# Assignment 02: Markov Models

1. **Temporal Subsampling of a Discrete Time Markov Process** (*10 points*)

   Suppose $X_1, X_2, \ldots$ forms a Markov process (that is **not** necessarily homogeneous for the purpose of this problem). Then recall that as per our definition,

   $$p(x_n|x_{n-1}, x_{n-2}, \ldots x_1) = p(x_n|x_{n-1}) \tag{1}$$

   for all $n \geq 1$. It can be seen that the definition in (1) is also equivalent[1] to the condition that

   $$p(x_n, x_{n-1}, x_{n-2}, \ldots x_1) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}) \tag{2}$$

   for all $n \geq 1$.

   (a) Show that for any positive integer $n$ and any $k < n$

   $$p(x_n, x_{n-1}, \ldots x_{n-k}) = p(x_{n-k})p(x_{n-k+1}|x_{n-k})p(x_{n-k+2}|x_{n-k+1})\cdots p(x_n|x_{n-1}). \tag{3}$$

   This result is equivalent to the condition that for any positive integer $n$ and any $k < n$

   $$p(x_n|x_{n-1}, x_{n-2}, \ldots x_{n-k}) = p(x_n|x_{n-1}). \tag{4}$$

   It is also immediately obvious that (3) and (4) imply (1) and (2), respectively. Thus the conditions in (1), (2), (3), and (4) are all equivalent and any of these can be used as the defining condition for a Markov process.

   (b) For $n = 4$, (2) becomes

   $$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3). \tag{5}$$

   Formally show that (1) also implies that

   $$p(x_1, x_3, x_4) = p(x_1)p(x_3|x_1)p(x_4|x_3). \tag{6}$$

   (c) From the result of the preceding part, conclude that

   $$p(x_4|x_3, x_1) = p(x_4|x_3). \tag{7}$$

   (d) Using the results from the preceding parts, formally show that

   $$p(x_1, x_2, x_4) = p(x_1)p(x_2|x_1)p(x_4|x_2). \tag{8}$$

   (e) From the result of the preceding part, conclude that

   $$p(x_4|x_2, x_1) = p(x_4|x_2). \tag{9}$$

   (f) By continuing this line of reasoning, we can argue that if $k$ is some positive integer and $n_1, n_2, \ldots n_k$ is any strictly increasing sequence of positive integers, then

   $$p(x_{n_1}, x_{n_2}, \ldots x_{n_k}) = p(x_{n_1})p(x_{n_2}|x_{n_1})p(x_{n_3}|x_{n_2})\cdots p(x_{n_k}|x_{n_k-1}). \tag{10}$$

   State the above relation in words.

[1] Using Bayes' rule, one can readily see that (1) implies (2). Can you also argue the converse?

2. **Markov Models for Text: Seuss and Saki** The files "spamiam.txt" and "saki_story.txt" available on the website have poetry and prose of specific genres[2]. For this problem, use the text in these files to empirically estimate probabilities and transition probabilities as indicated. Ignore any characters in these files other than the $26$ alphabets 'a'-'z' (use white space and carriage returns as indicated in specific parts). Also ignore any case distinctions among alphabets (example 'C' and 'c' are equivalent).

   **For each of the sub-parts indicated below, print the $100$ words that you generate in the form of a $10 \times 10$ array and circle any valid English words that you recognize.**

   (a) Assuming that the $26$ letters of the alphabet are equiprobable. Generate one hundred random $4$ letter words by selecting the $4$ individual letters of each word independently.

   (b) Estimate the probabilities of individual letters using "spamiam.txt". Generate one hundred random $4$ letter words by selecting the $4$ individual letters of each word independently according to the estimated probability distribution.

   (c) Again using the file "spamiam.txt", estimate the transition probabilities, $P(x_{n+1}|x_n)$, for all $26$ possible values of $x_n$ - the $n^{th}$ letter in a word and $x_{n+1}$ - the $(n+1)^{th}$ letter in a word (assume that these probabilities are independent of $n$). Also for this part and the next, for your estimation of transition probabilities, use only the letters inside a word for the computation and do not incorporate letters from adjacent words (with a blank in between). Generate one hundred random $4$ letter words by first generating a letter at random according to the probability mass function (pmf) in 2b and then generating remaining letters according to appropriate transition probabilities. *Note: You may default to the model of 2b if you end up with a situation where your estimate of $P(x_{n+1}|x_n)$ is zero for all values of $x_{n+1}$.*

   (d) Once again use the file "spamiam.txt", to estimate the transition probabilities $P(x_{n+1}|x_n, x_{n-1})$, for all possible values of the successive letters. Generate one hundred random four letter words using these estimated probabilities. Make reasonable assumptions that generalize what was indicated in 2c.

   (e) Repeat parts 2b-2d using the file "saki_story.txt".

   (f) Comment on your results.

   (g) **Extra Credit:** Estimate the entropy rate for each of the Markov models you developed and compare these both across models for a single data file and across the two data files. You may need to make suitable assumptions in order to determine your answers (which may be hard/impossible to validate).

[2] The first of these was obtained from: http://www.seuss.org/seuss/spam.i.am.html, which seems to be now defunct. The second is from http://www.gutenberg.net/, which you may choose to visit for additional information and more.