STA613/CBB540 HOMEWORK 4

DUE THURSDAY, FEB 19^{TH}

Your written answers should be as succinct as possible.

(1) *Sparse regression.* I have a collection of individuals with various levels of heart disease; this trait is encoded in variable y_i for the ith individual. For each individual, I also have information about their levels of LDL and HDL cholesterol, heart rate, systolic and diastolic blood pressure, and fasting glucose level, all encoded in a vector x_i for individual i. I would like to use the markers x to predict the heart disease levels y, but I do not think that every marker contributes to the model.
   (a) Consider a simple Lasso model: write out the definition of the model with the penalty term.
   (b) How would you estimate the parameters in this model? What parameters are being estimated?
   (c) How can I find the generalization error of a fitted model?
   (d) If I change my heart disease phenotype to a boolean scalar value (i.e., 0 for no heart disease, 1 for heart disease), should I change the model? If so, how?
   (e) If I think that all of the markers may be important for prediction, but I still want to use penalized regression, what model can I use? What assumptions are (implicitly) in this model regarding the regression coefficients?

(2) *Linear mixed models.*
   (a) What is a random effect? How is it different from a fixed effect?
   (b) In the [Segura et al.] paper reading for this topic, what is the (non-noise) random effect modeled in the LMM? How do they model the random effect?
   (c) Write the definitions (in terms of probabilities or ratios) of statistical power and FDR.
   (d) Write three sentences interpreting Figure 2 in this paper, paying special attention to power, FDR, the Lasso model (LM) and single-locus mixed model (MM) performance, and their multi-locus mixed model (MLMM) performance.

(3) *Mixture models: K-means and Expectation Maximization (EM).* In class, we discussed a Gaussian mixture model and associated K-means algorithm for estimating 'centroids', and the EM algorithm for estimating the mixture model parameters.
   (a) Write a program in R that performs K-means clustering.
   (b) Derive the M-step updates for μ_i, Σ_i, and π_i for EM in a mixture model (consider the expected complete log likelihood). Write another program in R that performs EM using the E-step we discussed in class.

(c) Download the data from the website and run these methods on the data. Let the number of clusters $K = 2$. How sensitive are the solutions to the starting points (try a number of them). Plot one set of results from K-means and one from EM (color the points by the hard or soft assignments – for soft assignments, use the most likely cluster).

(d) Try the same thing again with $K = 3$. Plot one set of results from each method.

(e) After playing around with the methods, what are the main practical differences between the two?