Chapter 9

Unsupervised machine learning

Unsupervised machine learning (a.k.a. cluster analysis) is a set of methods to assign objects into clusters under a predefined distance measure when class labels are unknown. Unsupervised machine analysis is usually more difficult than supervised machine learning because the class labels are unknown and consequently the performance and statistical properties of the methods are difficult to assess. Below we will introduce a few topics and methods that are useful in high-throughput genomic data analysis.

9.1 Distance and similarity measure

??check Chapter 6 of Modern Multidimensional Scaling by Borg and Groenen.

For any cluster analysis to work, a distance or dissimilarity measure should be defined so that cluster algorithms can assign objects near from each other into a cluster. Below, we describe a few types of distance measures and similarity measures. Taking expression microarray data as an example, denote by x_{gs} the expression intensity of gene g and sample s ($1 \leq g \leq G, 1 \leq s \leq S$). Denote by x_g the intensity vector of gene g: $x_g = \{x_{g1}, x_{g2}, \cdots, x_{gS}\}$.

9.1.1 Distance measures

Minkowski distance The Minkowski family of distance measures are defined for two gene vectors x_{g_i} and x_{g_j} as

$$d(x_{g_i}, x_{g_j}) = \left(\sum_{s=1}^{S} |x_{g_i s} - x_{g_j s}|^k \right)^{1/k}$$

for some integer k. Three commonly used distances follow below.
Manhattan distancess (a.k.a. city-block distance) ($k = 1$):

$$d(x_{g_i}, x_{g_j}) = \sum_{s=1}^{S} |x_{g_i s} - x_{g_j s}|$$

Euclidean distance ($k = 2$):

$$d(x_{g_i}, x_{g_j}) = \left(\sum_{s=1}^{S} |x_{g_i s} - x_{g_j s}|^2 \right)^{1/2}$$

Maximum distance ($k = \infty$; see exercise 1):

$$d(x_{g_i}, x_{g_j}) = \max_s |x_{g_i s} - x_{g_j s}|$$

Note that these distances are scale-dependent. For example, if the intensities are not properly normalized across arrays, some arrays with brighter signals (higher intensities) tend to dominate the analysis. The equal-distance bounds from origin are spherical for Euclidean distance, diamond-shape for Manhattan distance and square for maximum distance.
Mahalanobis distance The Minkowski distances implicitly assume uncorrelated variables. This is often not true in the data analysis. One solutation is to "decorrelate" the variables by Mahalanobis distance:

$$d(x_{g_i}, x_{g_j}) = \sqrt{(x_{g_i} - x_{g_j})\Sigma^{-1}(x_{g_i} - x_{g_j})'}$$

Note that Euclidean distance is a special case of Mahalanobis distance with identity covariance matrix. In cluster analysis, the matrix Σ is often estimated by pooling all within-group covariance matrixes.
Gower's distance Gower (1971) proposed an generalizable correlation coefficient when variables span from numeric, binary, categorical to ordinal. The general idea is the standardize the distance contribution of each variable to be between 0 and 1. For example, for numerical variables, the absolute differences are standardized by the largest possible value (i.e. the range).

9.1.2 Similarity measure

Pearson correlation and inner product Pearson correlation is defined as

$$r(x_{g_i}, x_{g_j}) = \frac{cov(x_{g_i}, x_{g_j})}{\sqrt{var(x_{g_i})} \cdot \sqrt{var(x_{g_j})}} = \frac{\sum_{s=1}^{S} (x_{g_i s} - \bar{x}_{g_i}) \cdot (x_{g_j s} - \bar{x}_{g_j})}{\sqrt{\sum_{s=1}^{S} (x_{g_i s} - \bar{x}_{g_i})^2} \cdot \sqrt{\sum_{s=1}^{S} (x_{g_j s} - \bar{x}_{g_j})^2}}$$

Pearson correlation is always between -1 and 1. If the gene vectors are standardized to zero mean and unit variance, Pearson correlation reduces to "inner product":

$$r(x_{g_i}, x_{g_j}) = \sum_{s=1}^{S} x_{g_i s} \cdot x_{g_j s}$$

As a result, Pearson correlation has an intuitive geometric interpretation that it measures the "$\cos\theta$" of two vectors. In the standardized case, Pearson correlation and Euclidean distance are equivalent in the sense that $d^2 = 2 - 2 \cdot r$, where d is the Euclidean distance and r is the Pearson correlation (Exercise 2).

When the underlying variables have a bivariate normal distribution, the statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

follows a Student's t-distribution with degree of freedom S-2 under null hypothesis (zero correlation). This also holds approximately even if the observed values are non-normal, provided sample sizes are not very small. Note that when the gene vectors are of very high dimension (e.g. several thousands), the hypothesis testing is easily rejected even if the estimated Pearson correlation is very low.

Spearman rank correlation
Pearson correlation is known to be sensitive to outliers. This can be a major problem in high-throughput genomic data analysis where noises and measurement errors are peppered in the data. A more robust alternative is to replace the original raw intensities by the ranks. This leads to the Spearman rank correlation. It can be shown that if no rank ties, the Spearman rank correlation can be written as

$$r(x_{g_i}, x_{g_j}) = 1 - \frac{6 \cdot \sum_{s=1}^{S} d_s^2}{n \cdot (n^2 - 1)}$$

where $d_s = x_{g_i s} - x_{g_j s}$ (Exercise 3).

To test whether Spearman rank correlation is zero or not, one can use permutation test or apply the t-test used in Pearson correlation. Another Fisher transformation approach is to use the z-statistic

$$z = \sqrt{\frac{S-3}{1.06}} \cdot \operatorname{arctanh}(r) = \sqrt{\frac{S-3}{1.06}} \cdot \frac{1}{2} \ln \frac{1+r}{1-r}.$$

Under null hypothesis (zero correlation), z-statistic approximately follows a standard normal distribution. Note that Pearson correlation focuses to detect linear correlation. Spearman correlation has an advantage to detect non-linear association because of the rank statistics used. For example, if the two vectors have quadratic relationship, Spearman correlation generates correlation close to 1 while Pearson correlation cannot.

Other correlation measures for non-numeric variables When data are not from numerical variables, different correlation measures have been developed (e.g. biserial, polyserial correlations or Kendall's tau correlation). More details will be added in the future.

9.2 Clustering methods

9.2.1 Existing popular methods

Hierarchical clustering Hierarchical clustering is one of the most popular clustering algorithm in genomic research and it is probably also the most misused one. The algorithm merges the "nearest" two nodes at each iteration. For G objects, $G-1$ iterations will construct a hierarchical tree. To define the "nearest" two nodes (each node contain one or multiple objects), different linkage can be pre-specified. Commonly used linkages include single linkage (compute distance of the nearest pair), complete linkage (comput distance of the furthest pair), average linkage (compute average distance of all pairs) and centroid linkage (compute distance between centroids of two nodes). Single linkage tends to form elongated clusters. Complete and average linkage tends to find more spherical clusters. Note that the ordering of objects in the graphical presentation of a hierarchical tree is not unique. The ordering can sometimes be misleading in visualization. Since hierarchical clustering is an iterative local agglomerative algorithm, mistakes can accumulate in the clustering process of thousands of genes and the result has been found inferior than global optimization methods such as K-means or model-based clustering.

Likelihhod-based inference: model-based clustering and K-means

Many clustering methods are based on global optimization of a criterion that measures compatibility of the clustering result to the data. K-means and mixture Gaussian model-based clustering are examples of this category. In the statistical literature, clustering is often obtained through likelihood-based inference including the mixture maximum likelihood (ML) approach and the classification maximum likelihood (CML) approach (see Celeux and Govaert 1993; Ganesalingam 1989). In the ML approach, the R^d-valued vectors $x_1, \cdots, x_n$ are sampled from a mixture of densities, $f(x) = \sum_{j=1}^{k} \pi_j f(x, \theta_j)$, where π_j is the probability that the data is generated from cluster j and each $f(x, \theta_j)$ is the density for cluster j from the same parametric family with parameter θ_j. The log-likelihood to be maximized is

$$L = \log \left\{ \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f(x_i, \theta_j) \right\}$$

and clustering is obtained by the assignment of each x_i to the cluster with greatest posterior probability. For more details refer to Fraley and Raftery (2002); McLachlan *et al.* (2002).

In the CML approach, the partition $C = \{C_1, \cdots, C_k\}$, where C_j's are disjoint subsets of $X = \{x_1, \cdots, x_n\}$, is considered as an unknown parameter and is directly pursued in the optimization. Two types of CML criteria under different sampling schemes have been discussed in the literature. The first criterion samples data of $n_1, \cdots, n_k$ observations in each cluster, where n_j's are fixed and unknown and $\sum n_j = n$. Following the convention of Celeux and Govaert (1992), the $\mathcal{C}_1$-CML criterion takes the form (see Scott and Symons, 1971)

$$\mathcal{C}_1(C, \theta) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \log f(x_i, \theta_j).$$

The second type assumes that observations are sampled at random from the mixture distribution and thus $n_1, \cdots, n_k$ is a multinomial distribution of sample size n and probability parameters $\pi = \{\pi_1, \cdots, \pi_k\}$. The $\mathcal{C}_2$-CML criterion leads to (see Symons, 1981)

$$\mathcal{C}_2(C, \pi, \theta) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \log\{\pi_j f(x_i, \theta_j)\}.$$

It is easily seen that $\mathcal{C}_2$-CML can be viewed as a penalized $\mathcal{C}_1$-CML or that $\mathcal{C}_1$-CML is a special form of $\mathcal{C}_2$-CML with an implicit equal proportions

assumption (Bryant, 1991; Celeux and Govaert, 1992). Below, we restrict f to be Gaussian distributed with $\theta_j = (\mu_j, \Sigma_j)$ and the $\mathcal{C}_1$-CML criterion becomes

$$\mathcal{C}_1(C, \theta) = f(x|C, \theta) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \log f(x_i|\mu_j, \Sigma_j) \tag{9.1}$$

where $f(x_i|\mu_j, \Sigma_j) = \frac{\exp\{-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)\}}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}$.

In addition to likelihood-based inference, many clustering methods have utilized heuristic global optimization criteria. K-means (Hartigan and Wong, 1979) is an effective clustering algorithm in this category and is applied in many applications due to its simplicity. In the K-means criterion, objects are assigned to clusters so that the within cluster sum of squared distance is minimized.

$$L = \sum_{j=1}^{k} \sum_{x_i \in C_j} |x_i - \bar{x}^{(j)}|^2$$

where $\bar{x}^{(j)}$ is the mean of objects in cluster j. It can be shown that that K-means is actually a simplified form of the $\mathcal{C}_1$-CML sampling scheme under the Gaussian assumption when the covariance matrixes are identical and spherical in all clusters. The optimization of K-means is an NP-hard problem. One simple solution is by EM algorithm (Exercise ??). Hartigan and Wong (1979) provides a very faster algorithm. K-means algorithm are generally faster ($\sim O(N \cdot K)$) than other clustering algorithms but all these algorithms encounter local optimization issues.

Note that K-means clustering requires objects to locate in an Euclidean space (so that the cluster centers can be calculated). This can be relaxed by replacing cluster centers by cluster medoids and it leads the K-medoids (or Partitioin around medoids; PAM).

$$L = \sum_{j=1}^{k} \sum_{x_i \in C_j} |x_i - \tilde{x}^{(j)}|^2$$

where $\tilde{x}^{(j)}$ is the medoids of objects in cluster j. K-medoids is useful in many applications when only dissimilarity measures are definded for each pair of objects (Similar to PCA versus MDS in dimension reduction).

9.2.2 Some advanced methods

Penalized and weighted K-means In Tseng (2007), a general class of loss function extended from K-means is proposed for clustering purposes:

$$W(C;k,\lambda)=\sum_{j=1}^{k}\sum_{x_i\in C_j} w(x_i;\underset{\sim}{P})\cdot d(x_i,C_j)+\lambda|S| \tag{9.2}$$

where $w(\cdot;\cdot)$ is a function of the weighting factor, $\underset{\sim}{P}$ is the prior information available, $d(x,C_j)$ calculates the dispersion of point x in cluster C_j, $|\cdot|$ represents size of the set and λ is a tuning parameter representing the degree of penalty of each noise point. The weighting factor $w(\cdot;\cdot)$ is used to incorporate prior knowledge of preferred or prohibited patterns of cluster selections. Minimizing equation (9.2) with given weight function $w(\cdot;\cdot)$, distance measure $d(\cdot,\cdot)$, k and λ produces a clustering solution. We denote by $C^*(k,\lambda)=\{C_1^*(k,\lambda),\cdots,C_k^*(k,\lambda),S^*(k,\lambda)\}$ the minimizer of $W(C;k,\lambda)$. Proposition 1. lists several useful properties of this formulation (Exercise 4). In particular, λ is inversely related to the number of noise points (i.e. $|S|$). This is a desirable property to control tightness of the resulting clusters in practice.

Proposition 1. (a) Similar to K-means if $k_1>k_2$, then $W(C^*(k_1,\lambda);k_1,\lambda)\leq W(C^*(k_2,\lambda);k_2,\lambda)$. (b) If $\lambda_1>\lambda_2$, then $|S^*(k,\lambda_1)|\leq|S^*(k,\lambda_2)|$. (c) If $\lambda_1>\lambda_2$, then $W(C^*(k,\lambda_1);k,\lambda_1)>W(C^*(k,\lambda_2);k,\lambda_2)$.

Note that K-means and K-medoids are special cases of the general class of PW-Kmeans. We can consider a simpler penalized K-means form without weights:

$$W_P(C;k,\lambda_0)=\sum_{j=1}^{k}\sum_{x_i\in C_j}|x_i-\bar{x}^{(j)}|^2+\eta^2\cdot\lambda_0|S| \tag{9.3}$$

where λ_0 is a tuning parameter, $\eta=H/\sqrt[S]{k}$, H is defined as the average pairwise distance of the data and S is the dimensionality of the data. The purpose of H and $\sqrt[S]{k}$ is to avoid the scaling problem of the penalty term. Under this formulation, the selection of λ_0 is invariant under data scaling and different k. In contrast to K-means, P-Kmeans provides flexibility of not assigning all points into clusters and allows a set of noise (or scattered) points, S. These points are defined as noises that do not tightly share common patterns with any of the clusters in the data. For clustering problems in complex data such as gene clustering in expression

profiles, ignoring scattered points has been found to dilute identified patterns, make more false positives and even distort cluster formation and interpretation. Similar to K-means, we can find relationships between penalized K-means and classification likelihood. If the scattered points in S are uniformly distributed in the hyperspace V (i.e. generated from a homogeneous Poisson process), then the $\mathcal{C}_1$ CML criterion becomes

$$f(x|C,\theta) = \prod_{j=1}^{k} \prod_{x_i \in C_j} f(x_i|\mu_j, \Sigma_j) \prod_{x_i \in S} \frac{1}{|V|}, \tag{9.4}$$

where $|V|$ is the hypervolume of V (see a similar model in Fraley and Raftery, 2002). Assume $\Sigma_j = \sigma_0^2 I$. We find that maximizing (9.4) is equivalent to minimizing (9.3) if $\lambda_0 = 2\sigma_0^2 \cdot (1/\eta)^2 \cdot \log|V|$. This relationship provides good guidance for the selection of λ_0.

Tight clustering and other heuristic methods methods Tight gene modules with similar gene expression patterns often imply gene co-regulation or share related biological functions and are basic elements in many genomic exploratory data analyses. In the Tight Clustering method (Tseng and Wong, 2005), it directly identifies small and tight clusters in the data and allows a set of scattered genes without being clustered. The method utilizes resampling techniques to obtain consistent tight clusters in repeated subsampling evaluations. Since the method is based on resampling techniques, the computation demand is higher and the result can depend on the sampling seed or small data perturbation.

Many other heuristic clustering methods exist. For example, Consensus Clustering (Monti et al, 2003) and CLICK (Sharan and Shamir, 2000) are popular algorithms for gene clustering in microarray analysis.

Bayesian clustering Bayesian clustering has been developed for microarray data analysis (Medvedovic et al., 2002 and Qin 2006). One important advantage is that the number of clusters does not need to be specified *a priori* but can be inferred from the posterior distribution while a major disadvantage of Bayesian clustering is the assumptions made behind the Bayesian modeling.

9.3 Estimate the number of clusters??

Estimating the number of clusters is a difficult problem due to the lack of true underlying class labels. Methods for this purpose will be added in the future.

9.4 Clustering evaluation??

As discussed above, evaluating performance of clustering methods is a difficult task. More methods on this topic will be added in the future. **adjusted Rand index**

Exercise:

1. Show that the maximum distance is a special case of Minkowski family when $k \to \infty$.
2. Show that when gene intensity vectors x_{g_i} and x_{g_j} are standardized to zero mean and unit variance, $d^2 = 2-2\cdot r$ where d is the Euclidean distance and r is the Pearson correlation of the two genes.
3. Show that Spearman correlationi can be calculated in the simpler form $r(x_{g_i}, x_{g_j}) = 1 - \frac{6\cdot\sum_{s=1}^{S} d_s^2}{n\cdot(n^2-1)}$ if no rank tie exists.
4. Prove Proposition 1 in Penalized and weighted K-means.
5. Prove that K-means is a special case of $\mathcal{C}_1$-CML criterion in 9.1 under Gaussian assumption and when the covariance matrixes are identical and spherical.