Bioinformatics – Lecture Notes

Announcements

Reference: Microarray Data Analysis and Visualization by Arun Jagota

5 days of classes including today.

Final Projects due May 2 (2 weeks from today).

Class 26 – April 18, 2002

I. Cluster Analysis Applied to Microarray Data

Recall that microarray data can be thought of as gene expression patterns or sample expression patterns. These can be each considered to be vectors. The first thing we have to do before applying cluster analysis is to find a distance between the various expression pattern vectors. This is done using similarity/dissimilarity measures such as Euclidean distance, Mahalonobis distance, or linear correlation coefficients. Once a distance matrix is computed, the following clustering algorithms can be used. The clusters formed can differ significantly depending upon the distance measure used.

A. Hierarchical Clustering

This is the class of clustering methods we talked about previously ie. UPGMA, WPGMA, etc.

B. k-Means Clustering

An alternate method of clustering called k-means clustering, partitions the data into k clusters and finds cluster means $\mathbf{m}_i$ for each cluster. In our case, the means will be vectors also. Usually, the number of clusters k is fixed in advance. To choose k something must be know about the data. There might be a range of possible k values. To decide which is best, optimization of a quantity that maximizes cluster tightness ie. minimizes distances between points in a cluster. A possible measure is given after the method description.

The method is as follows:

1) Pick the initial means $\mathbf{m}_1, \mathbf{m}_2, \ldots \mathbf{m}_k$. To do this it helps to know something about what the clusters might look like.
2) Assign each data vector $\mathbf{x}$ in the data set D to the cluster C_i that has a mean $\mathbf{m}_i$ that is closest to $\mathbf{x}$ using the similarity measure.
3) Re-calculate the mean of all the clusters C_i

4) Repeat steps 2) and 3) until sufficient convergence is reached ie, the means to not change much.

In order to choose k, we want to find k that minimizes the following measure of cluster tightness is

$$\sum_{i=1,k} \frac{1}{|C_i|} \sum_{x \in C_i} d(x, m)$$

C. Self-organizing Maps

This is basically an application of neural networks to microarray data. Assume that there is a 2-dimensional grid of cells and a map from a given set of expression data vectors in R^n, ie, there are n nodes in the input layer and a connection neuron from each of these to each cell. Each cell (i, j) gets it own weight from n input neurons. The weight vector μ_{ij} is the mean of the cluster associated with cell (i, j). Each data vector **d** gets mapped to the cell (i, j) that is closest to **d** using Euclidean distance.

In order to train the network, the mean vectors μ_{ij} for the cells (i, j) must be learned. This is done by the following procedure:

1) The vectors μ_{ij} are given initial values randomly.
2) For each data vector **d**, the following steps are performed:
    a. The cell (i, j) that is closest to **d** is found.
    b. The vector for every cell (i', j') that is close to (i, j) is updated by

$$m_{i',j'}(t+1) = m_{i',j'}(t) + \eta d(ij, i'j')(d - m_{i',j'}(t))$$

where t increases as the learning process procedes and $d(ij, i'j')$ is the distance between cell (i, j) and cell (i', j').

II. Hidden Markov Models and Microarray Data

We can use Hidden Markov models for pattern recognition in the study of micorarray data. Suppose that we want to consider gene expression data from a tissue sample and want to know if it is control or different from the control (diseased, experimentally altered, responding to drug, etc.). Consider the gene expression data vector as a set of emissions, one for each vector coordinate. Each emission has a value that is defined by some probability distribution function. This can be continuous, or can even discrete. To make it discrete, the data should be preprocessed to indicate, up-regulation, down-regulation, or no significant change.

Using the procedures from lecture earlier, a set of data can be used to fit a HMM with two states normal and not normal. Transition and emission probabilities can

be estimated. Then the posterior probabilities calculated to determine the state of each tissue sample.

III. Finding Genes Expressed Unusually Different in a Population

The following section addresses the question: Is gene g expressed unusually in the sample?

The first thing to do is to come up with a formal mathematical definition for what unusual is. Assume that the microarray data is log transformed ratio data. If a histogram is constructed of the data, it should yield roughly a normal distribution. Anything that is out near either tail can be considered to be unusually expressed. Note that this can be either a high or low expression level.

Calculate the Z-score for the data point considered

$$Z = \frac{e_g - m}{s}$$

where e_g is the expression level, μ is the mean and σ is the standard deviation. The Z value will give an indication of the how far the data is toward the tail (α - level).

IV. Finding Genes Expressed Significantly in a Population

In order to determine if a gene is significantly up- or down-regulated in a population relative to a control, statistical methods such as hypothesis testing have to be used.

Suppose that we think that gene x is associated with a cancer. Assume that this gene has been monitored in n tissues with cancer. Also, assume that the log ratios of the gene expression level of gene x compared to normal tissue in the n tissue samples is given by $e_x^1, e_x^2, ...e_x^n$.

Now we formalize our question. Is gene x sufficiently up-regulated in the cancerous tissues in this group with respect to the control for use to be able to infer that it is up-regulated relative to the control in the population of all tissues with this cancer C?

One way of approaching this problem is the calculate the average of the e_x^i's, that is $\bar{e}_x = \frac{1}{n}\sum_i e_x^i$. We could conclude that if this is greater than 0 we have up-regulation. However, is this enough about 0.

Let us re-phrase the question: Is $\bar{e}_x$ sufficiently above 0 to conclude that gene x is up-regulated in the cancerous tissue relative to the control?

A. Using Statistical Inference

1. Normal Distribution

To do this we use *statistical inference*. Assume that the size of the group n > 30. Then we can invoke the *central limit theorem* to assume that the group mean $\bar{e}_x$ is normally distributed with mean equal to the population mean μ_x and the standard deviation is $\sigma_x/\sqrt{n}$, where σ_x is the population standard deviation. We can then calculate the Z-statistic

$$Z = \frac{\bar{e}_x - \mu_x}{\sqrt{\sigma^2/n}}$$

The question can now be put in terms of statistics: Is Z sufficiently greater than 0 to conclude that gene x is up-regulated in the cancerous population relative to control. Remember that Z ~ N(0, 1).

We formulate the *null hypothesis* H_0: $\mu_x = 0$ that means the opposite of what we want to find, that is, it states that the mean is not greater than 0. This makes the *alternative hypothesis* H_A: $\mu_x > 0$. Our Z statistic is

$$Z = \frac{\bar{e}_x}{\sqrt{\sigma^2/n}}$$

If $Z > z_\alpha$, then we reject the null hypothesis in favor of the alternative with confidence $1-\alpha$. Otherwise, we fail to reject the null hypothesis.

2. Student's T Test

What happens if n < 30 and we do not know the population standard deviation? In this case, we use the student's t test.

$$t = \frac{\bar{e}_x - \mu_x}{s_x/\sqrt{n}}$$

where s_x is the sample standard deviation.

We proceed as before with the same null hypothesis and alternative and get

$$t = \frac{\bar{e}_x}{s_x / \sqrt{n}}$$

We then go to a student's t table and see if $t > t_{\alpha, n-1}$. If this is true, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

3. Non-parametric Tests

The student's t test assumes that the data is normally distributed. If we do not want to make that assumption, we can use a non-parametric test such as the *sign test*.

Let $\{e_{x,i}\}$, i = 1, 2, …, n be the set of the expression levels of gene x in individual i in a group of size n drawn from the population. We want to test if m_x, the median expression level of the gene in the population is zero or not. Thus, the H_0: $m_x = 0$ and H_A: $m_x \neq 0$. First, we drop all i: $e_{x,i} = 0$. Let n be the size of the remaining group. Let $n_+ = \{i: e_{x,i} > 0\}$ and $n_- = \{i: e_{x,i} < 0\}$. Under H_0, the probability is ½ that $e_{x,i} > 0$ and ½ that $e_{x,i} < 0$ because the sample group is drawn randomly from the population.

For any confidence level α, we calculate the largest k such that

$$\Pr[k \leq j \leq n-k \ positives] = \frac{1}{2^n} \sum_{j=k}^{n-k} \binom{n}{j} \geq \alpha$$

If $n_+ \leq k$ or $n_- \geq n-k$, then we reject H_0 with confidence level α.

Similarly if H_0: $m_x = 0$ (≥ 0) and H_A: $m_x < 0$

$$\Pr[k \leq j \ positives] = \frac{1}{2^n} \sum_{j=k}^{n} \binom{n}{j} \geq \alpha$$

If $n_+ \leq k$, then we reject H_0 with confidence level α.

4. Is any gene up-regulated in the population?

Suppose that we have a set of genes S from a micorarray experiment. S can be the whole set or a partial set. We want to know if at least one gene in the set S is up-regulated. This is done by using a *compound test*, with H_0: $\wedge_{x \in S} (m_x = 0)$ and H_A: $\vee_{x \in S} (m_x > 0)$. To do this, we need to do |S| many simple tests, one for each gene x in S. If one of the tests rejects the null hypothesis, we reject H_0.

There is one caveat. When we do the simple tests at level α, and reject H_0, we cannot say this with confidence $1-\alpha$. If we assume that the tests are

independent, there are two possible correction schemes. The *Bonferroni correction* uses

$$\alpha_S = \frac{\alpha}{|S|}$$

This is a very simple correction. A more precise correction is the *Sidak correction*

$$P_{corrected}(\text{simple test}) = 1 - (1 - P(\text{simple test}))^{|S|}$$

If the tests are not independent, in either case, we have an overly conservative correction.

5. Ranking and Filtering Genes by How Upregulated they are.

The best way of doing this is by the P-values of their simple tests. This is because the P-value is instantly readable and gives the probability that the null hypothesis is rejected erroneously. Second, it is the natural measure from which to establish a cut-off to select top-ranking genes. The cut-off will be based on the amount of risk one is willing to take.

6. Estimating the Mean Expression Level of Gene x in the Population

Suppose that we want to estimate the mean expression level μ_x of a gene x in a population of tissues with cancer C from the observed expression levels of x in a group which is a small subset of the population. One estimate is the group mean $\bar{e}_x$. This however will have some error associated with it. It is helpful to put some error bounds.

If we have a group with $n \geq 30$, we can assume that the group is approximately normally distributed with mean μ_x and standard deviation $\sigma_x / \sqrt{n}$. Then a 1- α confidence interval for μ_x is

$$\bar{e}_x \pm z_{\alpha/2} \sigma_x / \sqrt{n}$$

If we have a group with $n < 30$, we use the student's t test which yields a 1- α confidence interval for μ_x is

$$\bar{e}_x \pm t_{\alpha/2, n-1} s_x / \sqrt{n}$$