6.874 Computational Systems Biology

Mid-term exam

March 20, 2008

(2 points) Your name and MIT ID:

[Blank space for name and MIT ID]

THIS IS AN OPEN BOOK, CLOSED COMPUTER EXAM.

NOTE Questions marked with "G – graduate only" are graduate level questions. The undergraduate version has a total 40 points, and the graduate version includes 14 extra points. If you are registered for 6.874 you should answer all of the questions on the exam. If you are registered for 6.807 or 7.90, your answers to the graduate questions will be graded and used for extra credit in the course.
1.1 (10 points) We were faced with the problem of clustering $n$ gene profiles $x_i, i = 1, \ldots, n$, where each profile consisted of $m$ measurements (expression values). Following the lectures, we decided to try Gaussian mixture models for this task, written as

$$P(x|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j N(x_i; \mu_j, \sigma^2 I)$$

The parameters $\theta$ include the mixing proportions $p_1, \ldots, p_k$, the mean profiles $\mu_1, \ldots, \mu_k$, as well as the overall variance $\sigma^2$. We use EM to estimate the mixture model twice, once with $k = 2$ components and the other time with $k = 3$ components. The resulting parameter values were $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, respectively. In both cases the resulting clusters looked fine so we were unclear about how to select between the two models. Can you give an equation that would permit us to compare which of the two models best describes the available expression data? Please provide the equation in the terms of this problem.

1.2 (6 points) (G – graduate only) Provide two reasons why we might not end up selecting the right model on the basis of the equation you provided above.
2.1 **(10 points)** By clustering expression data with our favorite clustering algorithm (affinity propagation) we identified a set of 22 interesting genes. We compared this set to genes bound by a transcription factor relevant in our biological context and found that 20 of our genes were bound by the factor. This seemed like a significant outcome since the factor binds only 30 of the complete set of 6000 genes in yeast. Under the null hypothesis that the 30 genes bound by the factor are drawn uniformly at random from the total set of genes, write down an expression for the probability that our result or even greater overlap would be due to chance alone.

2.2 **(4 points) (G – graduate only)** Would the result be more or less significant in a human context? In other words, if 20 of our 22 genes were bound by a factor that binds only 30 out of the total 25,000 or so human genes?
3.1 (10 points) You expose a population of yeast cells to heat shock and at ten minute intervals you perform expression analysis of 10 genes for a total of 200 minutes. The resulting expression data (starting at an observation at minute 0) forms a $10 \times 21$ matrix $X$. We are interested in finding the dominant vector direction (a $10 \times 1$ column vector) in which the expression of the 10 genes change over time. How do we extract this from the data matrix $X$? Express your answer in the terms of this problem.

3.2 (4 points) (G – graduate only) How do we identify the dominant vector direction in which the time course profiles change (a $21 \times 1$ column vector) across the genes?
4.1 (10 points) Topic models can be used to decompose expression data to identify how genes are activated across conditions. The caricature data matrix below shows how a set of 6 genes are active (1) or inactive (0) under three different experimental conditions.

We will use \( d^t \), \( t = 1, 2, 3 \), to represent the data from each experiment, and \( \hat{n}_g(d^t) \) for the number of expression events for gene \( g \) in condition \( t \). In the example below, \( \hat{n}_g(d^t) \) is either 0 or 1.

The probability of generating the data matrix from the topic model is then

\[
P(d^1, d^2, d^3|\theta) = \prod_{t=1}^{3} \prod_{g=1}^{6} \left( \sum_{z=1}^{m} \theta_t^g \theta_g^z \right) \hat{n}_g(d^t)
\]

<table>
<thead>
<tr>
<th>genes</th>
<th>conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 0</td>
<td>1 1 0</td>
</tr>
<tr>
<td>1 1 0</td>
<td>1 1 0</td>
</tr>
<tr>
<td>0 1 1</td>
<td>0 1 1</td>
</tr>
<tr>
<td>0 1 1</td>
<td>0 0 0</td>
</tr>
</tbody>
</table>