BioE 145 Assignment 1 Due Tuesday 2/2 11:59 PM PST

Note: Turn in your submission to this assignment in Gradescope by Tuesday February 2nd, 11:59 PM PST. Attach a PDF printout of your completed IPython notebook from lab as an appendix, as well as any code used to find your answers to the following questions.

IPython notebook: Google Colab Assignment 1 [Solutions]

Part 1: Linear Algebra Review

1. Rank and Eigenvectors

   (a) Determine the eigenvalues and eigenvectors for the following matrices.

$$\begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix}, \begin{bmatrix} 6 & 9 \\ 4 & 6 \end{bmatrix} \tag{1}$$

   For the first matrix:

$$\lambda_1 = 10, \quad v_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix} \qquad \lambda_2 = 2, \quad v_2 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \tag{2}$$

   For the second matrix:

$$\lambda_1 = 12, \quad v_1 = \begin{bmatrix} 0.83205 \\ -0.83205 \end{bmatrix} \qquad \lambda_2 = 0, \quad v_2 = \begin{bmatrix} 0.5547 \\ 0.5547 \end{bmatrix} \tag{3}$$

   (b) For both of the above, determine the rank and dimension of the null space.

   First matrix: rank = 2, dim(N) = 0
   Second matrix: rank = 1, dim(N) = 1

2. Show that a symmetric matrix A, with dimension n by n, can be expressed as

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T \tag{4}$$

   where v_i is a vector in $\mathbb{R}^n$ and λ_i is a scalar in $\mathbb{R}$. What do they represent in relation to A?

   Hint: Use the following expansion for an n by n diagonal matrix D.

$$D = \sum_{i=1}^{n} D_{i,i} e_i e_i^T \tag{5}$$

where $\mathrm{e_i}$ is the $\mathrm{i^{th}}$ vector in the standard basis of $\mathbb{R}^n$.

A symmetric matrix can be written as

$$A = QDQ^T \tag{6}$$

$$A = Q\left(\sum_{i=1}^{n} D_{i,i}e_ie_i^T\right)Q^T \tag{7}$$

$$A = \sum_{i=1}^{n} D_{i,i}Qe_ie_i^TQ^T \tag{8}$$

$$A = \sum_{i=1}^{n} D_{i,i}Q_iQ_i^T \tag{9}$$

$$A = \sum_{i=1}^{n} \lambda_iv_iv_i^T \tag{10}$$

where $\mathrm{Q_i}$ is the $\mathrm{i^{th}}$ column of Q.

λ_i is an eigenvalue of A and v_i is the corresponding eigenvector of A.

BioE 145 Assignment 1 Due Tuesday 2/2 11:59 PM PST

Part 2: Probability Review

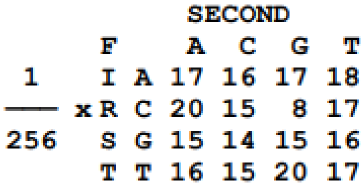
1. You are analyzing the nucleotide composition of an animal genome and you notice a modest AT bias. The genome is 23/50 AT (46%) and 23/50 GC (54%). You then start to look at the frequencies of dinucleotides.

   (a) If nucleotides are independent, what is the probability of 5'-CpA-3'dinucleotides? What is the probability of 5'-TpC-3'dinucleotides?

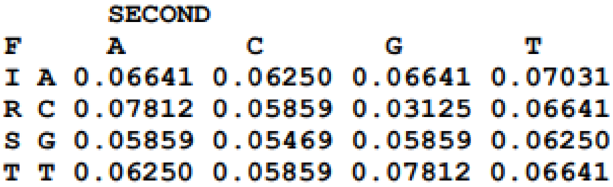
$$\text{P(CpA)} = \text{P(C)} \times \text{P(A)} = \frac{27}{100} \times \frac{23}{100} = 0.0621 \tag{1}$$

$$\text{P(TpC)} = \text{P(C)} \times \text{P(A)} = \frac{23}{100} \times \frac{27}{100} = 0.0621 \tag{2}$$

   (b) You find the dinucleotide frequencies don't match the values calculated for independent nucleotides. For instance, the probability of 5'-CpA-3'dinucleotides is 20/256 (7.813%) whereas the probability of 5'-TpC-3'dinucleotides is 15/256 (5.860%). The dinucleotide frequencies overall are given below, in fractions of 256 and as decimal numbers:



```
             SECOND
        F     A  C  G  T
 1      I  A 17 16 17 18
 ---  x R  C 20 15  8 17
256     S  G 15 14 15 16
        T  T 16 15 20 17
```



```
            SECOND
F      A         C         G         T
I A 0.06641 0.06250 0.06641 0.07031
R C 0.07812 0.05859 0.03125 0.06641
S G 0.05859 0.05469 0.05859 0.06250
T T 0.06250 0.05859 0.07812 0.06641
```

      i. Verify the single-nucleotide probabilities of A and C in two different ways.

      Examine A and C as both the first and the second nucleotide.

$$\text{P(A)} = \sum_{i \in A,C,G,T} P(A,i) \approx \frac{27}{100} \tag{3}$$

$$\text{P(A)} = \sum_{i \in A,C,G,T} P(i,A) \approx \frac{27}{100} \tag{4}$$

$$\text{P(C)} = \sum_{i \in A,C,G,T} P(C,i) \approx \frac{23}{100} \tag{5}$$

$$\text{P(C)} = \sum_{i \in A,C,G,T} P(i,C) \approx \frac{23}{100} \tag{6}$$

      ii. What are the probabilities of each nucleotide following a T?

For $i \in A, C, G, T:$

$$\text{P}(i \text{ after T}) = \frac{\text{P(T, i)}}{\text{P(T)}} \tag{7}$$

$$\text{P(A after T)} = \frac{16}{256} \times \frac{100}{27} = 0.23148 \tag{8}$$

$$\text{P(C after T)} = \frac{15}{256} \times \frac{100}{27} = 0.21701 \tag{9}$$

$$\text{P(G after T)} = \frac{20}{256} \times \frac{100}{27} = 0.28935 \tag{10}$$

$$\text{P(T after T)} = \frac{17}{256} \times \frac{100}{27} = 0.24595 \tag{11}$$

iii. What are the probabilities of each nucleotide preceding a G?

For $i \in A, C, G, T:$

$$\text{P}(i \text{ before G}) = \frac{\text{P(i, G)}}{\text{P(G)}} \tag{12}$$

$$\text{P(A after G)} = \frac{17}{256} \times \frac{100}{23} = 0.28872 \tag{13}$$

$$\text{P(C after G)} = \frac{8}{256} \times \frac{100}{23} = 0.13587 \tag{14}$$

$$\text{P(G after G)} = \frac{15}{256} \times \frac{100}{23} = 0.25476 \tag{15}$$

$$\text{P(T after G)} = \frac{20}{256} \times \frac{100}{23} = 0.33967 \tag{16}$$

iv. What biological phenomenon could explain the dinucleotide patterns you see?

- CpG dinucleotides are less common, while TpG and CpA dinucleotides are more common
- Can arise from cytosine methylation at CpG sites
- Repair of cytosine to uracil deamination is less efficient with 5-methylcytosine, which convert to thymine
- Produces C to T and G to A changes at CpG sites

2. A bag contains two dice. One is a fair die that rolls 1 through 6 with equal probability. The other is a weighted die that has a one-third chance of rolling a 6, and never rolls a 1. You reach in to the bag, pick one of the two dice (either with equal probability), and roll it.

(a) Write a table of the joint probabilities of picking the fair or the weighted die, and rolling each number.

Joint	1	2	3	4	5	6
Fair	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
Weighted	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$

(b) Compute the conditional probabilities $P(six|fair), P(six|weighted)$.

$$\mathrm{P}(six\,|\,\text{fair}) = \frac{\mathrm{P}(six\,\&\,\text{fair})}{\mathrm{P}(\text{fair})} = \frac{1}{12} \times \frac{2}{1} = \frac{1}{6} \tag{17}$$

$$\mathrm{P}(six\,|\,\text{weighted}) = \frac{\mathrm{P}(six\,\&\,\text{weighted})}{\mathrm{P}(\text{weighted})} = \frac{1}{6} \times \frac{2}{1} = \frac{1}{3} \tag{18}$$

(c) Compute the conditional probabilities $P(fair|six), P(weighted|six)$.

$$\mathrm{P}(\text{fair}\,|\,six) = \frac{\mathrm{P}(six\,\&\,\text{fair})}{\mathrm{P}(six)} = \frac{1}{12} \times \frac{12}{3} = \frac{1}{3} \tag{19}$$

$$\mathrm{P}(\text{weighted}\,|\,six) = \frac{\mathrm{P}(six\,\&\,\text{weighted})}{\mathrm{P}(six)} = \frac{1}{6} \times \frac{12}{3} = \frac{2}{3} \tag{20}$$

Part 3: Maximum Likelihood Estimator

1. The Maximum Likelihood Estimator (MLE) finds the model, or set of parameters, that maximizes the probability of the data. In other words, it maximizes the likelihood of some model θ given the data we obtain and seek to fit a model to. Defining the likelihood as

$$\mathcal{L}(\theta; \mathcal{D}) = p(\text{data} = \mathcal{D} \,|\, \text{true model} = h_\theta) \tag{1}$$

then the MLE is

$$\hat{\theta}_{MLE} = \arg\max \mathcal{L}(\theta; \mathcal{D}) \tag{2}$$

(a) Given data $\mathcal{D}$ that is a set of outputs $y_1, \ldots, y_n$ arising from inputs $x_1, \ldots, x_n$, write out the MLE above as a probability of the y_is and x_is.

Note: Each output y_i is conditioned on the input x_i (as well as the model)

$$\hat{\theta}_{MLE} = \arg\max p(\text{data} = \mathcal{D} \,|\, \text{true model} = h_\theta) \tag{3}$$

$$\hat{\theta}_{MLE} = \arg\max p(y_1,\, y_2,\, ...,\, y_n \,|\, x_1,\, x_2,\, ...,\, x_n,\, h_\theta) \tag{4}$$

(b) The data is related via $y_i = h_\theta(\mathrm{x}_i) + Z_i$ where $h_\theta(\mathrm{x}_i)$ is fixed and Z_i is i.i.d Gaussian, see (3). What is the conditional probability of y_i conditioned on x_i and θ?

$$Z_i \sim \mathcal{N}(0,\, \sigma^2) \tag{5}$$

$$y_i \sim \mathcal{N}(h_\theta(\mathrm{x}_i),\, \sigma^2) \tag{6}$$

(c) Using the answers to (a) and (b), show the MLE estimate $\hat{\theta}_{MLE}$ can be written as follows using the log-likelihood.

$$\hat{\theta}_{MLE} = \arg\min \sum_{i=1}^{n} (y_i - h_\theta(\mathrm{x}_i))^2 \tag{7}$$

$$\hat{\theta}_{MLE} = \arg\max\, p(y_1,\, y_2,\, ...,\, y_n \,|\, x_1,\, x_2,\, ...,\, x_n,\, h_\theta) \tag{8}$$

$$\hat{\theta}_{MLE} = \arg\max \prod_{i=1}^{n} p(y_i \,|\, x_i,\, h_\theta) \tag{9}$$

$$\hat{\theta}_{MLE} = \arg\max \;\; -n\log\sqrt{2\pi}\sigma - \left(\sum_{i=1}^{n} \frac{(y_i - h_\theta(x_i))^2}{2\sigma^2}\right) \tag{10}$$

$$\hat{\theta}_{MLE} = \arg\max \;\; -\left(\sum_{i=1}^{n} \frac{(y_i - h_\theta(x_i))^2}{2\sigma^2}\right) \tag{11}$$

$$\hat{\theta}_{MLE} = \arg\min \sum_{i=1}^{n} \frac{(y_i - h_\theta(x_i))^2}{2\sigma^2} \tag{12}$$

$$\hat{\theta}_{MLE} = \arg\min \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 \tag{13}$$

Hints:

i. As logs are monotonic functions, taking the log of (2) allows us to find the same optimizer θ

ii. The probabilities of all y_i can be treated as independent, and can be expanded as a product, e.g. $p(y_1,\, y_2) = p(y_1)\cdot\, (y_2)$

iii. Use the answer from (b) with the formula for a normal distribution

iv. For an optimization problem, constants (added or, if positive, multiplied) don't affect the optimization and can be dropped or ignored

v. The arg max of a term is equivalent to the arg min of the negated term, i.e. arg max x = arg min $-x$