Lecture 8: Clustering and Typical Methods

Shukai Gong

1 K-means Clustering

A sample should be closer to the centroid of its cluster than to the centroid of other clusters. In this sense, a heuristic K-means is realized by

Classic K-means

We find the centroids of the clusters in a heuristic way:

1. Initialize K centroids $\{\boldsymbol{c_k}\}_{k=1}^{K}$ randomly from the observed data.
2. Repeat the following steps until convergence:
   (a) Assign each data point to the nearest centroid $\mathcal{C}$: $\forall \boldsymbol{x_n}, \boldsymbol{x_n} \in \mathcal{C}_k$ if $k = \arg\min_{j\in\{1,\cdots,K\}} d(\boldsymbol{x_n}, \boldsymbol{c_j})$.
   (b) Update the centroids by averaging samples within the cluster: $\boldsymbol{c_k} = \frac{1}{|\mathcal{C}_k|}\sum_{\boldsymbol{x_n}\in\mathcal{C}_k} \boldsymbol{x_n}$. (**This is essentially the barycenter of the cluster**)

2 Spectral Clustering

Classic K-means Clustering has several drawbacks:

- It doesn't work well for linearly inseperable data;
- It faces curse of dimensionality (as it primarily relies on Euclidean distance).

Spectral clustering is introduced to overcome this.

Spectral Clustering

We first use the spectrum (eigenvalues) of the **similarily matrix** $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ of the data $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \in \mathbb{R}^{N\times D}$ to perform **dimensionality reduction** and then apply K-means on the reduced space.

Here $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is the similarity matrix of the data $\boldsymbol{X}$, where $a(\boldsymbol{x_i}, \boldsymbol{x_j}) = a_{ij} \in [0, 1]$ is the similarity between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. The spectrum of $\boldsymbol{A}$ is obtained by EVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \sum_{n=1}^{N} \lambda_n \boldsymbol{u_n}\boldsymbol{u_n}^\top$

Steps of Spectral Clustering

Given a set of data $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \in \mathbb{R}^{N\times D}$

1. Construct the similarity matrix $\boldsymbol{A} = [a(\boldsymbol{x_i}, \boldsymbol{x_j})] \in \mathbb{R}^{N\times N}$.
2. Compute the Laplacian matrix $\boldsymbol{L}$ of $\boldsymbol{A}$: $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{A}\boldsymbol{1_N}) - \boldsymbol{A}$.
3. Normalize the Laplacian matrix $\boldsymbol{L}$: $\boldsymbol{L}_{\mathrm{norm}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$ where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{A}\boldsymbol{1_N})$.
4. Conduct EVD on the Laplacian matrix $\boldsymbol{L}_{\mathrm{norm}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, $0 = \lambda_1 \le \cdots \le \lambda_N$.
5. Truncate the eigenvectors of $\boldsymbol{U}$ to get $\boldsymbol{U}_L \in \mathbb{R}^{N\times L}$.
6. Apply K-means on the rows of $\boldsymbol{U}_L$ to get the clustering.

In this sense, spectrum clustering is essentially a two-step process: (1) **Laplacian Eigenmap**; (2) K-means.

3 Evaluation of Clustering

Measurements using external ground truth information are called *external evaluation*, while those not relying on the ground truth are called *internal evaluation*.

3.1 When Ground Truth is Available

Purity

Denote $\Omega = \{\omega_1, \cdots, \omega_K\}$ as the set of K clusters where each ω_k contains the indices of the samples in the k-th cluster. Denote $\mathcal{C} = \{c_1, \cdots, c_J\}$ as the J classes (ground truth)where each c_j contains the indices of the samples in the j-th class. The purity is computed by assigning each cluster to the class which is most frequent in the cluster, and calculating the averaged accuracy of the assignment,

$$\text{Purity}(\Omega, \mathcal{C}) = \frac{1}{N} \sum_{k=1}^{K} \max_{j \in \{1, \cdots, J\}} |\omega_k \cap c_j|$$

where N is the number of samples.

[Drawback]: When the number of clusters K is much larger than the number of classes J, the purity may be high even if the clustering is bad.

Normalized Mutual Information (NMI)

Denote

- $P(\omega_k) = \dfrac{|\omega_k|}{N}$ and $P(c_j) = \dfrac{|c_j|}{N}$ as the probability of a sample belonging to the k-th cluster.
- $P(c_j) = \dfrac{|c_j|}{N}$ as the probability of a sample belonging to the j-th class.
- $P(\omega_k \cap c_j) = \dfrac{|\omega_k \cap c_j|}{N}$ as the probability of a sample belonging to both the k-th cluster and the j-th class.

The NMI is defined as

$$\text{NMI}(\Omega, \mathcal{C}) = \frac{2I(\Omega, \mathcal{C})}{H(\Omega) + H(\mathcal{C})}$$

where

- $I(\Omega, \mathcal{C}) = \sum\limits_{k,j} P(\omega_k \cap c_j) \log \dfrac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)}$ is the mutual information between Ω and $\mathcal{C}$
- $H(\Omega) = -\sum\limits_{k} P(\omega_k) \log P(\omega_k)$ and $H(\mathcal{C}) = -\sum\limits_{j} P(c_j) \log P(c_j)$ are the entropies of Ω and $\mathcal{C}$.

[Note]: NMI achieves a tradeoff between the quality of the clustering and the number of clusters.

Rand Index (RI)

Rand Index describes the percentage of pairwise decision correctness. Consider $\frac{N(N-1)}{2}$ pairs of samples,

- **True Positive (TP)**: The percentage of the paired samples in the same cluster and the same class.
- **True Negative (TN)**: The percentage of the paired samples in the different cluster and the different class.
- **False Positive (FP)**: The percentage of the paired samples in the same cluster and the different class.
- **False Negative (FN)**: The percentage of the paired samples in the different cluster and the same class.

and Rand Index is defined as

$$\text{RI} = \frac{TP + TN}{TP + TN + FP + FN}$$

Some other measurements are listed below:

- **Precision, Recall, F1 Score:**

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1 Score} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Jaccard Index:** $\text{JI} = \frac{TP}{TP + FP + FN}$.
- **Dice Index:** $\text{DI} = \frac{2TP}{2TP + FP + FN}$.
- **Fowlkes-Mallows Index:** $\text{FMI} = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} = \sqrt{\text{Precision} \times \text{Recall}}$.

3.2 When Ground Truth is Unavailable

Davies-Bouldin Index

Denote $\boldsymbol{c}_i$ as the centroid of the i-th cluster, and $s\sigma_i$ as the average distance of all the samples in the i-th cluster to the centroid $\boldsymbol{c}_i$. The Davies-Bouldin Index is defined as

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(\boldsymbol{c}_i, \boldsymbol{c}_j)} \right)$$

The principal of DBI is to encourage low intra-cluster distances and high inter-cluster distances. $\sigma_i + \sigma_j$ measures the compactness of the clusters, while $d(\boldsymbol{c}_i, \boldsymbol{c}_j)$ measures the separation between the clusters. In general, the smaller the DBI, the better the clustering.

Dunn Index

Dunn Index aims at identifying dense and well-separated clusters, it's defined as the ratio between the

minimal inter-cluster distance to the maximal intra-cluster distance,

$$\mathrm{DI} = \frac{\min\limits_{1\le i<j\le K} d(\boldsymbol{c}_i, \boldsymbol{c}_j)}{\max\limits_{1\le i\le K, \boldsymbol{x}_n\in\mathcal{C}_i} d(\boldsymbol{c}_i, \boldsymbol{x}_n)}$$

Silhouette

Given the i-th cluster $\mathcal{C}_i$ and the j-th sample $\boldsymbol{x}_j \in \mathcal{C}_i$, the averaged distance of the sample to other samples in the same cluster is

$$a(\boldsymbol{x}_j) = \frac{1}{|\mathcal{C}_i| - 1} \sum_{\boldsymbol{x}_k\in\mathcal{C}_i, \boldsymbol{x}_k\neq\boldsymbol{x}_j} d(\boldsymbol{x}_j, \boldsymbol{x}_k)$$

The smallest averaged distance to the samples in **other clusters** is

$$b(\boldsymbol{x}_j) = \min_{k\in\{1,\cdots,K\}, k\neq i} \frac{1}{|\mathcal{C}_k|} \sum_{\boldsymbol{x}_k\in\mathcal{C}_k} d(\boldsymbol{x}_j, \boldsymbol{x}_k)$$

The silhouette of $\boldsymbol{x}_j$ is defined as

$$s(\boldsymbol{x}_j) = \frac{b(\boldsymbol{x}_j) - a(\boldsymbol{x}_j)}{\max\{a(\boldsymbol{x}_j), b(\boldsymbol{x}_j)\}} = \begin{cases} 1 - \dfrac{a(\boldsymbol{x}_j)}{b(\boldsymbol{x}_j)} & a(\boldsymbol{x}_j) < b(\boldsymbol{x}_j) \\ 0 & a(\boldsymbol{x}_j) = b(\boldsymbol{x}_j) \\ \dfrac{b(\boldsymbol{x}_j)}{a(\boldsymbol{x}_j)} - 1 & a(\boldsymbol{x}_j) > b(\boldsymbol{x}_j) \end{cases}$$

Setting the number of clusters as k, the averaged silhouette value of all the data points measures the **tightness** of the clusters.

$$\overline{s_k} = \frac{1}{N} \sum_{j=1}^{N} s(\boldsymbol{x}_j)$$

Setting the number of clusters from 1 to K, the silhouette coefficient is defined as

$$\mathrm{SC} = \max_{k\in\{1,\cdots,K\}} \overline{s_k}$$

References

- CSC 311: Introduction to Machine Learning — Lecture 11 k-Means and EM Algorithm