Model-based clustering

Raphael Hoffmann

CSE 527 Lecture Notes, 10/27/03

1 General Idea

We assume that each cluster is generated by a multivariate normal distribution. Thus, our total dataset can be regarded as a mixture of different distributions. We can assign two variables to every cluster k: the distribution's mean vector μ_k and its covariance matrix Σ_k. If we know the number of distributions (=clusters) and their means and (co-)variances, we can calculate the probability that a specific point belongs to a certain cluster. Conversely, if we know the correct cluster for every point, we can calculate the means and variances of the distributions for every cluster. A model based approach then works similar to k-means:

1. Initialize, e.g. by assigning points randomly to clusters
2. Calculate means and covariances for every cluster
3. Calculate probabilities of cluster membership for every point and assign points to clusters with highest probability
4. Goto 2

Note: The Gaussian Mixture Model might be an oversimplification of reality, however it is nevertheless useful.
One problem of the algorithm is how to deal with points that lie in different clusters with roughly equal probability (hard assignnment, probabilistic assignment, no classification?).

2 Mathematical Background

2.1 Variance and Covariance in 2-dimensional space

Variance $var_x = E((x - \mu_x)^2)$
Covariance $cov_{xy} = E((x - \mu_x)(y - \mu_y))$
If x and y are independent, then $cov_{xy} = 0$. The covariance is a measure of the degree to which we can predict one value (e.g. x) if we know the other (e.g. y).

2.2 Multivariate Gaussian distributions

$$f(x) = \frac{1}{\sqrt{(2\pi)^N det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{1}$$

where $x = (x_1, .., x_N)$ and Σ is the covariance matrix, which is defined by

$$\Sigma = \begin{pmatrix} cov_{11} & cov_{12} & \cdots \\ cov_{21} & cov_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

The covariance matrix is symmetric and its diagonal entries are the variances. It can be written as the following composition $\Sigma = \lambda D A D^T$, where
λ is a real, that specifies the volume,
A is a diagonal matrix, that specifies the shape (how ellipsoid?) and
D is a unit matrix, that specifies the orientation of the cloud.

3 Models

The $\Sigma_k = \lambda_k D_k A_k {D_k}^T$ decomposition of the covariance matrix to a cluster k allows us to create simpler models with fewer parameters. For example:

1. Equal Volume Spherical Model (EI): $\Sigma_k = \lambda I$
   Here, all I have to calculate are the means and the common variance λ. Using this model is approximately equivalent to using k-means.

2. Unequal Volume Spherical Model (VI): $\Sigma_k = \lambda_k I$
   The clusters are spherical, but can have different sizes.

3. Diagonal Model: $\Sigma_k = \lambda_k B_k$, where B_k is diagonal, $|B_k| = 1$. The clusters are elliptical, but parallel to the axes.

4. EEE Elliptical Model: $\Sigma_k = \lambda DAD^T$. The clusters are elliptical, but the same covariance structure applies to all clusters.

5. Unconstrained Model (VVV): $\Sigma_k = \lambda_k D_k A_k {D_k}^T$.

4 The Algorithm: Expectation Maximization (EM)

We need an initial assignment of points to clusters (or alternatively initial estimates of the means and variances). After that we iterate between the E and M steps:

- 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

The algorithm can get stuck in a local minimum (use random starts!). However, it is usually not a big problem in practice.

5 Model selection

5.1 The Bayesian Information Criterion (BIC)

With the BIC it is possible to evaluate the odds for one model against another model. More precisely, the BIC expresses the likelihood that our dataset was generated by a given model (Notation: $p(D|M_k)$, where D represents the dataset, and M_k the model). Since models with more parameters will generally "fit" the data better, the BIC also includes a penalty for the number of parameters so that we can at least roughly assess whether the improved fit justifies the increased number of parameters.

$$BIC_k = 2 \log p(D|\hat{\Theta}_k, M_k) - v_k \log(n) \approx 2 \log p(D|M_k)$$

v_k : number of parameters to be estimated in model M_k, and $\hat{\Theta}_k$ is the maximum likelihood estimate of the parameters Θ_k
However, the integrated likelihood $p(D|M_k)$ is hard to evaluate.

5.2 Comparison of model-based clustering and heuristical approaches

Clusters were evaluated by using the BIC and Adjusted Rand index. The Adjusted Rand index compares clusters with external critera. The BIC scores do not require external criteria. The quality of clusters found by model-based clustering were compared to those found by the CAST and k-means algorithms. The used data sets comprised two real datasets

- Ovarian cancer data
- Yeast cell cycle data

and two synthetic data sets

- Gaussian mixture: Multivariate normal distributions with the sample covariance matrix and mean vectors of eachs class in the ovary data were generated. This design destroys the specifics of the distributions, but keeps the covariances.
- Randomly resampled ovary data: Here, the specifics of the distributions (means, variances) were kept, but the covariances ignored.

Since there was not enough data, it was impossible to maintain both, the specifics of the distributions and the covariances.
Randomly resampled data: The Adjusted Rand and the BIC both overwhelmingly favored the diagonal model (model-based clustering).
Real ovary data: The Adjusted Rand favored EEE (model-based clustering) with 4 clusters. Also, the BIC scores of EEE and the diagonal model had a local maximum at 4 clusters.

Standardized yeast cell cycle data: The Adjusted Rand favored EI (model-based clustering) with 5 clusters. BIC selected EEE at 5 clusters.

In general, on the synthetic data sets, model-based clustering was better than leading heuristic based clustering algorithms. On real data sets, the Adjusted Rand indices were comparable to those of CAST, with the additional advantage that BIC gave some indication of an appropriate number of clusters.