Model-Based Clustering

Review of Partitional Clustering, K-Means:

1. Decide # of clusters, K
2. Assign initial estimates for the center of each of K clusters
3. Assign each point to its nearest center
4. Recalculate the center based on these assignments
5. Iterate 3 & 4 until convergence (reduce sum of squares)

Model Based Clustering and Data Transformations of Gene Expression Data

1. Motivation
2. Model-Based Clustering
3. Validation
4. Summary and Conclusion

I. Motivation: Various clustering methods (K-Means, Hierarchical Average Link, etc.) yield varying solutions. However, these methods often do not arrive at an obvious solution. Model-based clustering allows us to "fit" data to a more obvious model.

II. Model-Based Clustering: Based on the idea that each cluster is generated by a multivariate normal distribution. It is also called the "Gaussian Mixture Model" because it consists of a mixture of several normal distributions.

a. Each cluster, k, has two parameters:
   - Mean vector μ_k
   - Covariance matrix Σ_k

b. General Approach to Model-Based Clustering (similar to K-Means):
   i. Initialize by randomly assigning points to clusters
   ii. Calculate parameters (mean and covariance) for each cluster
   iii. Calculate probabilities of cluster membership for each point and assign points to clusters with highest probability
   iv. Return to (ii)

   Note: If we know the number of clusters (distributions) and their parameters, we can calculate the probability that a given point belongs to a specific cluster. Conversely, if we know the correct cluster for every point, we can calculate the parameters for each cluster.

c. Statistics Review

Variance, Covariance, and Correlation

$$\text{var}(x) = E((x - \bar{x})^2)$$

$$\text{cov}(x, y) = E((x - \bar{x})(y - \bar{y}))$$

If large and positive → x and y move in the same direction at the same time
If large and negative → x and y move oppositely
If 0 → x and y are independent

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

Correlation, cor(x,y) is a value between –1 and +1.

Univariate and Multivariate Gaussian Distributions

Univariate:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\bar{x})^2/\sigma^2}$$

Multivariate:

$$\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\bar{x})^T(\Sigma^{-1})(x-\bar{x})}$$

Where Σ is the variance/covariance matrix:

$$\Sigma_{i,j} = E((x_i - \bar{x}_i)(x_j - \bar{x}_j))$$

Examples of Multivariate Gaussian Data:

1. var(x) = var(y) = 1; cov = 0 → circle
2. var(x) = 1, var(y) > 1, cov = 0 → ellipse
3. cov > 0 → Data points converge to the line x = y for increasing values of covariance.

d. Covariance Models: Linear algebra to decompose the variance matrix

$\Sigma_k = \lambda_k D_k A_k D_k^T$, where
λ is a scalar that specifies the volume.
A is a diagonal matrix that specifies the variances (i.e. the shape of the cloud; how spread-out is the ellipsoid in each dimension?).
D is a unit matrix that specifies the degree of rotation of the cloud.

Examples: Make models with fewer parameters than needed for the fully general covariance structure

- Equal Volume Spherical Model (EI): $\Sigma_k = \lambda I$
  - Assumes all clusters are spherical with the same variance; essentially K-means
- Unequal Volume Spherical Model (VI): $\Sigma_k = \lambda_k I$
  - The variance of spherical clusters can differ, i.e. different volumes
- Diagonal Model: $\Sigma_k = \lambda_k B_k$, where B_k is diagonal, $|B_k| = 1$
  - The variances can differ, giving rise to elliptical clusters
- EEE Elliptical Model: $\Sigma_k = \lambda DAD^T$
  - The clusters are elliptical, but the same covariance structure applies to all
- Unconstrained Model (VVV): $\Sigma_k = \lambda_k D_k A_k D_k^T$
  - Variable volume, shape, and orientation

Note the tradeoff: The more general (flexible) the model, the more parameters necessary.

e. EM (Expectation Maximization) Algorithm: General approach to maximum likelihood

- Iterate between E and M Steps (essentially the same method as K-means)
  - E step: Compute the probability of each observation belonging to each cluster using the current parameter estimates (means and variances).
  - M step: Estimate model parameters using the current group membership probabilities.
- If parameters are known, estimate clusters; if clusters are known, estimate parameters
- Slight refinement over K-means: As opposed to hard assignments, points can be weighted as members of multiple clusters based on calculated probabilities.
- Guaranteed to converge to a local optimum, but not necessarily a global optimum. (This generally works well in practice, though.)

f. Advantages of Model-Based Clusters

- Higher quality clusters
- Flexible models (whereas K-means only allows spherical clusters)

g. Model Selection: A principled way to choose the right model and the right number of clusters

- Relies on the Bayesian Information Criterion (BIC): Allows us to calculate the probability that our data set came from a given model. A large BIC score indicates strong evidence for the corresponding model.
- Definition of BIC Score

$$2\log p(D \mid M_k) \approx 2\log p(D \mid \hat{\theta}_k, M_k) - \upsilon_k \log(n) = BIC_k$$

where $p(D|M_k)$ is the probability of our dataset (D) given the model M_k, θ_k<hat> is the maximum likelihood estimate of the parameter θ_k, and v_k is the number of parameters to be estimated in the model M_k. The subtractive term penalizes for increasing the number of parameters.

Note: The BIC score is actually an approximation of integrated likelihood $p(D|M_k)$, which is difficult to solve (approximation is simply the first few terms of something like a Taylor series expansion).

III. Validation

a. Methodology

- Apply methods to data sets with external criteria: Are we getting good answers and does the BIC score lead us to those?
- The Adjusted Rand index compares clusters with external criteria (if adjusted Rand index = 1, there is perfect agreement; two random partitions of the same data have an expected index of 0)
- The quality of clusters found by model-based clustering can be compared to those found by the CAST and k-Means algorithms.

b. Gene Expression Data Sets

Real Data Sets

- Ovarian cancer data set encompassing 100,000 clones
  - Subset of data analyzed: 235 clones from 24 experiments (cancer versus normal tissue samples)
  - 235 clones correspond to 4 genes (so we expect 4 clusters)
- Yeast cell cycle data
  - 17 time points
  - Subset of 384 genes associated with 5 phases of the cell cycle (so we expect 5 clusters)

Synthetic Data Sets (both based on ovarian cancer data)

- Randomly re-sampled ovary data: For each class, randomly sample expression levels in each experiment independently.
  - Preserves the specifics of the distribution (means, variances); destroys covariance structure
- Gaussian mixture: Generate multivariate normal distributions with the sample covariance matrix and mean vector of each class in the ovary data.
  - Preserves the covariance structure; destroys the specifics of the distributions

Results

- Randomly Re-sampled Ovary Data: Both the adjusted Rand and BIC scores favored the diagonal model (model-based clustering) with 4 clusters, as expected.
- Square Root Ovary Data: The adjusted Rand favored EEE (model-based clustering) with 4 clusters; BIC analysis identified EEE and the diagonal model to have local maxima at 4. However, the global maximum indicated the VI model with 8 clusters (but 8 can be split from 4, so this is a sensible solution).
- Standardized Yeast Cell Cycle Data: The adjusted Rand favored EI with 5 clusters and BIC selected EEE with 5 clusters.

c. BIC Scores for Clustering of Alpha-Factor Data with Noise Mixture Models: Adds another component – Given the low probability of outliers, this method allows for minimal distortion of clustering due to outliers.

IV. Summary and Conclusion

- Synthetic Data Sets: Model-based clustering is better than leading heuristic based clustering algorithms.
- Real Data Sets: Adjusted Rand indices are comparable to CAST, but BIC gives a good indication of the number of clusters.