BioE 145/245 Assignment 1 Due Monday 1/30 11:59 PM

Note: Turn in your submission to this assignment in Gradescope by Monday January 30, 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

Part 1: Linear Algebra Review

1. Rank and Eigenvectors

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

$$\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}, \begin{bmatrix} 4 & 8 \\ 2 & 4 \end{bmatrix} \tag{1}$$

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

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

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T \tag{2}$$

   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.

$$D = \sum_{i=1}^{n} D_{i,i} e_i e_i^T \tag{3}$$

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

Part 2: Probability Review

1. 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.

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

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

2. You are developing a test for a disease that affects 0.5% of the population. The test is fairly accurate and will detect an infected person 98% of the time. However, it has a false positive rate of 1%.

   (a) What is the probability that someone has the disease, given that they tested positive?

   People showing symptoms of the disease are more likely to have it. Assume an infected person has a 90% chance of having a cough, while an uninfected person has a 10% chance of having a cough.

   (b) In the overall population, are coughing and testing positive independent variables?

(c) Given that someone is infected, are coughing and testing positive independent variables?

(d) What is the probability that someone is infected, coughing, AND tested positive?

(e) What is the probability that someone is coughing AND tested positive?
(Hint: There are two cases here - the person either is infected or isn't.)

(f) Bob is coughing and tested positive. What is the probability that Bob is infected?

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)

   (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?

$$Z_i \sim \mathcal{N}(0,\, \sigma^2) \tag{3}$$

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

$$\hat{\theta}_{MLE} = \arg\min \sum_{i=1}^{n} (y_i - h_\theta(\mathrm{x}_i))^2 \tag{4}$$

   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. $\arg\max x = \arg\min\, -x$