CSEP 590A: Computational Biology **Assignment #3**
due: Thursday, July 27

Due date is set for 7/27, but lecture five will make more sense if you have time to start before 7/20, and I will likely have some more to hand out this week. Turn this one in on paper; handwritten is fine, I don't recommend trying to typeset it. Extra credit is for extra practice and glory; it is not a big component of your grade.

1. **Bayes Rule:** In a certain population, an obese person has a 30 percent chance of having high blood pressure and a non-obese person has a 10 percent chance of having high blood pressure. Twenty percent of the population is obese. What is the conditional probability that a person is obese, given that the person has high blood pressure?

2. **Maximum Likelihood:** Let $x_1, x_2, \ldots, x_n$ be n samples of a normal random variable X with mean θ_1 and variance θ_2. In class I showed that the maximum likelihood estimates of θ_1 and θ_2 when both are unknown give a biased estimate of θ_2. What is the MLE of $\theta_2 = \sigma^2$ if $\theta_1 = \mu$ is assumed to be known? Extra Credit: Is it biased, i.e., does the expected value of $\hat{\theta}_2$ differ from θ_2?

3. **EM:** In class, I sketched the EM algorithm for the two-component Gaussian Mixture Model only in the special case when both subpopulations were assumed to share the same variance and the mixing proportions were assumed to be 50/50. Carry out the analysis for the general case where σ_1^2, σ_2^2 and $0 \leq \tau_1 \leq 1$ ($\tau_2 = 1 - \tau_1$) are arbitrary.

4. **Maximum Likelihood:** Suppose X is a discrete random variable with three possible outcomes, say A_1, A_2 and A_3. Let $\theta = (p_1, p_2, p_3)$ be the probabilities of outcomes A_1, A_2, A_3, resp., (where $p_1 + p_2 + p_3 = 1$, of course). Suppose you have collected n independent random samples $x_1, x_2, \ldots, x_n$ drawn from this distribution. Using the same basic approach as in the coin-flipping example in the class notes (Lec 4, slide 9), show that the maximum likelihood estimators for the parameters θ are $\hat{\theta} = (n_1/n, n_2/n, n_3/n)$, where n_i is the number of occurrences of outcome A_i among $x_1, x_2, \ldots, x_n$. Hint: The algebra is mildly easier if you happen to remember Lagrange multipliers, but it's certainly not essential. (FYI, this result generalizes to arbitrary multinomial distributions, not just 2 or 3 outcomes; see the slick proof in Chapter 11.)

5. **EM:** Recall that an *allele* of a gene is one variant of its DNA or protein sequence. Individuals generally carry two (possibly identical) alleles of each gene, one inherited from mother, one from father (genes on the X/Y chromosomes being exceptions). The ABO blood type gene has three common alleles in the human population: A, B and O. The blood type of an individual depends as follows on the pair of alleles that he or she has: type A if the pair is A/A or A/O; type B if the pair is B/B or B/O; type AB if the pair is A/B; type O if the pair is O/O. Let $p(A)$ be the fraction of A alleles in the population, $p(B)$, the fraction of B alleles and $p(O)$, the fraction of O alleles. These fractions are nonnegative and sum to 1. Under the standard assumption in genetics of independent assortment, the probability that an individual has a given pair of alleles is the same as the probability of obtaining that pair in two random draws from the set of all alleles in the population: for example, the probability of the pair A/B is $2p(A)p(B)$. In a sample of 20 individuals, 9 have blood type A, 2 have blood type B, 1 has blood type AB and 8 have blood type O. Derive the appropriate formulas needed to use the EM algorithm to determine the values of $p(A)$, $p(B)$ and $p(O)$ most likely to have given rise

to this data. Then run the algorithm for a few iterations on the given data. Try it with a couple of very different starting estimates for the parameters. You may write a program to do the iteration, do it by hand, or give a spreadsheet with the relevant formulas and "fill down" a few rows to iterate. If you use a spreadsheet, turn in a printout of the formulas as well as the numbers; I think CONTROL-backquote causes Excel to show all formulas. Hint: The parameters are $p(A), p(B)$ and $p(O)$, the observed data are the blood types of the individuals and the hidden data are the pairs of alleles possessed by the individuals. The solution to problem 4 will help. Depending on how you set up the likelihood function, you might (or might not) need the multinomial distribution from pg 300 of the text.

(If you'd like info on the genetics of the ABO blood group system, the 1930 Nobel prize in Physiology or Medicine, have a look at Wikipedia `http://en.wikipedia.org/wiki/Abo_blood_group` or OMIM `http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=110300`. In a nutshell, they are 3 alleles of a single gene on the ninth chromosome (9q34), which encodes a *glycosyltransferase* - an enzyme that modifies the carbohydrate content of the red blood cell antigens. The A and B alleles perform slightly (but immunologically significantly) different modifications; the O allele has a 1 base deletion, hence an altered reading frame, producing a very different protein with no apparent function at all, a so-called "null" allele. Aside from issues with blood transfusions, people with O blood type are apparently more susceptible to cholera. And, no, the "independent assortment" assumption for this gene is *not* well justified in the human population; prevalence is strongly dependent on geography. But we'll ignore that for this problem...)

Extra Credit Problems:

6. **Maximum Likelihood:** Suppose X is a random variable uniformly distributed between 0 and $\theta > 0$ for some unknown θ. Based on a sample $x_1, x_2, \ldots, x_n$ of X, what is the maximum likelihood estimator of θ? Is it biased?

7. **EM:** Generalize the EM algorithm from problem 3 to allow a fixed but arbitrary number $k \geq 1$ of components in the mixture, preferably allowing a choice of either a common variance σ^2 shared by all clusters, or a separate variance per cluster. Implement it and experiment with simulated data to see how well it recovers the parameters you used to generate the data. How quickly does the iteration converge? Does it ever seem to be converging to a local, not global, max? How well does it work with sparse data? Well-separated clusters? Highly overlapping clusters?