1, \ldots, X_n)$ by considering conditional probabilities that related neighboring symbols only (e.g. X_j and X_{j+1}) and then construct codes based on these conditional probabilities.

Marginal Probability

If we have a set of n random variables and with *joint probabilities* $p(X_1, X_2, \ldots, X_n)$, then we can define a marginal probability function on any subset of the variables by summing over all possible occurances of the remaining variables. (Recall MATH 323.)

For example, the marginal probability of X_7 is:

$$p(X_7) \;=\; \sum_{i_1=1}^{N} \cdots \sum_{i_6=1}^{N} \sum_{i_8=1}^{N} \cdots \sum_{i_n=1}^{N} p(X_1 = i_1,\; X_2 = i_2, \ldots, X_6 = i_6,\; X_7,\; X_8 = i_8, \ldots, X_n = i_n)$$

Here we have $n-1$ summations, that is, we are missing the summation over X_7. For a particular value such as, $X_7 = i_3$, we have

$$p(X_7 = i_7) \;=\; \sum_{i_1=1}^{N} \cdots \sum_{i_6=1}^{N} \sum_{i_8=1}^{N} \cdots \sum_{i_n=1}^{N} p(X_1 = i_1,\; X_2 = i_2, \ldots, X_6 = i_6,\; X_7 = i_7,\; X_8 = i_8, \ldots, X_n = i_n)$$

Notice that to specify the function $p(X_7)$ we need to specify only $N-1$ values, since

$$\sum_{j=1}^{N} p(X_7 = j) \;=\; 1$$

and specifying $N-1$ of the values is sufficient to specify all N.

Similarly, the marginal probability over the variables X_7, X_8 is the joint probability function:

$$p(X_7, X_8) \;=\; \sum_{i_1=1}^{N} \cdots \sum_{i_6=1}^{N} \sum_{i_9=1}^{N} \cdots \sum_{i_n=1}^{N} p(X_1 = i_1, \ldots, X_6 = i_6,\; X_7, X_8,\; X_9 = i_9 \ldots, X_n = i_n)$$

Again, if we examine this marginal function at a particular value of $(X_7 = i_7,\; X_8 = i_8)$, then we would write

$$\begin{aligned} & p(X_7 = i_7, X_8 = i_8) \\ = \quad & \sum_{i_1=1}^{N} \cdots \sum_{i_6=1}^{N} \sum_{i_9=1}^{N} \cdots \sum_{i_n=1}^{N} p(X_1 = i_1, \ldots, X_6 = i_6,\; X_7 = i_7,\; X_8 = i_8,\; X_9 = i_9 \ldots, X_n = i_n) \end{aligned}$$

Notice that there are $n-2$ summations here. There is no summation over X_7, X_8. Also note that we can specify this joint probability function by specifying $N^2 - 1$ values.

Stationarity property (also known as shift invariance)

One common assumption for many types of data is that the probability of any subsequence occuring doesn't depend on where in the sequence the subsequence occurs. For example, when we say that a `q` is likely to be followed by `u`, we have not specified where in the sequence the `q` occurs. *Anytime* we have a `q`, it is likely followed by a `u`. Similarly, although `s` is a relatively common letter in English text, if `thr` occurs, then it is unlikely to be followed by `s`.

We say the probability function $p(X_1, \ldots, X_n)$ is *stationary* (or shift invariant) if

$$p(X_1 = i_1, X_2 = i_2, \ldots, X_m = i_m) = p(X_{1+l} = i_1,\ X_{2+l} = i_2,\ \ \ldots, X_{m+l} = i_m)$$

for all $(i_1, i_2, \ldots, i_m)$, and for all l, m such that $l + m \leq n$. That is, the probability of a given subsequence occuring does not depend on where it occurs.

Note that stationarity implies that random variables X_i are identically distributed (since this is the case $m = 1$). However, stationarity does not imply that the random variables are independently distributed. Indeed it is the lack of independence that we will be interested in.

If a probability function is stationary, then any marginal probability function is also stationary. The reason is that a marginal probability is just a sum of (stationary) joint probabilities.

For example, consider $p(X_1, X_5, X_{30})$. If $p(X_1, \ldots, X_n)$ is stationary, then

$$p(\ X_1, X_5, X_{30}) \ = \ p(X_2, X_6, X_{31}) \ = \ p(X_3, X_7, X_{32}) \ = \ etc$$

This means the following. If we plug in any particular values for X_1, X_5, X_{30}, say (i_2, i_3, i_1) respectively, then

$$p(\ X_1 = i_2, X_5 = i_3, X_{30} = i_1) \ = \ p(X_2 = i_2,\ X_6 = i_3,\ X_{31} = i_1) = etc$$

Assuming $p(X_1, \ldots, X_n)$ is stationary is very strong. For example, suppose you concatenated all the files in your user directory on SOCS unix system into one big file. This big file would contain various emails that you have received as well as source code in various languages (Java, C, etc) as well as perhaps a set of images or videos or audio files. These files all have very different statistical properties. So, if we wanted to capture the probabilities, we would be best off NOT using a stationary model. On the other hand, if you only concatenated your C files together, then a stationary model would be more appropriate.

Similarly, if you set up a microphone in class and recorded the sound in the room over an entire day (and repeated this over many days) then the result files would NOT be stationary. For example, there would be little sound before 10:30 on MWF since there is no class in our room before and directly afterwards. However, between say 10:25 and 11:25, you would describe the files as stationary, namely, they would contain the sound of my voice.

Conditional Probability

Let $\mathcal{E}$ be the alphabet (or "sample space") consisting of the set of N^n sequences of symbols from the base alphabet $\{1, \ldots, N\}$ (or $\{i_1, \ldots, i_N\}$). Suppose that $\mathcal{E}_1 \subseteq \mathcal{E}$ and $\mathcal{S}_2 \subseteq \mathcal{E}$ are two events, namely subsets of the sample space. Then, the conditional probability $p(\ \mathcal{E}_1 \mid \mathcal{E}_2\ )$ is defined:

$$p(\mathcal{E}_1 \mid \mathcal{E}_2\ ) \ \equiv \ \frac{p(\mathcal{E}_1 \cap \mathcal{E}_2\ )}{p(\mathcal{E}_2\ )}$$

We define conditional probabilites on events defined by the random variables $X_1, \ldots, X_n$. For example,

$$p(X_8 = i_1 \mid X_7 = i_3) \;=\; \frac{p(X_8 = i_1, X_7 = i_3)}{p(X_7 = i_3)}$$

More generally, we can define the function $p(X_8|X_7)$. This function requires that we specify $N(N-1)$ values, namely, for each value of X_7, we would need to specify $N-1$ values of X_8. The reason for only needing to specify $N-1$ values of X_8 is that, for any fixed i_7, the probabilities $p(X_8 = i_8|X_7 = i_7)$ must add up to 1. Thus, for example, if one were to specify the probabilities of the first $N-1$ values of X_8, the Nth would be determined by

$$p(X_8 = N|X_7 = i_7) \;=\; 1 - \sum_{i_8=1}^{N-1} p(X_8 = i_8|X_7 = i_7).$$

Conditional Entropy

Earlier we defined the joint entropy $H(X_1, X_2, \ldots, X_n)$ of a vector i.e. sequence of random variables. Let's next write the joint entropy in terms of conditional and marginal probabilities. We present the idea using just two variables X_1, X_2.

$$\begin{aligned}
& H(X_1, X_2) \\
= \;& \sum_{i_2=1}^{N}\sum_{i_1=1}^{N} p(X_2 = i_2, X_1 = i_1) \log \frac{1}{p(X_2 = i_2 \ , \ X_1 = i_1)} \\
= \;& \sum_{i_2=1}^{N}\sum_{i_1=1}^{N} p(X_2 = i_2, X_1 = i_1)(\log \frac{1}{p(X_2 = i_2 \mid X_1 = i_1)} + \log \frac{1}{p(X_1 = i_1)}) \\
= \;& \sum_{i_1=1}^{N} p(X_1 = i_1) \sum_{i_2=1}^{N} p(X_2 = i_2|X_1 = i_1) \log \frac{1}{p(X_2 = i_2 \mid X_1 = i_1)} + \sum_{i_1=1}^{N} p(X_1 = i_1) \log \frac{1}{p(X_1 = i_1)} \\
= \;& \sum_{i_1=1}^{N} p(X_1 = i_1) H(X_2|X_1 = i_1) + H(X_1)
\end{aligned}$$

Define the *conditional entropy* of X_2 given X_1,

$$H(X_2|X_1) \;\equiv\; \sum_{i_1=1}^{N} p(X_1 = i_1) H(X_2|X_1 = i_1)$$

Then,

$$H(X_1, X_2) = H(X_2|X_1) + H(X_1)$$

which is very nice. In the context of data compression, we can think of this as follows: a lower bound on the number of bits that we need to encode a pair a random variables is equal to the lower bound on the number of bits we need to encode one of them, plus the lower bound on the number of bits we need to encode the second one given we know the first one.

ASIDE: Your intuition might be that conditional entropy $H(X_2|X_1)$ should be defined as

$$\sum_{i_1=1}^{N}\sum_{i_2=1}^{N} p(X_2 = i_2|X_1 = i_1) \log \frac{1}{p(X_2 = i_2 \mid X_1 = i_1)}$$

However, this quantity is

$$\sum_{i_1=1}^{N} H(X_2|X_1 = i_1)$$

which is the sum of the conditional entropies of X_2 given $X_1 = i_1$, summed over all i_1. This quantity doesn't take account of the (possibly) non-uniform probabilities of $p(X_1 = i_1)$.