Model Selection for Clustering Decision Tree

Yale Chang

1 Problem

Given dataset $X \in \mathbb{R}^{n \times d}$, where $n$ is the number of samples, $d$ is the number of features, the goal of clustering is to group $n$ samples into $K$ distinct categories (clusters) so that samples in the same cluster are similar to each other while samples in different clusters are dissimilar. There are three problems involved in specifying a clustering algorithm: objective function, searching strategy, model selection.

2 Objective Function

Based on the intuition that samples in the same cluster should be similar, different clustering objectives can be defined.

2.1 K-means

The objective of spectral clustering is

$$
\min_Y \sum_{k=1}^{K} \sum_{i : y_i = k} ||x_i - \mu_k||^2_2
$$

where $\mu_k = \frac{1}{n_k} \sum_{i : y_i = k} x_i$ is the centroid of $k$-th cluster.

2.2 Hierarchical Clustering

The objective of hierarchical or agglomerative clustering is [1]

$$
\min_{\alpha \in \mathbb{R}^{n \times d}} \frac{1}{2} ||\alpha - X||_F^2
$$

subject to $\sum_{i < j} I_{\alpha_i \neq \alpha_j} \leq t$.

2.3 Spectral Clustering

Given similarity matrix $S \in \mathbb{R}^{n \times n}$ and cluster number $K$, the objective of spectral clustering is

$$
\min_{U \in \mathbb{R}^{n \times K}} \text{Tr} \left( U^T D^{-1/2} (I - S) D^{-1/2} U \right)
$$

subject to $U^T U = I_{K \times K}$.
2.4 Information Theoretic Clustering

The objective of information-theoretic clustering could be one of the following:

- Maximizing mutual information
  \[
  \max_Y MI(X,Y) \quad (4)
  \]

- Minimizing conditional entropy \( H(Y|X) \)
  \[
  \min_Y H(Y|X) \quad (5)
  \]

- Minimizing conditional entropy \( H(X|Y) \)
  \[
  \min_Y H(X|Y) \quad (6)
  \]

2.5 Gaussian Mixture Model (GMM)

Follow maximum likelihood principle

\[
\max_{Z,\pi,\mu,\Sigma} \log p(X|\pi,\mu,\Sigma) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \quad (7)
\]

3 Searching Strategy

After formulating clustering as an optimization problem, it is necessary to design a searching strategy to find the optimal solution that optimizes the objective function. For example, K-means algorithm iterates between assigning samples to cluster centroids and updating cluster centroids. Hierarchical clustering takes a greedy strategy by building clusters in a bottom-up way. Spectral clustering finds the global optimum by solving eigen-decomposition of the normalized Laplacian matrix. GMM utilizes expectation maximization (EM) algorithm, which iterates between computing expectation of latent assignment variable \( Z \) and model parameter \( \theta_k (k = 1, \cdots, K) \), to find a local optimum.

4 Model Selection

In the discussion above, we assume the number of clusters are known as a prior, which is not necessarily true in solving practical problems. Model selection for clustering aims to automatically find the correct number of clusters from data. We will review a few existing model selection approaches.

4.1 GAP Statistic

GAP statistic [2] aims to find the right \( K \) for K-means and hierarchical clustering methods by detecting elbow in dispersion vs. cluster number plot.

4.2 Eigen Gap

Eigen gap [3] could be used to find the right \( K \) for spectral clustering.
4.3 Akaike information criterion (AIC)

\[ AIC = 2m - 2\log(L) \]  \hspace{1cm} (8)

where \( L \) is the maximized value of the likelihood function for the model and \( m \) is the number of estimated parameters in the model. The best model is the one that achieves the smallest AIC value.

4.4 Bayesian information criterion (BIC)

\[ BIC = m \cdot \log(n) - 2\log(L) \]  \hspace{1cm} (9)

4.5 Bayes Factor

Given observed data \( D \), the plausibility of the two different models \( M_1 \) and \( M_2 \), parametrised by model vectors \( \theta_1 \) and \( \theta_2 \) is assessed by the Bayes factor \( K \) given by

\[ K = \frac{p(D|M_1)}{p(D|M_2)} = \frac{\int p(\theta_1|M_1)p(D|\theta_1, M_1)d\theta_1}{\int p(\theta_2|M_2)p(D|\theta_2, M_2)d\theta_2} \]  \hspace{1cm} (10)

4.6 Minimum Message Length (MML)

MML [4] specifies that the minimal encoding of the model and the objects given the model is the best. For each different model we can calculate the total encoding length. In terms of Bayes theorem, we wish to maximize the posterior distribution \( p(M|D, \theta) = \frac{\pi(M|\theta)p(D|M, \theta)}{p(D)} \), take the logarithm of the expression above

\[ -\log p(M|D, \theta) = -\log p(M|\theta) - \log p(D|M, \theta) + \text{const} \]  \hspace{1cm} (11)

According to Shannon coding theorem, \( -\log p(M|\theta) \) is the minimum length in bits to encode the model and \( -\log p(D|M, \theta) \) is the minimum length in bits to encode the objects given the model. The MML criterion only defines a goodness measure for a model with an inherent bias towards simple models. It does not indicate how to search the model space.

4.7 Minimum Description Length (MDL)

MDL [5] is similar to MML in that they both try to capture the principle of Occam’s razor, i.e., prefer simpler models to explain the data. MDL and MML both choose the hypothesis minimizing the code length of the data. However, there are a few differences between them. First, MML always uses two-part codes, so that MML automatically selects both a model family and parameter values. Second, while the MDL codes such as \( P_{\text{min}} \) minimize worst-case relative code length (regret), the two-part codes used by MML are designed to minimize expected absolute code length. Here the expectation is taken over a subjective prior distribution defined on the collection of models and parameters under consideration. Indeed, Wallace and his co-workers stress that their approach is fully (subjective) Bayesian.
5  MDL-based pruning for decision tree clustering

Objective function: minimizing $H(Y|X)$ in information-theoretic clustering;
Searching strategy: decision tree construction, each leaf represents a cluster;
Model selection: MDL

In the MDL framework, given dataset $X$, we need to find a hypothesis $H$ such that $L(H) + L(X|H)$ is minimized.

References


