homework 04: a plague of sand mice

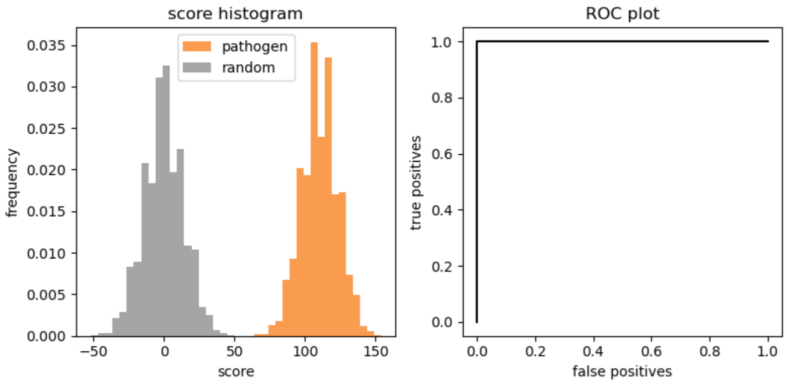
A strange new disease is sweeping through the sand mouse population. Your lab is trying to obtain genomic and transcriptomic sequence of the pathogen, by sequencing RNA from infected mice. Pathogen sequences make up a small fraction of the RNA-seq reads you obtain (1%), mixed with a large majority of sand mouse sequences (99%). You and Moriarty are both working on developing a method for classifying individual RNA-seq reads as "pathogen" versus "sand mouse", based on the read's sequence composition alone.

The lab has collected datasets that are available to you for training and testing methods:

- pathogen.fa: 10,000 pathogen sequences of length 200, in FASTA format.
- sandmouse.fa: 10,000 sand mouse sequences, also of length 200 and in FASTA format.

Moriarty's method

Moriarty is already crowing about his secret proprietary machine learning method. He shows the following (apparently spectacular) result:



The histogram on the left shows results of scoring the 10,000 pathogen sequences in pathogen.fa, compared to scoring 10,000 random sequences of uniform base composition (which are here if you want them). The ROC plot on the right shows that his method achieves perfect discrimination... Moriarty says.

1. test Moriarty's method yourself

After some prying on your part, Moriarty reluctantly discloses that his secret proprietary machine learning method is to score +1 for each A or T, and -1 for each C or G.

You note that Moriarty showed results for discriminating against *random* sequences (of uniform base composition), but the problem is to discriminate against *sand mouse* sequences.

Implement Moriarty's secret proprietary machine learning method. Test it on the sequences in pathogen.fa and sandmouse.fa. Make a plot of your own, showing a histogram and a ROC plot like in Moriarty's figure, but for performance in discriminating pathogen from sand mouse sequences, using matplotlib.

What do you conclude? Why are your results different from Moriarty's?

2. make your own method

You think you can do better, using higher order Markov models.

- Implement a **second order Markov model** (i.e. with parameters $P(x_i | x_{i-1}, x_{i-2})$.)
- Train (i.e. estimate parameters for) two models: one on pathogen.fa, and one on sandmouse.fa.

- Implement a routine that calculate the log-odds score for a sequence read, given the read sequence and the two 2nd-order Markov models.

- Plot the histogram and ROC plot for your 2nd-order Markov model method, for results of discriminating the reads in pathogen.fa from sandmouse.fa.

Because you're estimating the parameters of your models from the example data, you should train and test on separate data sets. Use half of each sequence set for training, and half for testing.

3. how good is your method?

Suppose the lab needs to achieve 90% sensitivity (i.e. detecting 90% of pathogen sequences).

- What score threshold would you need to set?

- What is your estimate for the false positive rate at that threshold?

Remember that in an RNA-seq sample from an infected sand mouse, 1% of the reads are from the pathogen, and 99% of the reads are from the mouse. Using your estimates above, if you use your method to classify reads, what proportion of the reads that you label as "pathogen" are actually false positives from the sand mouse? (This is your *false discovery rate*, FDR.)

(Use Python code to answer these questions, using the results of your testing above.)

turning in your work

Submit your Jupyter notebook page (a `.ipynb` file) to the course Canvas page under the Assignments tab. Please name your file `<LastName><FirstName>_<psetnumber>.ipynb`; for example, mine would be `EddySean_04.ipynb`.

Remember that we grade your psets as if they're lab reports. The quality of your text explanations of what you're doing (and why) are just as important as getting your code to work. We want to see you discuss both your analysis code, and your biological interpretations.