Stats 230, Homework 3

Due date: February 28

1. Consider a problem of fitting a logistic regression model to data.

   (a) Write a function that can fit logistic regression to data using two optimization methods:

   - Gradient descent/steepest ascent method
   - Newton-Raphson/Fisher scoring/IRLS method

   Your function should return MLEs of regression coefficients, their corresponding asymptotic confidence intervals, and a vector of the log-likelihoods "visited" by the chosen optimization option.

   (b) Apply your function to the heart-disease data: `https://hastie.su.domains/ElemStatLearn/datasets/SAheart.data` (description is here: `https://hastie.su.domains/ElemStatLearn/datasets/SAheart.info.txt`).

   (c) Compare convergence of the two optimization options by plotting iterations vs. log-likelihood values and benchmark run times of the two optimization algorithms.

2. Implement the EM algorithm For the ABO blood type example discussed in class and apply it to the the data

$$\mathbf{n} = (n_A, n_{AB}, n_B, n_O) = (6, 4, 55, 35).$$

   Package your implementation into a function that takes data and initial values of the allele frequency probabilities p_A, p_B, and p_O as inputs. Document and add this function to your class package on github.

3. Consider the occasional dishonest casino hidden Markov model (HMM) with hidden states 1=fair die, 2=loaded die, transition probabilities

$$\mathbf{P} = \begin{pmatrix} 0.98 & 0.02 \\ 0.05 & 0.95 \end{pmatrix},$$

   emission probabilities

$$\mathbf{E} = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{2} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \end{pmatrix},$$

   and initial distribution $\boldsymbol{\nu}^T = (\frac{1}{2}, \frac{1}{2})$.

   (a) Simulate $(\mathbf{x}_{1:100}, \mathbf{y}_{1:100})$ from this HMM and plot the hidden and observed states as time series.

(b) Implement the forward and backward and algorithms for the occasional dishonest casino example. Run these algorithms on the observed states generated in part (a), compute marginal probabilities of hidden states at each time and plot them together with the true simulated states.

(c) Pretend you don't know the initial distribution, transition probabilities, and emission probabilities corresponding to the loaded die. In other words, you only know the first row of the emission probability matrix. Implement the Baum-Welch algorithm to estimate all unknown parameters and apply it to data from part (a). Report estimated parameters.