**Homework 5:**

Please submit your answers in a single text document, and include a plaintext version of your code (.py, etc). Jupyter notebooks are hard to read without opening. Also, please don’t compress your files into a .zip (which requires us to download them instead of grading them online). This homework is due Sunday evening before midnight. If you turn it in early we will be more able to grade it and return comments prior to the midterm.

Answer all of the questions below.

**Q1.** Suppose you have a HMM with two states: **Variable**, in which the states are random, and **Repeating**, which produces sections of ‘AAA…’ repeats. The transition and emission probability tables are given below. For boundary conditions, assume initial P(*π1 = V) =* 1/2, and g(n) = 1 for both hidden states (that is, the transition from all k to the "End" state has a probability 1 at position 5*)*

For the following string, X, calculate the marginal probability of the third hidden state being ‘Repeating’. That is, *P{π3 = R | X}*?

X = ATAAA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Emission Probabilities: | A | C | G | T |
| Variable | .25 | .25 | .25 | .25 |
| Repeating | .91 | .03 | .03 | .03 |

|  |  |  |
| --- | --- | --- |
|  | Variable | Repeating |
| Variable | .75 | .25 |
| Repeating | .10 | .90 |

You may do this by hand or implement it in code. If you do it by hand, show your work. If you do it in code, don’t use any pre-packaged functions (you may however use your own code from prior weeks.) Regardless, provide the output probability as part of your write-up.

**Q2.** We are interested in knowing the expression of some gene, X, in a specific cell type. In 10 samples, we sequenced the transcriptome for 1000 reads each, and got the following counts for gene X.

24 33 42 30 44 38 27 39 47 51

We want to model expression of gene X as a binomial distribution with some probability p, and n=1000, in order to estimate the true expression probability of gene X.

Use each method below to estimate the parameter, p, the probability of seeing gene X

1. **Method of Moments\***
2. **Least Squares**
3. **Maximum Likelihood**
4. **Bayesian Method: Maximum a Posteriori (MAP) estimate**

**\***Only use the first moment, incorporating the second moment (variance) into your estimate of p is beyond the scope of the class (and it fits the data poorly)

**Note: these aren’t options, you should do each of these methods, a-d.** For each, you may do this analytically or with numerical estimation in the programming language of your choice (or on paper if you prefer). If you do this analytically, show all your work.

As the parameter space for this problem is very small, there is no need for any pre-packaged functions for optimization, you may however use functions to calculate the probability from a normal distribution (and you may also use standard functions, such as max(), np.where() and/or np.argmax(), in order to find the maximum value.) Graphing is not required, but is certainly encouraged, as it will help you appreciate what is happening in these methods. Naturally, you may use any packages you like to plot the data). For the Maximum Likelihood, especially if you do this analytically, remember that you can maximize the log likelihood.

**Q3.** We have two DNA strings:

S1 = AGCTA

S2 = ACCGA

Compute the probability of these two strings after k=2 generations, given an ancestor with equal probability at all bases and the Jukes-Cantor transition matrix provided below. You should assume that each position is independent.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | C | G | T |
| A | .85 | .05 | .05 | .05 |
| C | .05 | .85 | .05 | .05 |
| G | .05 | .05 | .85 | .05 |
| T | .05 | .05 | .05 | .85 |

Hint: This is the same as taking the joint probability at each position, conditioned on the ancestral probabilities. Taking the second position as an example, we want

P(G and C | a) \* P(a) + P(G and C | c) \* P(c) +  P(G and C | g) \* P(g) + P(G and C | t) \* P(t)

where the lower case letters represent the ancestral states at that position.

**Something** **to keep you up at night this Halloween:** The real cell type for Q2 (technically, a python script) produces counts based on a rounded normal distribution with a mean of 40 and sigma = 10), which is a rather different model from the one we assumed. *Spooky.*