

**Computational Genomics: Assignment No. 3**  
**due on January 27th, 2003**

**General Guidelines:**

This assignment is part of your final grade in the course. It should be done *independently*, either individually or in *pairs*, without any help from others. Duplicated and copied works will be given zero grade. Using articles or books is perfectly acceptable as long as you include the reference in your relevant answer.

**Credit:** Every one of the first four questions is worth 20 points. The RH question is worth 30 points. Solve all items for 110 points + one golden star. Starred (\*) section is harder, but you can accumulate the 100% grade even without solving it.

**Concluding remarks:** We enjoyed the course and hope you did too...

**Hidden Markov Models**

1. (**Warm up**) We are given a  $k$ -sided die, and we toss it  $n$  times. Outcomes of different tosses are independent. We are told that out of  $n$  tosses,  $n_1$  have result 1,  $n_2$  have result 2, ...,  $n_k$  have result  $k$ . What are the maximal likelihood estimates of the die probabilities  $p_1, p_2, \dots, p_k$ ? ( $p_i$  is the probability of getting result  $i$ , and clearly  $p_1 + p_2 + \dots + p_k = 1$ )? Prove your answer.
2. Prove that The EM (Baum-Welch) algorithms for HMM parameter estimation is monotonically improving. Namely, show that using the expected transition and emission counts as the new set of HMM parameter we cannot decrease the likelihood of data  $x$  given the model  $\theta : P(x|\theta)$ .
3. HMMs can be used for alternative representation of alignment problems, for example for pairwise alignment with affine gap penalties. The idea is to form a finite state automaton, with states indicating existence of a gap in either sequence, or a match between the sequences. An edge between two states is weighted such that the cost of path through the automaton equals the score of the alignment implicitly defined by it. The automaton can be further represented as an HMM, with costs (of state-transition or of output emission) replaced by probabilities (see also Durbin, chapter 4).
  - (a) Write the formulae presenting alignment costs as log-odds.

- (b) Generalize the gap alignment HMM of the above to an HMM corresponding to an alignment of three sequences,  $S_1, S_2, S_3$  with an affine gap penalty  $a_i l_i + b_i$  for a gap of length  $l_i$  in  $S_i$ . Draw the HMM.
- (c) Calculate the probability of three given sequences according to this model (summed over all alignments). In other words, calculate the likelihood of the the three sequences.

### Supervised Learning and Support Vector Machines (SVMs).

4 An  $n - 1$  dimensional hyperplane in  $R^n$  is specified by a *normal vector*  $\mathbf{w} \in R^n$  and a “shift”  $b \in R$ . A point  $\mathbf{z} \in R^n$  is *on the hyperplane* iff  $\mathbf{z} \cdot \mathbf{w} + b = 0$ , where  $\cdot$  denotes the inner product  $\mathbf{z} \cdot \mathbf{w} = \sum_{i=1}^n z_i w_i$ . The corresponding *decision function*  $f(\mathbf{x})$  is defined for points *not* on the hyperplane. It determines on which side of the hyperplane a point resides. Formally  $f(\mathbf{x}) = \text{sign}(\mathbf{x} \cdot \mathbf{w} + b)$ .

Suppose our “training set” contains  $m$  labeled points  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ , where each point  $\mathbf{x}_i$  is in  $R^n$ , and each label  $y_i \in \{-1, 1\}$  (indicating if a certain property of interest holds or not). Furthermore, assume the training set is *linearly separable*, namely there is an  $n - 1$  dimensional hyperplane that separates the “yes” instances ( $y_i = 1$ ) from the “no” instances ( $y_i = -1$ ).

Given a labelled training set  $S \subseteq R^n \times \{-1, 1\}$  that is linearly separable, find the “best” separating hyperplane. Suppose  $\mathbf{w}, b$  specifies a separating hyperplane for  $S$ . By rescaling  $\mathbf{w}$  and  $b$  we can guarantee that the closest point(s) to the hyperplane,  $\mathbf{x}$ , satisfies  $y_i (\mathbf{x} \cdot \mathbf{w} + b) = 1$ . In this case, the *margin* (the distance of the closest point in  $S$  to the hyperplane) equals  $1/\|\mathbf{w}\|$ .

The simplest version of SVM is the “hard margin” or “maximal margin” optimization problem. Given a linearly separable labeled set  $S$ , find the separating hyperplane with largest margin.

- (a) For the simple set in 2D  $S^- = \{(0, 0), (0, 2), (2, 0)\}$ ,  $S^+ = \{(5, 5), (0, 4), (4, 0)\}$  find the separating hyperplane with largest margin. Explicitly specify  $\mathbf{w}, b$ , and the resulting margin.
- (b) Why might this “hard margin” version of SVM be inappropriate for real data? Describe qualitatively an example that *is* linearly separable, yet there is a good reason to take a hyperplane that does not even separate the “yes” and “no” instances of  $S$ .
- (c) (\*) For a set  $S$  whose points are in 2D ( $\mathbf{x}_i \in R^2$ ), write explicitly the hard margin optimization problem. Do you see a simple way to solve it, which avoids the

general machinery of convex quadratic optimization, typically employed in SVM software?

### Radiation Hybrid Mapping:

5 The first step in ordering markers using data gathered from radiation hybrids experiments is to estimate the breakage probability between every pair of markers. Let  $a$  and  $b$  be two markers, and let  $\theta$  denote their breakage probability. That is,  $\theta$  is the probability that at least one radiation induced breakpoint occurs between  $a$  and  $b$ . In this question you will compute two different estimations for  $\theta$ .

Let  $n^{--}$  denote the number of hybrids that contain both markers, and define  $n^{-+}, n^{+-}, n^{++}$  similarly. In class we showed that the probability of getting the outcome  $\langle n^{--}, n^{-+}, n^{+-}, n^{++} \rangle$  given that the breakage probability is  $\theta$  is,

$$[q(1 - p\theta)]^{n^{--}} [pq\theta]^{n^{-+} + n^{+-}} [p(1 - q\theta)]^{n^{++}} \quad (1)$$

where  $p$  is the retention probability, and  $q = 1 - p$ .

- a. Use the maximum likelihood method to find  $\hat{\theta}$ , the value of  $\theta$  that maximizes (1).
- b. Assume that the values of  $n^{--}, n^{-+}, n^{+-}, n^{++}$  equal exactly their expected values (e.g.,  $n^{-+} = mpq\theta$ ).

Compute the value of your estimator,  $\hat{\theta}$ , under this assumption.

Another (simpler) approach for estimating the value of  $\theta$  is the following. Denote by  $n^=$  the number of hybrids in which either both markers appear or both are absent (i.e.,  $n^= = n^{--} + n^{++}$ ). Similarly, denote by  $n^{\neq}$  the number of hybrids in which exactly one of the markers appears (i.e.,  $n^{\neq} = n^{-+} + n^{+-}$ ).

- c. Compute the probability for the  $\langle n^=, n^{\neq} \rangle$  given that the breakage probability is  $\theta$ .
- d. Use the maximum likelihood method to find  $\bar{\theta}$ , the value of  $\theta$  that maximizes the expression you found in c. Notice that  $\hat{\theta} \neq \bar{\theta}$ .
- e. As in b, assume now that the values of  $\langle n^=, n^{\neq} \rangle$  equal exactly their expected values. Compute the value of  $\bar{\theta}$  under this assumption.

### 6 And the (timely) golden star question is:

- (a) In every language your course staff is familiar with, a "second" has the same two meanings: The ordinal number following the first one, as well as a short time unit. Any reason for this universal linguistic phenomena?

- (b) Why are there 60 minutes in an hour? (Hint: the answer is in your fingertips).
- (c) Why are there 24 hours in a day?
- (d) Why can we move in either direction along the  $X, Y, Z$  axis, but along the time axis we can only move forward?