Data Mining: Lecture 3

Cluster Analysis II

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology

13.09.2022

Affinity Propagation

- Clustering algorithm based on the idea of "message passing" between data point. Proposed in 2007 by B.J. Fray et.al.
- Finds "exemplars" (cluster representatives) in iterative way.
- Does not require number of clusters as hyperparameter.

Affinity Propagation: Algorithm I

- Let $X = \{x_1, \ldots x_n\}$ be the set of data points.
- We will use **negative** squared distance to determine the similarity.

$$s(x_i, x_k) = -||x_i - x_k||^2$$

- Compute the similarity matrix S which elements are pairwise distance between all the elements of X

Affinity Propagation: Algorithm II

- Initialize (with zeros) two $n \times n$ matrices; R - the *responsibility* matrix and A - the *availability matrix*.
- This matrices may be used as $log - probability$ matrices or tables.
- The elements of the matrices will be denoted as $r(i,k)$ and $a(i,k)$
  - $r(i,k)$ describe (quantify) how well x_k would suite x_i as exemplar (representative) and will be update on each iteration as follows:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\}.$$

  - $a(i,k)$ describe (quantify) how well it would suite x_i to chose x_k as exemplar (representative) and will be update on each iteration as follows:

$$a(i,k) \leftarrow s(i,k) - \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\left(0, r_{i',k}\right)\right) \quad i \neq k$$

- Stop once the convergence criteria is satisfied. Exemplars are the positive elements of the main diagonal of the matrix $C = R + A$.

Affinity propagation example

1. Initial data

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Point 1	3	4	3	2	1
Point 2	4	3	5	1	1
Point 3	3	5	3	3	3
Point 4	2	1	3	3	2
Point 5	1	1	3	2	3

2. Similarity matrix

	Point 1	Point2	Point 3	Point 4	Point 5
Point 1	0 → -22	-7	-6	-12	-12
Point 2	-7	0 → -22	-17	-17	-22
Point 3	-6	-17	0→ -22	-18	-21
Point 4	-12	-17	-18	0→ -22	-3
Point 5	-17	-22	-21	-3	0→ -22

3. Responsibility matrix

	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	-16	-1	1	-6	-11
Point 2	10	-15	-10	-10	-15
Point 3	11	-11	-16	-12	-15
Point 4	-9	-14	-15	-19	9
Point 5	-14	-19	-18	14	-19

4. Availability matrix

	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	21	-15	-16	-5	-10
Point 2	-5	0	-15	-5	-10
Point 3	-6	-15	1	-5	-10
Point 4	0	-15	-15	14	-19
Point 5	0	-15	-15	-19	-19

5. Criterion matrix

	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	5	-16	-15	-11	-21
Point 2	5	-15	-25	-15	-25
Point 3	5	-26	-15	-17	-25
Point 4	-9	-29	-30	-5	-10
Point 5	-14	-34	-33	-5	-10

Affinity Propagation: Interpretations

- Matrices R and A are considered as real valued messages.
- Possible convergence/stopping criteria are: Limit on the number of iterations, cluster assignments remain unchanged for a given number of iterations etc.

Mean shift

- Non parametric unsupervised technique.
- Does not depend on the cluster shape.
- Idea is to create a path for each point shifting it towards local density maximum.
- Requires one hyper parameter R, which is referred as *bandwidth* and represent the radius of the neighbourhood.

Mean shift: Algorithm I

- Repeat until converge:
  - For each point compute mean or weighted mean of its R - neighborhood.
  - Associate original point with the mean. (Repeat the previous step for all the mean points). This would create a path for each point.
- Mean points will converge to a limited number of regions. Path of each point will lead to one region or ideally a point.
- Each region represent the cluster. Use path of each point to get its cluster label.
- Return cluster labels.

Mean shift: Mean calculation

- Mean
- Weighted mean. Weight may be calculated as follows:
  - $$w(d) = \begin{cases} 1, & \text{if} \quad d \leq R \\ 0, & \text{if} \quad d > R \end{cases}$$
  - Or Gaussian kernel
    $$w(d) = e^{-\frac{d}{2\sigma^2}}$$
- Other kernels are possible

Spectral clustering I

- Inspired by mass-spring system. Each mass is associated with an observation point and spring with the similarity between two points.
- Based on the analysis of spectrum (eigenvalues) of the similarity matrix.
- Compute the similarity matrix S whose elements $s_{i,j}$ and define diagonal matrix D as follows:

$$D = \sum_j s_{i,j}$$

- Compute Laplacian matrix of S as follows:
  -

$$L = D - S$$

  - **NOTE!!!** Alternative definition of the Laplacian exists, for example, D may be defined as *degree* matrix, leading:

$$L_{i,j} = \begin{cases} d_i & \text{if} \quad i = j \\ -s_{i,j} & \text{if} s_{i,j} > \epsilon \\ 0 & \text{if} s_{i,j} \leq \epsilon \end{cases}$$

  here ε plays the role of a threshold.

Spectral clustering II

- Normalize laplacian matrix. Different ways are possible:
  - $$L^{\text{norm}} = I - D^{-1/2} S D^{-1/2}$$
  - Denote the eigenvector corresponding to the second smallest eigenvalue by v
  - Left normalized Lapalacian (random walk):

    $$L^{\text{rw}} = D^{-1} L = I - D^{-1} S$$

    denote u the eigenvector corresponding to the largest eigenvalue of the random walk normalized adjacency matrix $D^{-1}S$.
- Note that $D^{-1/2} v = u$

Spectral clustering III

- For binary clustering problem: Positive elements of v indicate points belonging to one class and negative to another.
- For arbitrary number of clusters:
  - Compute k eigenvectors corresponding to k smallest eigenvalues.
  - Stuck the vectors into clusters and treat it as similarity matrix.
  - Apply K-means to produce clusters.
- Return labels.

BIRCH I

- Balanced Iterative Reducing and Clustering using Hierarchies.
- Idea is to create a compact summary of the data set and then perform clustering on it instead the entire data set.

Definition

Let X be and n - dimensional space. Clustering feature of the cluster X_i is defined as triple $\mathrm{CF} = \{N, \mathrm{LS}, \mathrm{SS}\}$. Where N is the number of points in the cluster. LS is the linear sum of the points in the cluster and SS sum of squares of points in the cluster.

BIRCH II

- CF- tree is the tree representation of the data set.
- Nodes are limited by the number of elements they contain.
- Diameter of each leaf has to be less than a threshold.
- For each point, start from the root and choose the child branch most appropriate for the given point. Continue until the leaf node.
- At the leaf node verify if given point may be absorbed without violating threshold conditions.
- If absorbtion is not possible split the node by choosing two most distant points as seeds.
- Once no points is left, return clustering labels.

Wards linkage

- Agglomerative clustering techniques.
- The clusters to be merged are selected based on minimizing the increase in total within cluster variance.
- Squared Euclidean distance is used between the points

Clustering quality evaluation

Note! in real life problems ground truth information is not available!!!. Rare exceptions are possible. Nevertheless, when testing new clustering techniques testing on matchmarking data sets is required and then ground truth information is available.

Rand index

Ground truth information is required. No assumptions about the clustering structure is required. Measures the similarity between two assignments.

- Let C be the ground truth clustering assignment and K assignment produced by the algorithm: then

$$\mathrm{RI} = \frac{a+b}{C_2^n}$$

  where n is number of samples.
- May not be usable in the case of large numbers of clusters.

$$\mathrm{ARI} = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

Mutual information based score

Measures the agreement between ground truth and algorithm assignments.

- Let U and V are the clustering label assignments of N observation points.
- Mutual information is defined as follows:

$$\mathrm{MI(U,V)} = \sum_{\mathrm{i=1}}^{\mathrm{|U|}} \sum_{\mathrm{j=1}}^{\mathrm{|V|}} \mathrm{P(i,j)} \frac{\mathrm{|U_i \cup V_j|}}{\mathrm{N}} \log\left(\frac{\mathrm{P(i,j)}}{\mathrm{P(i)P'(j)}}\right)$$

- Normalized mutual information:

$$\begin{aligned} \mathrm{NMI}(U,V) &= \frac{\mathrm{MI(U,V)}}{(H(U),H(V))} \\ H(U) &= -\sum_{i=1}^{|U|} P(i)\log(P(i)) \\ H(V) &= -\sum_{j=1}^{|V|} P'(j)\log(P'(j)) \end{aligned}$$

where $P(i) = |U_i|/N$ and $P'(j) = |V_j|/N$

Homogenity and comleteness

- Define entropy of the clusters and conditional entropy of the clusters given ground truth data.

$$
\begin{aligned}
H(C) &= -\sum_{c=1}^{|C|} \frac{n_c}{n} \log \frac{n_c}{n} \\
H(C|K) &= -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{k} \log \left(\frac{n_{c,k}}{n_k} \right)
\end{aligned}
$$

- Homogeneity and completeness scores are defined as follows:

$$
\begin{aligned}
h &= 1 - \frac{H(C|K)}{H(C)} \\
c &= 1 - \frac{H(C|K)}{H(K)}
\end{aligned}
$$

V - measure

In many applications harmonic mean of homogeneity and completeness may be useful. It is referred as v-measure and defined as:

$$v = 2\frac{hc}{h+c}$$

Fowlkes-Mallows scores

Is based on the notions of precision and recall (borrowed from the classification quality assessment) and defined as geometric mean of the pairwise precision and recall.

$$FMI = \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})}}$$

Here TP stands for true positive, FN - false negative, and FP - false positive.

Calinski-Harabasz

- Does not require ground truth data.
- Define between group and within-cluster dispertion matrices.

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c - q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q)^T - c_e$$

where c_q is the center of cluster C_q, n_q is the number of points in cluster C_q

- Calinski-Harabasz index is defined as follows:

$$s = \frac{\mathrm{tr(B_k)}}{\mathrm{tr(W_k)}} \times \frac{n_E - k}{k - 1}$$

where n_E is the power of the data set to be clustered.

Davies-Bouldin

- Does not require ground truth data.
- Denote s_i the average distance between the points and centroid of the corresponding cluster.
- Denote $d_{i,j}$ distance between the centroid of clusters i and j.
- Define the similarity as $R_{i,j} = (s_i + s_j)/d_{ij}$.
- Davies - Bouldin index is then defined as follows:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{i,j}$$

Cluster Purity. NB! Not unsupervised any more!!!

- Let m_{ij} represent the number of data points from class (ground-truth cluster) i that are mapped to (algorithm determined) cluster j.
- Denote number of data points in true cluster i are by N_i,the number of data points in algorithm-determined cluster j by M_j.

$$N_i = \sum_{j=1}^{k_d} m_{ij}; \qquad M_j = \sum_{i=1}^{k_t} m_{ij};$$

- For a given algorithm-determined cluster j, the number of data points P_j in its dominant class is: $P_j = \max_i m_{ij}$.
- Purity index is defined

$$P_a = \frac{\sum_{j=1}^{k_d} P_j}{\sum_{j=1}^{k_d} M_j}.$$

Gini index

- Gini index for algorithm determined cluster j is defined:

$$G_j = 1 - \sum_{i=1}^{k_t} \left(\frac{m_{ij}}{M_j} \right)^2 .$$

- Average Gini index is defined as follows:

$$G = \frac{\sum_{j=1}^{k_d} G_j M_j}{\sum_{j=1}^{k_d} M_j}.$$

Computational exercises:

- program your own implementation of any clustering techniques discussed today.
- program your own implementation of any clustering performance metrics discussed today.