ECE 3077, Summer 2014

Homework #2

Due Thursday May 29, in class

Reading: B&T 1.3–1.4

1. Using you class notes, prepare a 1-2 paragraph summary of what we talked about in class in the last week. I do not want just a bulleted list of topics, I want you to use complete sentences and establish context (Why is what we have learned relevant? How does it connect with other things you have learned here or in other classes?). The more insight you give, the better.

2. Suppose that a point (x, y) is chosen from the square $S = [-1, 1] \times [-1, 1]$ using the uniform probability law. Note that since the area of S here is 4, we have

$$\mathrm{P}\left((x, y) \in A\right) = \frac{\mathrm{area}(A)}{4} \quad \text{for all} \quad A \subset S.$$

   (a) Sketch the regions of the plane corresponding to the events

$$A = \{|x| + |y| \leq 1\}$$
$$B = \{|x|^2 + |y|^2 \leq 1\}.$$

   (b) Calculate $\mathrm{P}\left(B \mid A\right)$.

   (c) Calculate $\mathrm{P}\left(A \mid B\right)$.

3. Suppose you are given a box with three drawers, one with two rare gold coins, one with a gold coin and a penny, and one with two pennies. You get to pick two coins from the box. Suppose that you begin by choosing a drawer at random and then pick one of the coins from the drawer at random. If your first draw is one of the gold coins, are you better off taking the other coin from the same drawer or randomly picking a coin from one of the other two drawers?

4. Tom (who is a boy) has two siblings, one older and one younger. We will assume that the their genders were determined by the equivalent of a fair coin flip, and that their genders are independent of one another.

   (a) What is the probability that Tom has two sisters?

   (b) You are told that Tom comes from a family with **at least one** girl. What is the probability, conditioned on this information, that Tom has two sisters?

   (c) You are told that Tom comes from a family where the **oldest child** is a girl. What is the probability, conditioned on this information, that Tom has two sisters?

5. Alice's cooler contains 13 lemonades and 5 Sprites. Bob's cooler contains 3 lemonades and 8 Sprites. Unfortunately, the two coolers look identical, making them indistinguishable from the outside.

   (a) Suppose Bob selects a cooler at random and then chooses a drink. What is the probability it is a lemonade?

   (b) Suppose that the drink he chose was indeed a lemonade. What is the probability he is choosing from is own cooler?

   (c) Now suppose he pulls out two more drinks, both of which are lemonades. Now what is the probability he is choosing from his own cooler?

6. M&Ms come in six colors: brown, yellow, green, blue, red, and orange. In a regular pack of M&Ms, these colors appear with equal probability and according to the uniform law.

   Mars Inc., which manufactures M&Ms, is running a contest in which one out of every 100,000 packages contain only green M&Ms; if you find such a package, you win a lavish prize. You purchase a pack, and begin taking out M&Ms...

   (a) Suppose that you pull out 5 M&Ms and they are all green. What is the probability that you are holding a winning package? [For the purpose of this problem, you may assume that the package contains an effectively infinite-supply of M&Ms so that for a normal package, the probability of a green M&M is 1/6 regardless of how many M&Ms you have already drawn.]

   (b) Suppose you pull out a 6th green M&M. Now what is the probability you are a winner?

   (c) What about after a 7th green M&M? Are you excited yet?

   (d) How many green M&Ms must you pull out to be at least 99% sure you are a winner?

7. How many five letter strings start with 't', end with 'n', and have no two consecutive letters the same. Explain your computations in detail. (Feel free to write a MATLAB script that checks your answer, though.)

8. The file `dickens-corpus.txt` contains the concatenation of the text of Charles Dickens' ten greatest novels[1] with all the punctuation, numerals, and special symbols removed, and all the capital letters converted to lower case. Only 27 characters appear: a–z and a single space. Even the newlines have been removed; all 2.2 million words are on a single line.

   We are going to use this large string of text, along with our MATLAB programming skills, to build a simple probability model for the English language.

   Reading from the text file is easy. The following code opens the file, and reads it into an array of type `char`:

[1] In order: A Tale of Two Cities, Great Expectations, David Copperfield, Oliver Twist, A Christmas Carol, Bleak House, The Pickwick Papers, Hard Times, Little Dorrit, and Nicholas Nickleby.

```
fid = fopen('dickens-corpus.txt');
A = fscanf(fid, '%c');
fclose(fid);
```

Despite the size of the file, that code should only take a few seconds to run.

As you may or may not know, ASCII code uses the decimal numbers 97–122 for the lower-case letters a–z. The whitespace is coded as 32. The following code replaces the entries in `A` with 1–26 for a–z and 27 for '_'.

```
iletters = find((A >= 97) & (A <= 122));
ispace = find(A == 32);
A(iletters) = A(iletters) - 96;
A(ispace) = 27;
```

(a) Using the code above and some of your own, generate a plot of the relative frequencies of each of the characters using the `stem` command. Turn in a printout of your code and the plot. What percentage of non-whitespace letters is 'e'? What about 'q'?

(b) Have MATLAB print out a list of the letters a–z in order of their relative frequencies, starting with the most frequent (so your list should start with 'e'). You will probably find the `sort` command useful. Also, if you create a character array of letter labels using

```
labels = [char(97:122)];
```

you can very easily index it in arbitrary order:

```
>> labels([2 15 1 20])

ans =

boat
```

Turn in your code and a printout of your list of numbers.

(c) Create a 27×27 matrix, call it H, and then populate it with conditional probabilities estimated by parsing the corpus such that

$$H(\gamma, \eta) = \mathrm{P}\left(\text{current letter is } \gamma \mid \text{previous letter was } \eta\right).$$

So $H(5, 1)$ is the probability that an 'e' follows an 'a', $H(26, 20)$ is the probability that a 'z' follows a 't', etc. What is the most likely letter to follow a 't'? What about an 'a'? What about '_' (i.e., what is the most likely letter to start a word)?

(d) Create a three stem plots which chart the probabilities of letters that follow a 'q', 'r', and 's'. Turn in these plots.