Introduction to Computational Biology
Lecture # 24: Unsupervised Analysis of Gene Expression Data

Yosef Prat
5.6.2007

We are facing the problem of gene expression data analysis. Our goal is to describe some attributes of the data, in a way that will allow us to understand the biological meanings. Two examples of approaching this problem are clustering and dimension reduction (as in PCA). We will focus on clustering.

1 Similarity based approach

This approach seeks the similarity or correlation between different genes. An example is the k-means clustering method. Another method is Hierarchical Agglomerative Clustering (HCA), an iterative algorithm that builds a similarity tree. We mentioned in previous lecture one kind of HCA, which called UPGMA. In UPGMA the distance between two clusters was defined by the mean-linkage:

Let $C_1 = \{g_1, \ldots, g_k\}, C_2 = \{g_{k+1}, \ldots, g_n\}$ be two clusters and $d(C_1, C_2)$ the distance between them:

$$d(C_1, C_2) = \frac{1}{k(n-k)} \sum_{i=1}^{k} \sum_{j=k+1}^{n} d(g_i, g_j)$$

Which means that the distance between two clusters is the average distance between elements in one cluster and elements in the other. Another possible definition for the distance between clusters is single-linkage:

$$d(C_1, C_2) = \underset{i,j}{\min} d(g_i, g_j)$$

Or complete-linkage:

$$d(C_1, C_2) = \underset{i,j}{\max} d(g_i, g_j)$$

Figure 1: Distance between clusters - the 3 linkage methods
2 Attribute based approach

This approach will be described by a statistical model. We discuss gene expression analysis but this model can be applied to different clustering needs. Let \( g \) represent a gene and \( a \) an array (results of one experiment in the microarray - e.g. hit shock). \( X_{ga} \) is the measurement of gene \( g \) in array \( a \). We are looking for clusters of genes. Let \( C_g \) be a random variable representing the cluster of gene \( g \). \( C_g \) is a hidden variable of course.

![Image](image.png)

Figure 2: micro-array

In order to be able to analyze the statistical model we use some simplifying assumptions.

**Assumption 1:** Genes were sampled i.i.d. For example, if we assume one measurement in \( \mathbb{R}^1 \) for each gene, from the distribution shown in Figure 3, we could expect some clusters around the red areas of the \( X \) values, even though the genes are sampled i.i.d.

![Image](image.png)

Figure 3: Some hypothetical distribution of genes measures

The probability for all of the matrix elements is therefore:

\[
P(X_{11}, ..., X_{mn}) = \prod_g P(X_{g1}, ..., X_{gn}) = \prod_g \left( \sum_{C_g} P(X_{g1}, ..., X_{gn}, C_g) \right)
\]

**Assumption 2:** given \( C_g \), For every \( a \), \( X_{ga} \) is independent on every \( X_{ga'} \).

**Assumption 3:**

\[
P(X_{ga} | C_g = c) \sim N(\mu_{ca}, \sigma_{ca}^2)
\]
The probability for the entire matrix is:

\[ P(X_{11}, \ldots, X_{mn}) = \prod_g \left( \sum_{C_g} P(C_g) \prod_a P(X_{ga} \mid C_g) \right) \]

Now we can ask what is the probability of being in a specific cluster given the values of a gene. using Bayes rule and the iid assumption:

\[ P(C_g = c \mid X_{g1}, \ldots, X_{gn}) \propto P(C_g = c) \prod_a P(X_{ga} \mid C_g = c) = P(c) \prod_a \frac{1}{\sqrt{2\pi} \sigma_{ca}} e^{-\frac{1}{2} \left( \frac{(x_{ga} - \mu_{ca})^2}{\sigma_{ca}^2} \right)} \]

For convenient only lets assume \( P(c) \) is uniform for every cluster, and that \( \forall c, \alpha \sigma_{ca} = 1 \), and we get:

\[ P(C_g = c \mid X_{g1}, \ldots, X_{gn}) \propto \prod_a e^{-\frac{1}{2} (x_{ga} - \mu_{ca})^2} = e^{-\frac{1}{2} \sum_a (x_{ga} - \mu_{ca})^2} = e^{-\frac{1}{2} \| \mathbf{x}_g - \mu_c \|^2_2} \]

Which means, that the probability of a gene to be in a cluster, is proportional to the negative exponent of its distance from the cluster expectancy vector. Hence, there is a higher probability that a gene will end up in a cluster closer (its center) to the gene.

3 Learning the parameters

Because of the likelihood being dependent on the hidden variables \( (C_g) \), an EM can be suitable for learning the parameters. (Improvement: apply k-means algorithm and use its results as the EM initial values)

Sufficient statistics would be the first and the second moments - \( \sum_i x_i, \sum_i x_i^2 \). Lets define three matrices:

\[ S^1_{ca} = \sum_g 1\{C_g = c\} \]
\[ S^2_{ca} = \sum_g 1\{C_g = c\} \cdot X_{ga} \]
\[ S^3_{ca} = \sum_g 1\{C_g = c\} \cdot X^2_{ga} \]

We are interesting in the expectations of these matrices because:

\[ E[1\{C_g = c\}] = P(C_g = c \mid X_{g1}, \ldots, X_{gn}) \]
\[ E[S^1_{ca}] = \sum_g P(C_g = c \mid X_g) \]
\[ E[S^2_{ca}] = \sum_g P(C_g = c \mid X_g) \cdot X_{ga} \]
\[ E[S^3_{ca}] = \sum_g P(C_g = c \mid X_g) \cdot X^2_{ga} \]

And we already expressed \( P(C_g = c \mid X_g) \) before. With these definitions estimator for the expectancy will be

\[ \hat{\mu}_{ca} = \frac{S^2_{ca}}{S^1_{ca}} \]

And estimator for the variance will be

\[ \hat{\sigma^2_{ca}} = \frac{S^3_{ca}}{S^1_{ca}} - \left( \frac{S^2_{ca}}{S^1_{ca}} \right)^2 \]
We still have the same problem as we had in the k-means algorithm: how do we choose k (number of clusters to look for). We can define some agglomerative model, that will decide to join clusters in certain stages (this can be done through the sufficient statistics). When clusters are joined together the likelihood drops, and then we run the EM again to gain likelihood, see Figure 4. Good points for stopping the process at would be before joining clusters (marked with red dots).

![Figure 4: The likelihood drops when clusters are merged](image-url)