

CS/INFO 3300

Homework 4

Due before class Monday 3/10

Going through your files you discover your notes from a birding trip, during which you saw a Canada goose, a bufflehead, a northern pintail, a red-tailed hawk, a pileated woodpecker, and a mallard. Unfortunately, you forgot to write down where you were! It has to be either Stewart Park (by Cayuga Lake) or Sapsucker Woods, but you can't remember which. Can we guess which location you saw these six birds?

Here's what you know. You go to Sapsucker Woods three times as often as you go to Stewart Park. Consulting eBird.org, you find the number of recent reported sightings of each bird at the two locations. At Stewart Park these are CG: 1000, BH: 30, NP: 75, RTH: 40, PW: 15, M: 1000. At Sapsucker woods the same numbers are CG: 400, BH: 1, NP: 25, RTH: 20; PW: 15, M: 350. In these problems you will use these numbers (your *training* data) to estimate the probability of these species at the two locations. (Ignore the fact that there are ~230 other species you might see.)

1. What is the probability that you were at Sapsucker Woods, NOT considering any bird sightings? (that is, calculate the prior probability that you went to Sapsucker Woods) (5 pts)
2. For each location and bird species, what is the probability that, if you select one of the sightings from that location at random, it will be an instance of that species? (12 pts)
3. Assume you are at Sapsucker Woods, and that bird sightings are independent. What is the probability of seeing a red-tailed hawk *and* a pileated woodpecker, using the estimated probabilities from the previous question? (that is, the conditional probability of the two bird sightings given your location) (8 pts)
4. What is the probability of seeing a red-tailed hawk and a pileated woodpecker *and* being at Sapsucker Woods? (that is, the joint probability of the sightings and the location) (5 pts)
5. Write a javascript function "logProb" that takes two arguments, a location and an array of bird species (for example ["CG", "BH", "NP"]), and returns the *log* of the probability of those events using the probabilities you estimated in the previous

questions. The function should support passing an empty array [] of sightings. (16 pts)

6. Now let's consider your mystery birding trip. Compute the probability of seeing those six species assuming you were at Sapsucker Woods and then again assuming you were at Stewart Park. Use the function you created in Q5 to calculate, for both locations, the log probabilities of seven observation arrays, starting with the empty array and adding one bird species each time: [], ["CG"], ["CG", "BH"], ..., ["CG", "BH", "NP", "RTH", "PW", "M"]. (14 pts, you may combine your answer with the answer for the next problem.)

7. Write a JSON array with one element for each of the observation arrays in the previous problem. Each element should be an object with three properties: the name of the bird species added (use "Prior" for the first element), and the log probability values you calculated for the two sites that you calculated in the previous problem. (5 pts)

8. Create a line plot that shows how each observed species changes the difference between the log probabilities of the sequence of species at the two locations. Make a d3 ordinal axis with the sequence of species ("Prior", "CG", "B", ..., "M") on the x-axis. Make a scale for the y-axis centered at 0 for the difference in log probabilities. Use d3 "line" to show the log probability of the observations so far assuming you were at Sapsucker Woods *minus* the log probability of the same observations assuming you were at Stewart Park. (35 pts)

9. Which site is more likely to have been the location of your trip? Which observation(s) has/have the most effect on the final difference in log probabilities? Which have the least effect? Does the prior affect your final guess? (5 pts)