CSE 527: Computational Biology HW #2 due: Monday, Nov. 16

Turn this one in on paper; handwritten is fine. (Typeset is also fine, but all the math may be a nuisance.) If you can't make it to class on the due date, you can mail it to me, or scan your paper and upload it to `https://catalysttools.washington.edu/collectit/dropbox/ruzzo/7491`.

Do any 6 problems below. Numbers 5 and/or 7 are especially recommended (but not the easiest). If desired, do some of the others for extra credit (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:** Suppose you scan a segment of genomic DNA triplet by (adjacent, non-overlapping) triplet, looking at whether it might code for a protein. Of course, some of those triplets will match amino acid codons and others will match stop codons. In other words, you can partition the segment into so-called "open reading frames" or ORFs, where one ORF consists of a maximal length (≥ 0) run of non-stop-triplets followed by one stop triplet. Assuming the segment is actually non-coding, you might expect the lengths of ORFs to be *geometrically* distributed, i.e., there is a $0 \leq p \leq 1$ such that for any $k > 0$, the probability of an ORF containing $k-1$ consecutive non-stop triplets followed by one stop is $(1-p)^{k-1} \cdot p$. (I.e., successive triplets are independently stops with probability p. I am assuming the scan starts with the very first triplet of the sequence, and ignoring the issue of looking at 1- and 2-nucleotide offsets from that, and/or the other strand.) Suppose such a scan encounters ORFs of length $x_1, x_2, \ldots, x_n$. What is the maximum likelihood estimator of p, assuming the geometric model above?

3. **Hypothesis Testing:** Continuing the framework of the previous problem, suppose a scan of another region of the same genome finds an ORF of length k ($k-1$ non-stop triplets followed by a stop triplet). Outline a simple statistical test, based just on this data, indicating whether or not that ORF is likely to be a real protein coding gene. Evaluate your test assuming the MLE for p was $3/64$ and $k = 100$.

4. **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 derived the maximum likelihood estimates of θ_1 and θ_2 when both are unknown. What is the MLE of the variance $\theta_2 = \sigma^2$ if $\theta_1 = \mu$ is assumed to be known?

   For example, suppose I draw a sample of 3 measuring 9, 10, 11. The sample mean is 10, sample standard deviation is $\sqrt{((9-10)^2 + (10-10)^2 + (11-10)^2)/3} = \sqrt{2/3} \approx .8$. This is not a surprising result if the population mean and varaince are ≈ 10 and ≈ 1, resp. But suppose I told you the sample was drawn from a population with mean 0. Drawing a sample of 9, 10, 11 is now much less likely, but is certainly possible, and is made more probable by increasing our estimate of the population variance (the *sample* variance is unchanged; it's defined in terms of the sample mean, not the population mean). So the question is: now that I know the population mean, what estimate of population variance make the data I just observed most likely? Give me the general formula, as well as the value for these specific observations.

Extra Credit: From the in-class discussion, when θ_1 and θ_2 are both unknown, the MLE gives a biased estimate of θ_2 (but not θ_1). In the case outlined in this problem, is your estimate of θ_2 biased, i.e., does the expected value of $\hat{\theta}_2$ differ from the actual θ_2 of the population from which you have sampled?

5. **EM:** In class, I sketched an EM algorithm for learning the parameters of a two-component Gaussian Mixture Model only in the special case when both subpopulations were assumed to share the same variance and the mixing proportions (τ_1/τ_2) were assumed to be 50/50. Carry out the analysis for the case where either σ_1^2, and σ_2^2 or $0 \leq \tau_1 \leq 1$ ($\tau_2 = 1 - \tau_1$) are arbitrary values to be estimated from the data along with μ_1, μ_2. Give a succinct (e.g., 1 or 2 sentence) English description of the result. Extra credit: Do the general case where all six parameters $\mu_j, \tau_j, \sigma_j^2, j = 1, 2$ are simultaneously estimated from the same data.

6. **Maximum Likelihood:** (a) 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 (Lecture 4, *circa* slide 13), 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 three variables are coupled, since $p_3 = 1 - p_1 - p_2$, so substitute for p_3 using this identity before you differentiate. (b) Generalize this to a variable with 4 possible outcomes. (You don't need to prove it, but FYI, it also generalizes to arbitrary multinomial distributions; see the slick proof in Chapter 11. Also, the algebra is mildly easier if you happen to remember Lagrange multipliers from your calculus class, but ignore this otherwise, it's not a big deal.) (c) Suppose you observe N nucleotide sequences sampled from the distribution defined by a fixed but unknown weight matrix of width k. Show that the MLE for the matrix parameters is just the set of nucleotide frequencies observed in each position.

7. **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. [If you want the jargon, O is *recessive*, A & B are *dominant* over O, and *co-dominant* with respect to each other. Your *genotype* is the pair of alleles you carry; your *phenotype* is the outward manifestation thereof. If you have two copies of the same allele, you are *homozygous* at that locus; *heterozygous* if you have two different alleles.]

   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 simplifying assumption that the population is in "Hardy-Weinberg equilibrium," 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, the first 9 have blood type A, the next 2 have blood type B, the next has blood type AB and the last 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, if possibe, please 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 phenotypes (blood types) of the individuals and the hidden data are the genotypes of (pairs of alleles possessed by) the individuals. I suggest you formulate the problem using 20 zero/one variables z_i to represent the hidden data, where $z_i = 1$ means individual i is homozygous; $z_i = 0$ means heterozygous. (Actually, z_i is unknown only for the first 11 individuals.) The solution to problem 6 may help; it is an analogous 3 parameter problem without hidden data. Depending on how you set up the likelihood function, you might (or more likely might not) need the multinomial distribution from pg 300 of the text.

(If you'd like to know more about the genetics of the ABO blood group system, the 1930 Nobel prize in Physiology or Medicine, 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) that 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, more or less explaining why the O allele is recessive or "silent." Aside from issues with blood transfusions, people with O blood type are apparently more susceptible to cholera. And, no, the "Hardy-Weinberg" 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...)

8. **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? If so, what is the unbiased estimator?

9. **EM:** Generalize the EM algorithm from problem 5 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. You might also generalize to multi-dimensional data (but note that you either need to make simplifying assumptions about the covariance structure of the data, or to estimate many more parameters). Implement it (or find code on the web; tell me where) 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? (This problem is very open-ended; you don't need to address all the questions above. Do what you can in a reasonable amount of time and tell me about it.)

10. **EM:** The method in the previous problem in the case of a common variance σ^2 shared by all clusters (and no covariance in the multidimensional case) is closely related to "K-means clustering." Lay out the similarities and differences. Perhaps write or find code for both and compare them on some simple simulated data sets.

11. **Parsimony:** Empirically, it has been observed that evolving nucleotide sequences are more likely to undergo transitions (substitution of one purine for the other, i.e. A↔G, or one pyrimidine for another, i.e., C↔T) than transversions (substition of a pyrimidine for a purine or vice versa). Redo the parsimony calculation shown in my lecture slides (Lecture 10, 11/2/09, approximately slides 11-15) assuming that the nucleotides at the leaves are TCAGA (in left-to-right order) and assuming that transitions have a cost of 1, whereas transversions cost 2. (Please show enough of your intermediate work so that I can see that you understand the method.)