Bioinformatics – Lecture Notes

Announcements

I need to collect project proposals, copies of the articles, and group meeting times.

Reference: Microarray Data Analysis and Visualization by Arun Jagota

Class 25 – April 16, 2002

I. Measuring Dissimilarity of Expression Data

We might want to compare two or more gene or sample expression patterns. This might be used to differentiate between diseased and normal cells or finding out the genetic similarity of tissues. To do this we need a *distance metric* or a *dissimilarity measure*.

Definition – A function is $d : R^n \rightarrow R$ is a *distance metric* if

1) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$ for all $\mathbf{x}$, $\mathbf{y} \in R^n$
2) $d(\mathbf{x},\mathbf{y}) \geq 0$ for all $\mathbf{x}$, $\mathbf{y} \in R^n$ with $d(\mathbf{x},\mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$
3) $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$ for all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in R^n$

Definition – A function $f(\mathbf{x},\mathbf{y})$ is a *dissimilarity measure* if $f(\mathbf{x},\mathbf{y}) > f(\mathbf{w},\mathbf{z})$ iff **x** is less similar to **y** than is **w** to **z**.

A. Distance Metrics

1. Euclidean Distance

This is the most common distance measure.

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

This should not be used if either
1) Not all components of the vectors being compared have equal weight.
2) There is missing data.

Preprocessing the data can often alleviate these problems.

We can also use the normalized Euclidean distance

$$d(x,y) = \frac{\sqrt{\sum_i (x_i - y_i)^2}}{\sqrt{n}}$$

2. Minkowski Distance

   This is similar to the Euclidian distance except that 2 is replaced by p.

$$d(x,y) = \sqrt{\sum_i | x_i - y_i |^p}$$

   To see why is might be useful, consider the case when p=1. Then every coordinate *i* contributes to the Minkowski distance. However, if p is large, only the component *i* with the largest difference contributes to the distance.

   There is also a normalized Minkowski distance.

$$d(x,y) = \frac{\sqrt{\sum_i | x_i - y_i |^p}}{\sqrt{n}}$$

3. Mahalonobis Distance (weighted Euclidean distance)

   This modification of the Euclidean distance allows different coordinates to be weighted differently.

$$d(x,y) = \sqrt{(x-y)^T C^{-1} (x-y)}$$

   where C^{-1} is a diagonal matrix that has unequal entries on the diagonal to assign the weights. If C^{-1} is the identity matrix, we have Euclidean distance.

4. Taxi-cab Distance

   This is a special case of the Minkowski distance with p=1

$$d(x,y) = \sqrt{\sum_i | x_i - y_i |}$$

   One way of thinking about this is that the distances are like traveling around city blocks.

   There is also Normalized Taxi-cab distance

$$d(x,y) = \frac{\sqrt{\sum_i |x_i - y_i|}}{\sqrt{n}}$$

5. Canberra Metric

   Sometimes it is useful to have a metric that is defined on [0,1]

$$d(x,y) = \frac{1}{n}\sum_i \frac{|x_i - y_i|}{x_i + y_i}$$

6. Bray-Curtis Coefficient

   This is also defined on the interval [0,1]

$$d(x,y) = \frac{1}{n}\frac{\sum_i |x_i - y_i|}{\sum_i x_i + y_i}$$

B. Non-metric Dissimilarity Measures

1. Maximum Coordinate Difference

   The following computes the maximum absolute distance along a coordinate

$$d(x,y) = \max_i |x_i - y_i|$$

2. Minimum Coordinate Difference

   The following computes the maximum absolute distance along a coordinate

$$d(x,y) = \min_i |x_i - y_i|$$

3. Dot Product

   This is a dissimilarity version of the dot product.

$$d(x,y) = -x \circ y$$

4. Pearson's Linear Dissimilarity

This is a dissimilarity version of Pearson's linear correlation ρ between two vectors. In this dissimilarity, $d(x,y) \in [0,1]$. Here 0 indicates a perfect similarity (positive correlation) and 1 indicates a maximum dissimilarity (negative correlation).

$$d(x,y) = \frac{1 - \boldsymbol{r}(x,y)}{2} \quad \text{with } \boldsymbol{r}(x,y) = \frac{(x - \bar{x}) \circ (y - \bar{y})}{\boldsymbol{s}_x \boldsymbol{s}_y}$$

Note that the linear correlation ρ is basically a normalized dot product by subtracting off the mean ($\bar{x}$ and $\bar{y}$) and dividing by the standard deviation $\boldsymbol{s}_x$ and $\boldsymbol{s}_y$.

The linear correlation (as well as the Euclidean distance), is a very frequently used method for comparing the expression pattern between two genes. It has some advantages over the Euclidean distance in that it can be rescaled easily ie. ρ(**x**,**y**)= ρ(c**x**,c**y**) for any positive constant c. This is not true for the Euclidean distance.

A slight modification yields Pearson's Absolute Value Dissimilarity

$$d(x,y) = 1 - |\,\boldsymbol{r}(x,y)\,|$$

In this score, 0 indicates that **x** and **y** are either maximum similarity or dissimilarity and 1 indicates **x** and **y** are uncorrelated.

5. Spearman Rank Dissimilarity

This is a dissimilarity version of Spearman rank correlation. The rank-vectors of **x** and **y** are given by r(**x**) and r(**y**), respectively. They basically list the place of each entry if the vector was listed in ascending order. If **x** = (1,5,3,8), r(**x**) = (1,3,2,4)

$$d(x,y) = \frac{1 - r_s(x,y)}{2} \quad \text{with} r_s(x,y) = \boldsymbol{r}(r(x), r(y))$$

This emphasized the order and ignores other details. If **x** and **y** are two gene expression patterns, this can be used to tell if the expression levels of these two genes both increase or decrease monotonically.

This give rise to the Spearman Absolute Value Dissimilarity

$$d(x,y) = 1 - |\,r_s(x,y)\,|$$

Here, d(**x**,**y**) = 0 if **x** and **y** have identical or maximally dissimilar rank vectors and d(**x**,**y**) = 1 when the rank vectors are random permutations of the order.

6. Coefficient of Shape Difference

$$d(x,y)=\sqrt{\frac{n}{n-1}(e(x,y)^2-q(x,y)^2}$$

where e(**x**,**y**) is the normalized Euclidean metric and

$$q(x,y)=\frac{1}{n}\left(\sum_i x_i-\sum_i y_i\right)$$

7. Cosine Dissimilarity

This is the dissimilarity version of the cosine dissimilarity measure. It assumes values on [0,1] and is equal to 0 when $\mathbf{x}=\pm\ \mathbf{y}$ and 1 when **x** and **y** are orthogonal.

$$d(x,y)=\frac{1-\cos(x,y)}{2}\text{ where }\cos(x,y)=\frac{x\circ y}{\sqrt{x\circ x}\sqrt{y\circ y}}$$

8. Weighted Dot Product

The weighted dot product is

$$xWy=\sum_i x_i w_i y_i$$

where **W** is the diagonal matrix that contains diagonal elements w_i the weight each of the components.

C. Measures for Binary Patterns

Sometimes we may want to discretize the data by preprocessing so that the data is converted into patterns of up and down regulation. Assume that the vectors **x** and **y** are binary expression patterns with position i assuming the value of +1 for up-regulation for the i^{th} component gene and –1 for the down-regulation of the i^{th} component gene.

1. Hamming Distance

If the data for **x** and **y** are tabulated in a table

$$T(x,y) = \begin{array}{c} +1 \\ -1 \\ \\ \end{array} \begin{array}{c} \left[\begin{array}{cc} \#i's : x_i = +1, y_i = +1 & \#i's : x_i = +1, y_i = -1 \\ \#i's : x_i = -1, y_i = +1 & \#i's : x_i = -1, y_i = -1 \end{array} \right] \\ \begin{array}{cc} +1 \qquad\qquad\qquad\qquad & -1 \end{array} \end{array}$$

The Hamming distance is given by

$$HD(x,y) = T_{+1,-1}(x,y) + T_{-1,+1}(x,y)$$

This measures gives a measure of when the up- and down-regulation patterns of **x** and **y** are dissimilar.

2. Dependence-based Dissimilarity Measures

Let X and Y be random variables on the set {+1,-1} and let P(X,Y) be their joint distribution. Assume that the gene expression pattern vectors **x** and **y** are created by randomly choosing values from P(X,Y). Then the table T(**x**,**y**) is a contingency table for the random sample. The degree of dependence between X and Y can be estimated by using a statistical hypothesis test. Let P be the probability of erroneously rejecting the null hypothesis of independence. Using this P-value of the dependence test, a similarity measure can be defined as

$$S(\mathbf{x},\mathbf{y}) = -\log P$$

The dependence based test will measure the degree of positive or negative correlation.

Example

$\mathbf{x}$ = (-1,-1,-1,+1,+1,+1) $\mathbf{y}$=(+1,+1,+1,-1,-1,-1)

By the Hamming measure **x** and **y** are maximally dissimilar. By the dependence based measure, **x** and **y** are maximally similar.

a. c^2-based measure

One possibility that can be used is a c^2-based measure. In this approach, a c^2-test for dependence between X and Y is done and P(c^2) is used at the P-value in the similarity measure. The P-value will be approximated by 1- a_c where

$$a_c = \arg\min_a \,(c^2(1,\mathbf{a}) \le c^2(x,y)$$

with $c^2(1,a)$ being the c^2 value from the distribution with 1 degree of freedom so that the area on the right of this value is α, and $c^2(x,y)$ is the c^2 value from the contingency table T(**x**,**y**) obtained as follows:

i) Under the hypothesis that X and Y are independent E_{XY} is calculated as

$$E_{XY} = \frac{E_x E_Y}{n}$$

where E_x is the sum of row x = +1 or –1 in the matrix T(**x**,**y**) and E_y is the the sum of column x = +1 or –1 in the matrix T(**x**,**y**)

ii) The c^2 value is computed as

$$c^2 = \sum_{\substack{x=+1,-1,\\ y=+1,=1}} \frac{(T_{xy} - E_{xy})^2}{E_{xy}}$$

Note that the c^2 test is an approximate test that is only valid when $E_{xy} \geq 5$.

b. Fisher Exact Test

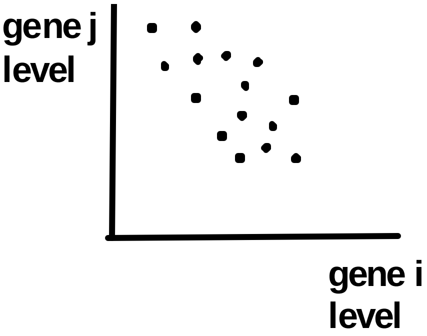
When the c^2 test cannot be used, the Fisher exact test should be used because it computes the exact probability of obtaining the counts in **T** under the hypothesis that X and Y are independent and with the row and column sums constrained by the values given in **T**.

$$\Pr[T \mid row\ and\ column\ sums] = \Pr[T_{+1,-1} \mid row\ and\ column\ sums]$$

$$= \frac{\binom{+1\ row\ sum}{T_{+1,-1}}\binom{-1\ row\ sum}{T_{+1,-1}}}{\binom{n}{+1\ column\ sum}}$$

II. Visualizing Micorarray Data

It is usually easiest to understand data if it can be represented in 2 or 3 dimensions. For example, a 2-D scatter plot of the expression levels of genes *i* and *j* over a number of samples can show the relationship between these two genes.

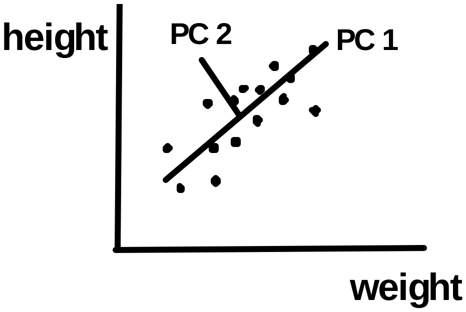


If the data is of higher dimensional, *principal components analysis* and *multidimensional scaling* can be used to help visualize the data.

A. Principal Components Analysis

In principal components analysis n-dimensional data is converted to d-dimensional data (d<<n) such that the components in the new space are uncorrelated and axis or dimensions of the new space are ordered with respect to the amount of variance they explain. The first component explains the most about the data. The second component is orthogonal to the first and explains more about the data and so on.

Example – Consider height and weight data for a group of individuals. This is 2-D data, but there is a correlation between height and weight. We can use this property to reduce the data to 1D. PC1 explains most of the data and PC2 explains the rest.



1. Implementation

Let D be the set of data points in R^n. The goal is to map these to a new d-dimensional space. Let $\bar{x}$ be the mean of the points in D and

$$\Sigma = \sum_{x \in D} (x - \bar{x})(x - \bar{x})^T$$

be the covariance matrix. The eigenvectors $\mathbf{v_i}$ of Σ (i = 1, 2, …n) are the basis vectors for the new space. The eigenvector with the largest eigenvalue is the first principal component, the second largest, the second principal component, etc.

A data point **x** in the original n-dimensional space can be mapped to the new d-dimensional space for $d \le n$ by the transformation

$$y_i = v_i^T x \quad \text{for i = 1, 2, … d}$$

which projects **x** onto the first d eigenvectors of Σ.

2. Notes:

   PCA changes the coordinate system so as to maximize the variance of the data along the coordinate axes
   It is a linear method. It uses the information from all coordinates.

   The similarities between the axes of the original coordinate axes are given in the correlation matrix.

3. Application to Microarray Analysis

   Sample Application 1 – If we want to compare the sample expression patterns from two groups (diseased vs normal, experimental vs control). If we have n genes, the each pattern is a point in n-dimensional space. Suppose we want to see if the sample expression patterns for these two groups cluster by group. We might want to perform PCA analysis and perform cluster analysis at the top three components.

   Sample Application 2 – On gene chips (such as the one made by Affymetrix), the same gene occupies multiple cells. In theory, the expression level of all cells with the same gene should be perfectly correlated. However, in practice, this is often not the case due to imperfections in the technology or hybridization of the sequence fragment to other genes in the target.

   PCA allows us to see how good the correlation among these cells is. To use PCA, we would hybridize k different samples on the same chip. For each sample, the expression levels of a gene x in the n cells is an n-dimensional vector. Hence, there are k points in n-dimensional space. Using PCA, if most of the variance is explained by the first principal component, the effective dimensionality of the data is 1 and there cells are highly correlated.

4. Limitations

Clustering by PCA effectively yields clusters as if the Euclidean distance metric had been used. Hence, it is possible that it might miss clusters.

The reduction of dimensionality uses all coordinates. If only a few genes out of a thousand differ between two samples (Application 1), clustering by PCA might not yield any meaningful results.

B. Multidimensional Scaling

Suppose that we have a set of data D of size m in an n-dimensional space and a m x m dissimilarity matrix **S** for all pairs of data based on a symmetric dissimilarity measure. Multidimensional scaling will try to map the points in D to a d-dimensional space where d<<n so that if two data items are dissimilar in n-space, they are dissimilar in d-space. Typically the Euclidean distance measure is used.

Example – Let $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$, be sample expression patterns in a high dimensional space. Suppose that we want to map these to a one-dimensional Euclidean space. Let d($\mathbf{s}_1$, $\mathbf{s}_2$) = 0.5, d($\mathbf{s}_2$, $\mathbf{s}_3$) = 0.4, and d($\mathbf{s}_1$, $\mathbf{s}_3$) = 0.1. Then the following is a reasonable mapping

$\mathbf{s}_1$ $\mathbf{s}_2$ $\mathbf{s}_3$
0.25 0.3

To quantify the goodness-of-fit of a mapping, a popular measure is

$$d_{MDS}(f) = \sum_{s_i, s_j} \left(\frac{d_{j(s_i), f(s_j)} - d(s_i.s_j)}{d_{j(s_i), f(s_j)}} \right)$$

where $f(s)$ is the location of a sample pattern s in the new space and d(s_i,s_j) is the Euclidean metric. We try to find a mapping that minimized d_{MDS}.

III. Cluster Analysis of Microarray Data