Flat clustering approaches

Colin Dewey
BMI/CS 576
www.biostat.wisc.edu/bmi576
colin.dewey@wisc.edu
Fall 2016
Flat clustering

- Cluster objects/genes/samples into $K$ clusters
- In the following, we will consider clustering **genes** based on their expression profiles
- $K$: number of clusters, a user defined argument
- Two example algorithms
  - $K$-means
  - Gaussian mixture model-based clustering
Notation for $K$-means clustering

- $K$ number of clusters
- $N_k$ Number of elements in cluster $k$
- $x_i$ $p$-dimensional expression profile for $i^{th}$ gene
- $X = \{x_1, \ldots, x_N\}$ is the collection of $N$ gene expression profiles to cluster
- $f_k$ Center of the $k^{th}$ cluster
- $C(i)$ Cluster assignment ($1$ to $K$) of $i^{th}$ gene
**K**-means clustering

- Hard-clustering algorithm
- Dissimilarity measure is the Euclidean distance
- Minimizes *within-cluster scatter* defined as

\[
\sum_{k=1}^{K} \sum_{i: C(i)=k} \left\| x_i - f_k \right\|^2
\]

- Center of cluster \( k \)
- Sum over all genes assigned to cluster \( k \)
$K$-means clustering

- consider an example in which our vectors have 2 dimensions
$K$-means clustering

- each iteration involves two steps
  - assignment of profiles to clusters
  - re-computation of the cluster centers (means)
**K-means algorithm**

- **Input:** \( K \), number of clusters, a set \( X = \{x_1, \ldots, x_N\} \) of data points, where \( x_i \) are \( p \)-dimensional vectors
- **Initialize**
  - Select initial cluster means \( f_1, \ldots, f_K \)
- **Repeat until convergence**
  - Assign each \( x_i \) to cluster \( C(i) \) such that
    \[
    C(i) = \arg\min_{1 \leq k \leq K} \| x_i - f_k \|^2
    \]
  - Re-estimate the mean of each cluster based on new members
**K-means: updating the mean**

- To compute the mean of the $k^{th}$ cluster

$$f_k = \frac{1}{N_k} \sum_{i : C(i) = k} x_i$$

- $N_k$: Number of genes in cluster $k$
- $x_i$: All genes in cluster $k$
$K$-means stopping criteria

- Assignment of objects to clusters don’t change
- Fix the max number of iterations
- Optimization criterion changes by a small value
**K-means Clustering Example**

Given the following 4 instances and 2 clusters initialized as shown.

\[
\text{dist}(x_i, x_j)^2 = \|x_i - x_j\|^2
\]

\[
\begin{align*}
\text{dist}(x_1, f_1)^2 &= 2, \quad \text{dist}(x_1, f_2)^2 = 13 \\
\text{dist}(x_2, f_1)^2 &= 2, \quad \text{dist}(x_2, f_2)^2 = 9 \\
\text{dist}(x_3, f_1)^2 &= 9, \quad \text{dist}(x_3, f_2)^2 = 2 \\
\text{dist}(x_4, f_1)^2 &= 61, \quad \text{dist}(x_4, f_2)^2 = 26
\end{align*}
\]

\[
\begin{align*}
f_1 &= \left(\frac{4 + 4}{2}, \frac{1 + 3}{2}\right) = (4,2) \\
f_2 &= \left(\frac{6 + 8}{2}, \frac{2 + 8}{2}\right) = (7,5)
\end{align*}
\]

\[
\begin{align*}
\text{dist}(x_1, f_1)^2 &= 1, \quad \text{dist}(x_1, f_2)^2 = 25 \\
\text{dist}(x_2, f_1)^2 &= 1, \quad \text{dist}(x_2, f_2)^2 = 13 \\
\text{dist}(x_3, f_1)^2 &= 4, \quad \text{dist}(x_3, f_2)^2 = 10 \\
\text{dist}(x_4, f_1)^2 &= 52, \quad \text{dist}(x_4, f_2)^2 = 10
\end{align*}
\]
**K-means Clustering Example**
*(Continued)*

assignments remain the same, so the procedure has converged.

\[
f_1 = \left(\frac{4 + 4 + 6}{3}, \frac{1 + 3 + 2}{3}\right) = (4.67, 2)
\]

\[
f_2 = \left(\frac{8}{1}, \frac{8}{1}\right) = (8, 8)
\]
Gaussian mixture model based clustering

• $K$-means is hard clustering
  – At each iteration, a data point is assigned to one and only one cluster

• We can do soft clustering based on Gaussian mixture models
  – Each cluster is represented by a distribution (in our case a Gaussian)
  – We assume the data is generated by a mixture of the Gaussians
  – Clustering may be accomplished via an Expectation-Maximization (EM) algorithm
Gaussian distribution

- Gaussian distribution

\[ f(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right) \]

\( \mu \): Mean
\( \sigma \): Standard deviation
Representation of Clusters

- in the EM approach, we’ll represent each cluster using a 
  $p$-dimensional multivariate Gaussian

$$
N_j(\tilde{x}_i) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_j|}} \exp \left[ -\frac{1}{2} (\tilde{x}_i - \tilde{\mu}_j)^T \Sigma_j^{-1} (\tilde{x}_i - \tilde{\mu}_j) \right]
$$

where

- $\tilde{\mu}_j$ is the mean of the Gaussian
- $\Sigma_j$ is the covariance matrix

this is a representation of a Gaussian in a 2-D space
Cluster generation

• We model our data points with a generative process
• Each point is generated by:
  – Choosing a cluster (1,2,…,k) by sampling from probability distribution over the clusters
    • Prob(cluster k) = P_k
  – Sampling a point from the Gaussian distribution N_j
EM Clustering

- the EM algorithm will try to set the parameters of the Gaussians, $\Theta$, to maximize the log likelihood of the data, $X$

$$\text{log likelihood}(X \mid \Theta) = \log \prod_{i=1}^{n} \Pr(\tilde{x}_i)$$

$$= \log \prod_{i=1}^{n} \sum_{k=1}^{K} P_k N_k(\tilde{x}_i)$$

$$= \sum_{i=1}^{n} \log \sum_{k=1}^{K} P_k N_k(\tilde{x}_i)$$
EM Clustering

- the parameters of the model, $\Theta$, include the means, the covariance matrix and sometimes prior weights for each Gaussian
- here, we’ll assume that the covariance matrix is fixed; we’ll focus just on setting the means and the prior weights
EM Clustering: Hidden Variables

- on each iteration of $K$-means clustering, we had to assign each instance to a cluster
- in the EM approach, we’ll use hidden variables to represent this idea
- for each instance $\vec{x}_i$ we have a set of hidden variables $Z_{i1}, \ldots, Z_{iK}$
- we can think of $Z_{ij}$ as being 1 if $\vec{x}_i$ is a member of cluster $j$ and 0 otherwise
EM Clustering: the E-step

• recall that $Z_{ij}$ is a hidden variable which is 1 if $N_j$ generated $\tilde{x}_i$ and 0 otherwise

• in the E-step, we compute $h_{ij}$, the expected value of this hidden variable

$$h_{ij} = E(Z_{ij} | \tilde{x}_i) = \Pr(Z_{ij} = 1 | \tilde{x}_i) = \frac{P_j N_j(\tilde{x}_i)}{\sum_{l=1}^{K} P_l N_l(\tilde{x}_i)}$$
EM Clustering: the M-step

- given the expected values $h_{ij}$, we re-estimate the means of the Gaussians and the cluster probabilities

$$\bar{\mu}_j = \frac{\sum_{i=1}^{n} h_{ij} \bar{x}_i}{\sum_{i=1}^{n} h_{ij}}$$

$$P_j = \frac{\sum_{i=1}^{n} h_{ij}}{n}$$

- can also re-estimate the covariance matrix if we’re not treating it as fixed
EM Clustering – Overall algorithm

• Initialize parameters (e.g., means)
• Loop until convergence
  – E-step: Compute expected values of $Z_{ij}$ values given current parameters
  – M-step: Update parameters using $E[Z_{ij}]$ values
    • Means
    • Cluster probabilities
EM Clustering Example

Consider a one-dimensional clustering problem in which the data given are:

\[ x_1 = -4 \]
\[ x_2 = -3 \]
\[ x_3 = -1 \]
\[ x_4 = 3 \]
\[ x_5 = 5 \]

The initial mean of the first Gaussian is 0 and the initial mean of the second is 2. The Gaussians both have variance = 4; their density function is:

\[
 f(x | \mu) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{2})^2}
\]

where \( \mu \) denotes the mean (center) of the Gaussian.

Initially, we set \( P_1 = P_2 = 0.5 \)
EM Clustering Example

\[ f(x | \mu) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2} \left( \frac{x - \mu}{2} \right)^2} \]

- \[ f(-4 | \mu_1) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2} \left( \frac{-4 - 0}{2} \right)^2} = .0269 \]
- \[ f(-4 | \mu_2) = .0022 \]
- \[ f(-3 | \mu_1) = .0646 \]
- \[ f(-3 | \mu_2) = .00874 \]
- \[ f(-1 | \mu_1) = .176 \]
- \[ f(-1 | \mu_2) = .0646 \]
- \[ f(3 | \mu_1) = .0646 \]
- \[ f(3 | \mu_2) = .176 \]
- \[ f(5 | \mu_1) = .00874 \]
- \[ f(5 | \mu_2) = .0646 \]
EM Clustering Example: E Step

\[ h_{11} = \frac{\frac{1}{2} f(x_1 | \mu_1)}{\frac{1}{2} f(x_1 | \mu_1) + \frac{1}{2} f(x_1 | \mu_2)} = \frac{.0269}{.0269 + .0022} = 0.924 \]

\[ h_{21} = \frac{\frac{1}{2} f(x_2 | \mu_1)}{\frac{1}{2} f(x_2 | \mu_1) + \frac{1}{2} f(x_2 | \mu_2)} = \frac{.0646}{.0646 + .00874} = 0.881 \]

\[ h_{31} = \frac{.176}{.176 + .0646} = 0.732 \]

\[ h_{41} = \frac{.0646}{.0646 + .176} = 0.268 \]

\[ h_{51} = \frac{.00874}{.00874 + .0646} = 0.119 \]

\[ h_{12} = \frac{\frac{1}{2} f(x_1 | \mu_2)}{\frac{1}{2} f(x_1 | \mu_1) + \frac{1}{2} f(x_1 | \mu_2)} = \frac{.0022}{.0269 + .0022} = 0.076 \]

\[ h_{22} = \frac{.00874}{.0646 + .00874} = 0.119 \]

\[ h_{32} = \frac{.0646}{.176 + .0646} = 0.268 \]

\[ h_{42} = \frac{.176}{.0646 + .176} = 0.732 \]

\[ h_{52} = \frac{.0646}{.00874 + .0646} = 0.881 \]
**EM Clustering Example: M-step**

\[
\mu_1 = \frac{\sum_i x_i \times h_{i1}}{\sum_i h_{i1}} = \frac{-4 \times .924 + -3 \times .881 + -1 \times .732 + 3 \times .268 + 5 \times .119}{.924 + .881 + .732 + .268 + .119} = -1.94
\]

\[
\mu_2 = \frac{\sum_i x_i \times h_{i2}}{\sum_i h_{i2}} = \frac{-4 \times .076 + -3 \times .119 + -1 \times .268 + 3 \times .732 + 5 \times .881}{.076 + .119 + .268 + .732 + .881} = 2.73
\]

\[
P_1 = \frac{\sum_i h_{i1}}{n} = \frac{.924 + .881 + .732 + .268 + .119}{5} = 0.58
\]

\[
P_2 = \frac{\sum_i h_{i2}}{n} = \frac{.076 + .119 + .268 + .732 + .881}{5} = 0.42
\]
EM Clustering Example

- here we’ve shown just one step of the EM procedure

- we would continue the E- and M-steps until convergence
Comparing K-means and GMMs

• K-means
  – Hard clustering
  – Optimizes within cluster scatter
  – Requires estimation of means

• GMMs
  – Soft clustering
  – Optimizes likelihood of the data
  – Requires estimation of mean (and possibly covariance) and mixture probabilities
Choosing the number of clusters

• Picking the number of clusters based on the clustering objective will result in $K=N$ (number of data points)

• Two common options:
  – Pick $k$ based on penalized clustering objective $G$
    \[
    \bar{K} = \arg\min_{K, \Theta} G(X, K, \Theta) + \lambda \cdot K
    \]
  – Pick based on cross-validation
    • Appropriate for EM clustering
Picking $K$ based on cross-validation

1. Split data set into three sets: Training set, Test set, and a third set.
2. Cluster each of the three sets separately.
3. Evaluate the results of each clustering on the test data.
4. Compute the objective based on the test data for each clustering.
5. Run the method on all data once $K$ has been determined.
Summary of flat clustering

• Two methods
  – K-means (hard clustering)
  – Gaussian mixture models (soft clustering)
• Both need users to specify K
• K can be picked based on penalized objective or cross-validation