CS 530: Geometric and Probabilistic Methods in Computer Science Homework 2 (Fall '15)

1. You are given a deck of ordinary playing cards from which all cards except the jacks, queens, kings, and aces have been removed. Because I hate the color red, I also have removed the queens and kings of $\diamondsuit$ and $\heartsuit$. The discrete r.v. X has outcomes $\{\clubsuit, \spadesuit, \diamondsuit, \heartsuit\}$ and the discrete r.v. Y has outcomes $\{\text{jack}, \text{queen}, \text{king}, \text{ace}\}$.

   (a) Compute the marginal p.m.f.'s, p_X and p_Y, and the joint p.m.f., p_{XY}.

   (b) Are X and Y statistically independent?

   (c) Compute the conditional p.m.f.'s, $p_{X|Y}$ and $p_{Y|X}$.

   (d) Compute the entropies H_X, H_Y, and H_{XY}.

   (e) Compute the conditional entropies $H_{X|Y}$ and $H_{Y|X}$.

   (f) Compute the mutual information, I_{XY}.

   (g) I draw a card. How much information do you receive if you are told the card is black? red?

2. A fair coin is flipped until the first head appears. Let X denote the number of flips required. Find the entropy, H_X, in bits. The following expressions may be useful: $\sum_{n=1}^{\infty} r^n = r/(1-r)$ and $\sum_{n=1}^{\infty} nr^n = r/(1-r)^2$.

3. The p.m.f. for the 12367 most frequently used words in English is approximately:

$$p(n) = \begin{cases} \frac{0.1}{n} & \text{for } 1 \leq n \leq 12367 \\ 0 & n > 12367. \end{cases}$$

   This remarkable fact is known as Zipf's law, and applies to many languages (Zipf, 1949). If we assume that English is generated by picking words at random according to this distribution, what is the entropy of English (per word)?

4. JPEG[1] is by far the most widely used compressed image format. However, unlike the GIF format, which uses a *lossless* compression method, JPEG compression decreases image quality, *i.e.*, it is a *lossy* method. In this exercise, JPEG will be viewed as an *information channel.* The grey values of the pixels of an image before and after JPEG compression will be considered to be samples of two non-independent discrete r.v.'s X and Y. Note that a pixel of an uncompressed image with 256 grey levels can contain at most 8 bits of information. The *lena* image on the class homepage is stored in a PGM format. This format does no compression. The *lena-jpeg* image is also stored in the PGM format. However, the *lena-jpeg* image has already undergone JPEG compression. Using pixels of the *lena* and *lena-jpeg* images, compute the following:

   - H_X - The entropy of the *lena* image.
   - H_Y - The entropy of the *lena-jpeg* image.
   - $H_{Y|X}$ - The channel noise.
   - $H_{X|Y}$ - The channel loss.
   - I_{XY} - The mutual information, *i.e.*, the amount of information which actually passes through the JPEG information channel.

5. Download the complete amino acid sequence for the chromosome of the bacterium, *Buchnera*, from the class webpage. Amino acids are represented by single characters from the set:

   $$\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, \$\}$$

   Using MATLAB, compute the distribution of the amino acids. Plot the distribution and compute the entropy of the *Buchnera* chromosome.

6. Assuming that the amino acid sequence of the *Buchnera* chromosome can be modeled as a first order Markov process, compute the conditional entropy per amino acid. [Hint: Use the amino acid sequence to build a 21×21 transition matrix, $P_{t\,|\,t-1}$ and compute the conditional entropy, $H_{t\,|\,t-1}$, as described in the class notes.]

[1] Joint Photograph Experts Group.

7. Using MATLAB, compute the eigenvectors and eigenvalues of the transition matrix you constructed in the last problem. Is the Markov process irreducible and aperiodic? If so, plot the limiting distribution and compare it to the distribution of amino acids you plotted in the first problem.

8. You are going to successively flip a coin until the pattern HHT appears; that is, until you observe two successive heads followed by a tail. When HHT appears the game ends. In order to calculate some properties of this game, you set up a Markov process with the following states: S, H, HH, and HHT, where S represents the starting point, H represents a single observed head, HH represents two successive heads, and HHT is the sequence you are looking for. Observe that if you have just tossed a tails, followed by a heads, a next toss of a tails effectively starts you over again in your quest for the HHT sequence. Set up the transition probability matrix.

9. Let $x_i^{(n)}$ denote the quality of the n-th item produced by a production system with $x_0^{(n)}$ being "good" and $x_1^{(n)}$ being "defective." Suppose that $\mathbf{x}$ evolves as a Markov chain whose transition probability matrix is:
$$\begin{bmatrix} .99 & .12 \\ .01 & .88 \end{bmatrix}$$
What is the probability that the fourth item is defective, given that the first item is defective?

10. Consider the Markov process whose transition probability matrix, $\mathbf{P}$, is given by:
$$\begin{bmatrix} .4 & .0 & .0 & .1 \\ .0 & .7 & .3 & .2 \\ .3 & .2 & .3 & .4 \\ .3 & .1 & .4 & .3 \end{bmatrix}$$
Suppose that the initial distribution is $\mathbf{x}^{(1)} = [.1\ .1\ .5\ .3]^{\mathrm{T}}$.

    (a) Compute $\mathbf{x}^{(2)}, \mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$

    (b) Compute $\mathbf{P}^2$ and $\mathbf{P}^3$.

    (c) Compute $\mathbf{x}$ where $\mathbf{x} = \mathbf{P}\mathbf{x}$.

11. Hazel and Naomi each have a brown paper grocery bag containing three exotic fruits. Initially, Hazel's bag contains three persimmons and Naomi's bag contains three kumquats. Once every day, Hazel and Naomi swap one fruit chosen at random from their bags.

    (a) Derive an expression for the probability that Hazel will have $k+1$ kumquats today given that she had k kumquats yesterday.

    (b) Derive an expression for the probability that Hazel will have k kumquats today given that she had k kumquats yesterday.

    (c) Derive an expression for the probability that Hazel will have $k-1$ kumquats today given that she had k kumquats yesterday.

    (d) Give the transition matrix for the Markov process.

    (e) Is the Markov process irreducible? aperiodic? Prove your answers.

    Hint: Observe that the number of kumquats which Hazel has in her bag always equals the number of persimmons which Naomi has in her bag and *vice versa.*