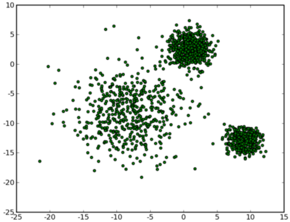
K-means 聚类算法

王石平 博士

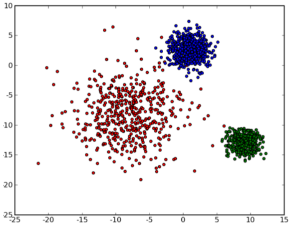
Outline

- Motivation
- K-means algorithm
- Face image clustering

An example

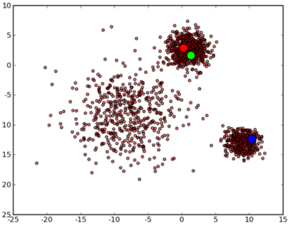


(a) Original data

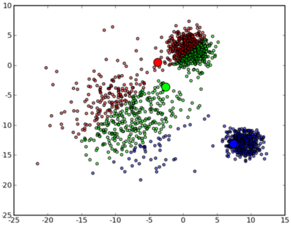


(b) Semantic clustering

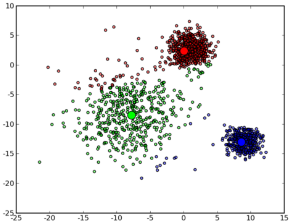
Iteration process



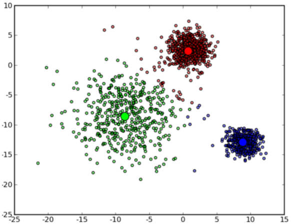
(c) 1th-iteration



(d) 2th-iteration



(e) 3th-iteration



(f) 4th-iteration

Algorithm 1 K-means algorithm

Input: data samples $X = \{x_1, \cdots, x_n\}$ and the number of clusters k.

Output: Clustering results of all samples.

```
Randomly generate k cluster centroids μ_1, ··· , μ_k;
repeat
    Assign each sample to the cluster having smallest distance
    to the corresponding cluster centeroid;
    Update cluster centroids with the means of clusters;
until convergence
return  Clustering result.
```

Algorithm 2 K-means algorithm

Input: data samples $X = \{x_1, \cdots, x_n\}$ where $x_i \in \mathbb{R}^d$ and the number of clusters k.

Output: Clustering results of all samples.

```
Initialize k cluster centroids μ_1, ..., μ_k where μ_i ∈ R^d;
Initialize clustering indicator array cluster ∈ R^n;
repeat
    Set cluster_i = argmin_j ||x_i − μ_j||;
    Update μ_j = (Σ_{i=1}^{n} 1(cluster_i=j) x_i) / (Σ_{i=1}^{n} 1(cluster_i=j));
until convergence
return cluster.
```

Initialize k cluster centroids $\mu_1, \cdots, \mu_k$ where $\mu_i \in \mathbb{R}^d$;

Initialize clustering indicator array $cluster \in \mathbb{R}^n$;

repeat

  Set $cluster_i = \arg\min_j ||x_i - \mu_j||$;

  Update $\mu_j = \frac{\sum_{i=1}^{n} 1(cluster_i = j) x_i}{\sum_{i=1}^{n} 1(cluster_i = j)}$;

until convergence

return $cluster$.

Question

- Objective function?
- Initialization sensitivity?
- Convergence?
- Performance evaluation?

Objective function of K-means clustering

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} ||x_i - \mu_j||$$

where $r_{ij} = 1$ if x_i is assigned to cluster j; otherwise, $r_{ij} = 0$. 计算可得

$$\mu_j = \frac{\sum\limits_{i=1}^{n} r_{ij} x_i}{\sum\limits_{i=1}^{n} r_{ij}}.$$

Clustering accuracy

Denote p_i and q_i be the obtained clustering label and the label provided by the dataset, respectively. The ACC is defined as follows

$$ACC = \frac{\sum_{i=1}^{n} \delta(p_i, map(q_i))}{n} \tag{1}$$

where $\delta(a, b) = 1$ if $a = b$; otherwise $\delta(a, b) = 0$. $map(\bullet)$ is the best permutation mapping function that matches the obtained clustering label to the equivalent label of the dataset.

Normalized mutual information

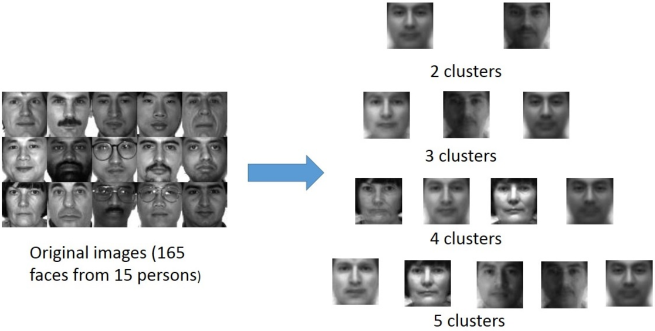
Given two random variables P and Q, NMI of P and Q is defined as

$$NMI(P,Q) = \frac{I(P;Q)}{\sqrt{H(P)H(Q)}} \tag{2}$$

where $I(P;Q)$ is the mutual information of P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q, respectively. Here, the clustering results $\widetilde{C} = \{\widetilde{C}_i\}_{i=1}^{\widetilde{c}}$ of and the ground truth labels $C = \{C_j\}_{j=1}^{c}$ of all samples are viewed as two discrete random variables. NMI is specified as

$$NMI(C,\widetilde{C}) = \frac{\sum_{i=1}^{\widetilde{c}} \sum_{j=1}^{c} |\widetilde{C}_i \cap C_j| log \frac{n|\widetilde{C}_i \cap C_j|}{|\widetilde{C}_i||C_j|}}{\sqrt{(\sum_{i=1}^{\widetilde{c}} |\widetilde{C}_i| log \frac{|\widetilde{C}_i|}{n})(\sum_{j=1}^{c} |C_j| log \frac{|C_j|}{n})}}. \tag{3}$$

K-means clustering result from face images



Thanks!