CS 369 Assignment 2 2019

Due 6:00 pm Wednesday April 24

Marked out of 28. 7.5% of the final grade.

Write your code in Python 3 and present your report with embedded code as a Jupyter Notebook.

In your report, include explanations of what you are doing and why you are doing it. Write in whole sentences and present your results clearly.

Submit your Jupyter notebook in the raw .ipynb form and as a .html file to Canvas by 6:00 pm on the due date.

1. *[4.5 marks total]* Use the `numpy.random` library to simulate random samples from the distributions below. When plotting histograms, make sure the axes are labeled appropriately and you don't choose far too many or far too few bins (something like $\sqrt{n}$ bins for n points is a good rule of thumb).

   For each problem i-iv below:

   a. *[0.5 marks each]* state the value of the parameters of the distribution that are appropriate for the problem;

   b. *[0.5 marks each]* make a histogram of the sampled values; and

   c. *[0.5 marks each]* compare the mean and variance of the sampled values to the theoretical mean and variance for the distribution with the given parameters.

   i. On average, 1 bird visits a particular tree every 10 mins to forage. Use the Poisson distribution to simulate the number of birds arriving at the tree every hour for each of 100 hours.

   ii. Assuming the average life span of a giant tortoise is about 100 years, use the exponential distribution to simulate the number of years each of 200 giant tortoises live.

   iii. A certain species of reptile lays exactly 30 eggs per year but on average 85% of the eggs are predated before hatching. Use the binomial distribution to simulate the number of offspring that each of 200 of these reptiles successfully hatch in a year.

2. *[5 marks total]* In this short question, you will write methods for generating samples from the exponential and Poisson distributions.

   (a) *[1.5 marks]* Using the inversion method, write function `rand_exp` that draws samples from the exponential distribution. It should take two parameters: `rate` which is the rate paramter of the exponential distribution (called λ in the notes) and `size` which is a positive integer number of samples to return. It should return an ndarray of length `size`.

(b) *[2.5 marks]* Using your `rand_exp` function from part (a), write a function `rand_poiss` that draws samples from the Poisson distribution. It should take two parameters: `rate` which is the rate parameter of the Poisson distribution (called λ in the notes) and `size` which is an integer number of samples to return. It should return an ndarray of length `size`.

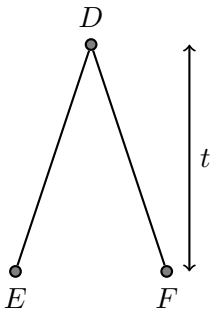
(c) *[1 mark]* Provide evidence that your implementations of `rand_exp` and `rand_poiss` are correct by making a histogram of 10000 samples from each function with the rate parameter set at $\lambda = 2$ and comparing them to histograms of samples drawn using analogous functions in the `numpy.random` library.

3. *[7.5 marks total]* The Jukes-Cantor model of DNA sequence evolution is simple: each site mutates at rate μ and when a mutation occurs, a new base is chosen uniformly at random from the four possible bases, $\{A, C, G, T\}$. If we ignore mutations from base X to base X, the mutation rate is $\frac{3}{4}\mu$. All sites mutate independently of each other.

   Thus we observe mutations at a site after an exponentially distributed waiting time with rate $\frac{3}{4}\mu$. At a mutation, choose from the 3 possible bases to mutate to with equal probability.

   A sequence that has evolved over time according to the Jukes-Cantor model has each base equally likely to occur at each site.

   Your programs should write sequences consisting of $\{A, C, G, T\}$, though it may be easier internally to translate the bases to integers, $\{1, 2, 3, 4\}$ for example.



Sequence D is the most recent common ancestor of sequences E and F.
The time since E split from F is t time units.

   (a) *[4.5 marks]* Write a method that simulates pairs of sequences that have diverged from a recent common ancestor t time units ago. Assume that evolution has occurred according to the Jukes Cantor model. The distribution for the sequence of the most recent common ancestor is uniform over the four possible bases at each site. The method should take sequence length, time t and mutation rate μ as inputs. It should return the ancestral sequence (D in the figure) and the descendant sequences (E and F in the figure). You may use methods `choice, exponential` and `poisson` from the `numpy.random` library.

   Simulate a pair of sequences of length 50 with $\mu = 0.01$ and $t = 10$. Print the resulting sequences along with the ancestral sequence. Report the number of sites at which each sequence differs from the ancestral sequence and from its sibling sequence (i.e., the number of sites difference between D and E, D and F and E and F.

   (b) *[3 marks]* Explain why you would expect the number of mutations that occur on a tree to be Poisson distributed with parameter $2tL\frac{3}{4}\mu$, where L is the

sequence length. Simulate 1000 pairs of sibling sequences of length 1000 with $\mu = 0.01$ and $t = 25$. For each simulated pair, count the number of sites at which they differ from each other. Report the mean and variance of the number of differing sites. Is this number Poisson distributed with parameter $2tL\frac{3}{4}\mu$? Explain why or why not.

4. *[10 marks total]* Suppose we wish to estimate basic secondary structure in protein (amino acid) sequences. The model we consider is a simplistic rendition of the model discussed in S C. Schmidler et al. (2004) Bayesian Segmentation of Protein Secondary Structure, doi:10.1089/10665270050081496

   We assume that at each point of the sequence, the residue is associated with one of three secondary structures: α-helix, β-strand and loops which we label H, S and T, respectively. To simplify the problem, we classify the amino acids as either hydrophobic, hydrophilic or neutral (B, I or N, respectively) so a sequence can be represented by this 3-letter alphabet.

   In a α-helix, the residues are 35% hydrophobic, 55% hydrophilic and 10% neutral. In a β-strand, the respective proportions are 55%, 15%, 30% and in a loop they are 10%, 10%, 80%.

   Assume that all secondary structures have geometrically distributed length with α-helices having mean 20 residues, β-strands having a mean of 12 residues and loops a mean of 10 residues. A β-strand is followed by an α-helix 40% of the time and a loop 60% of the time. An α-helix is followed by a β-strand 20% of the time and a loop 80% of the time and a loop is equally likely to be followed by a strand or a helix. At the start of a sequence, any structure is equally likely.

   (a) *[2 marks]* Derive the transition probabilities of a state to itself (e.g., a_{HH}) by considering that if L is geometrically distributed with parameter p then $E[L] = 1/p$. Make sure you use the parametrisation of the geometric distribution that takes values in $\{1, 2, \ldots\}$ and remember that $\sum_l a_{kl} = 1$ for any state k.

   (b) *[2 marks]* Sketch a diagram of the HMM (a hand-drawn and scanned picture is fine). In your diagram, show only state nodes and transitions. Show the emission probabilities using a separate table.

   (c) *[3 marks]* Write a method to simulate state and symbol sequences of arbitrary length from the HMM. Your method should take sequence length, a and e as arguments. Simulate and print out a state and symbol sequence of length 150.

   (d) *[3 marks]* Write a method to calculate the natural logarithm of the joint probability $P(x, \pi)$. Your method should take x, π, a and e as arguments.

   Use your method to calculate $P(x, \pi)$ for π and x given below and for the sequence you simulated in Q4c.

   $\pi =$ `S,S,S,S,T,T,S,S,S,S,S,S,S,H,H,H,H,H,H,H,H,H,H,H`

   $x =$ `N,N,I,B,N,B,N,B,B,B,N,N,B,B,B,B,I,I,I,I,I,B,N,B,I`