Model-based clustering

Machine Learning II 2019

2.1 From probabilities to distance conversion

Consider the two-phase meme generation procedure specified in the tutorial:

- For a minor variation $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_{s_i}, \sigma_1 \mathbf{I})$ and all choices are independent.
- For an abrupt change $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_{s_i}, \sigma_2 \mathbf{I})$ and all choices are independent.
- Probability of abrupt changes $\Pr[c_i = 1] = \varrho$ and all choices are independent.

Express log-likelihood $\log \Pr[\boldsymbol{x}_i \wedge c_i | s_i]$ and determine for which value c^* of c_i the value is maximised.

Solution

The log-likelihood is

$$\log \Pr[\boldsymbol{x}_i \wedge c_i | s_i] = \log \Pr[\boldsymbol{x}_i | c_i, s_i] + \log \Pr[c_i | s_i].$$

For a minor change ($c_i = 0$), this results in

$$p_0 = -\log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(x_i - x_{s_i})^T(x_i - x_{s_i}) + \log(1 - \varrho);$$

for an abrupt change ($c_i = 1$), this results in

$$p_1 = -\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(x_i - x_{s_i})^T(x_i - x_{s_i}) + \log \varrho.$$

The value c^* is the c_i which results in the bigger log-likelihood:

$$c^* = \arg\max_{c_i \in \{0,1\}} p_{c_i}.$$

3.1 - 3.3 Implementation of k-means and its advancements

The likelihood that $\boldsymbol{x}_i$ is generated by the cluster j can be expressed as

$$\Pr[\boldsymbol{x_i} | z_i = j, \boldsymbol{\Theta}] = \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}_j}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)\right)$$

where the covariance matrix $\boldsymbol{\Sigma}$ fixes the shape and scale of the cluster.

Implement the algorithm on `challenge-1.csv`, `challenge-2.csv` and `challenge-3.csv`. After you know a good estimate for the labels and cluster centres, you can find the mixture probabilities

$\lambda_j = \Pr[z = j]$ as fractions of corresponding cluster labels and test the applicability of the model by generating the same amount of data and comparing it visually with the original data. Do you believe that the model adequately describes the data?

Solution

The optimization algorithm consists of iteratively doing the following two steps until convergence:

Step 1. For each $\boldsymbol{x}_i$, find the label z_i such that

$$z_i = \underset{j \in \{1,\dots,k\}}{\arg\max} \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}_j}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)\right). \tag{1}$$

If $\boldsymbol{\Sigma}_j = \sigma_j \mathrm{I}$ and log-likelihood is used instead of likelihood, this simplifies to

$$z_i = \underset{j \in \{1,\dots,k\}}{\arg\min} \left(2\log\sigma_j + \frac{1}{2\sigma_j^2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_j)\right). \tag{2}$$

If $\forall j \in \{1, ..., k\} : \sigma_j = \sigma$ then this further simplifies to

$$z_i = \underset{j \in \{1,\dots,k\}}{\arg\min} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_j). \tag{3}$$

Step 2. For each $j \in \{1, ..., k\}$, find the maximum likelihood parameters $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$:

$$(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \underset{(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\arg\max} \prod_{i:z_i=j} \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right).$$

Simple k-means

The simple k-means algorithm where labels are given by equation 3 works well for `challenge-1.csv` (**Figure** 1). In `challenge-2.csv`, though it can correctly perform clustering, it fails to model the clusters adequately (**Figure** 2).

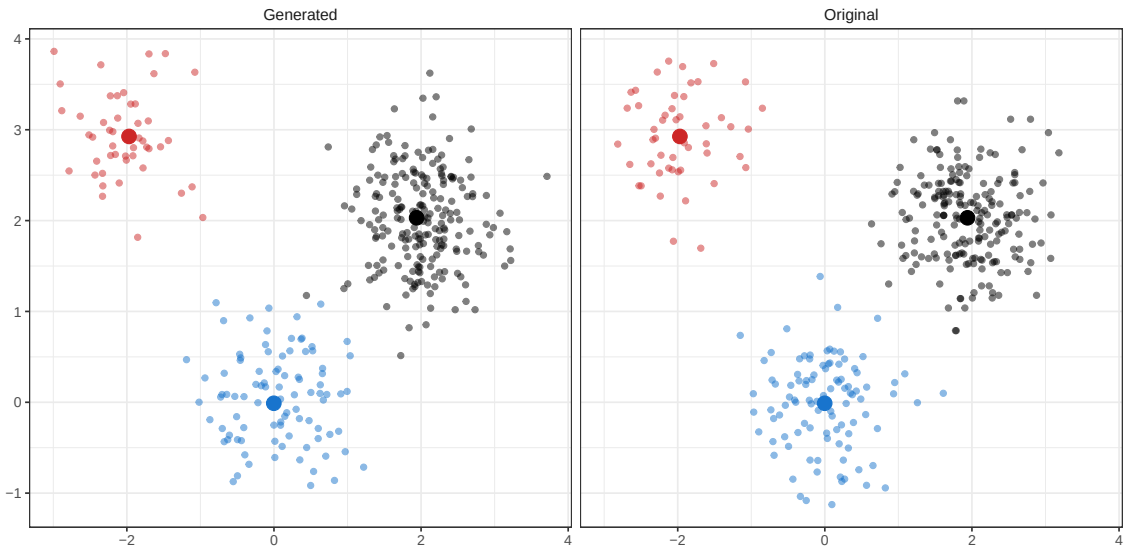

Figure 1: K-means with data *challenge-1.csv*. The model adequately describes the data.

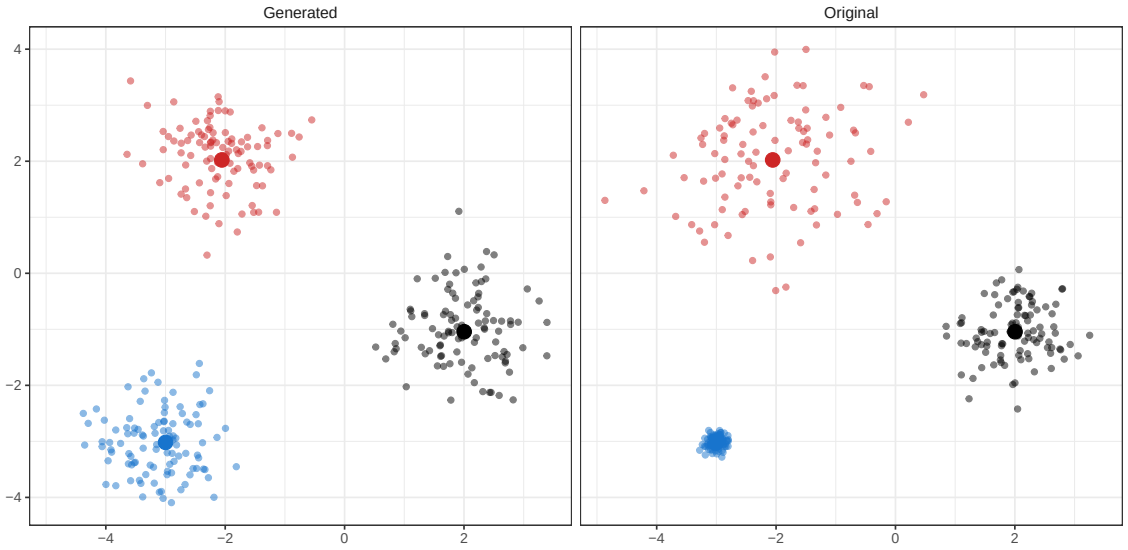

Figure 2: K-means with data *challenge-2.csv*. The model doesn't adequately describe the data as the variances of the generated clusters do not match the variances of the real clusters.

K-means+

The modified k-means (k-means+) where labels are given by equation 2 works well for `challenge-2.csv` (**Figure** 3). In `challenge-3.csv`, though it can decently perform clustering, it fails to model the clusters adequately (**Figure** 4).

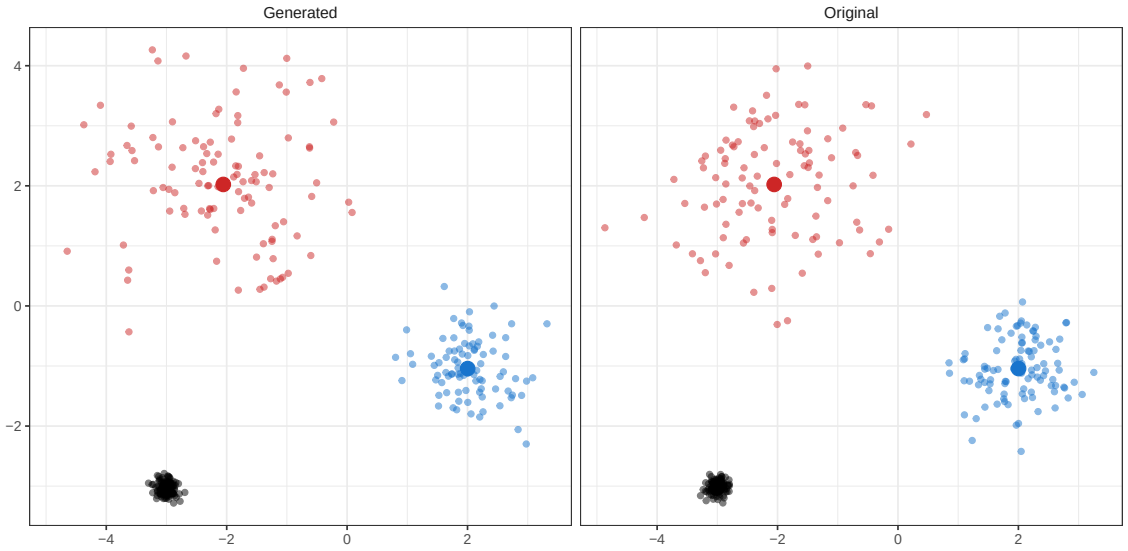

Figure 3: K-means+ with data *challenge-2.csv*. The model adequately describes the data.

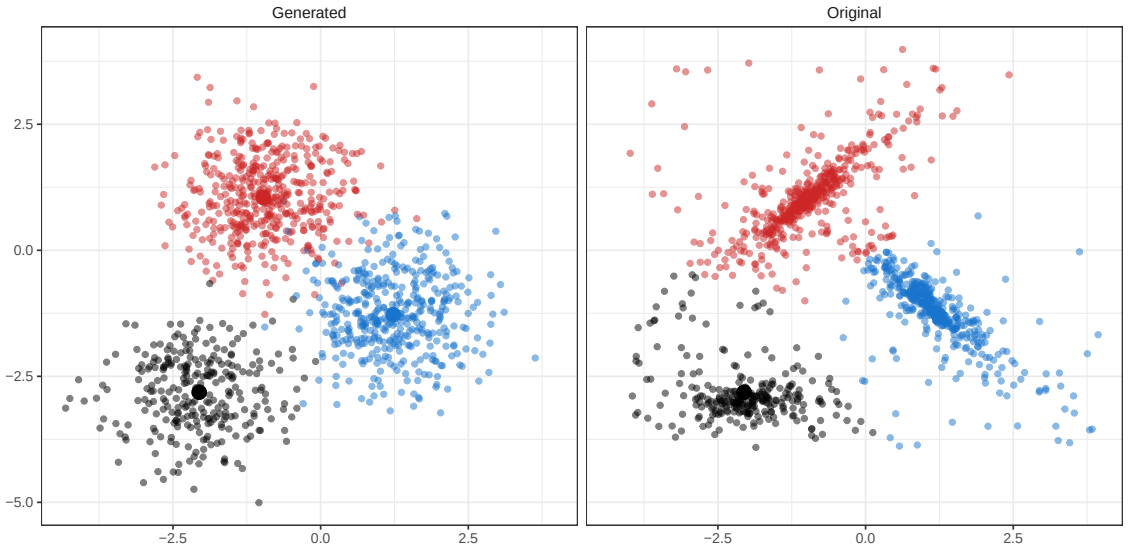

Figure 4: K-means+ with data *challenge-3.csv*. The model doesn't adequately describe the data as the shapes of the generated clusters do not match the shapes of the real clusters.

K-means++

The advanced k-means (k-means++) where labels are given by equation 1 works reasonably well even for `challenge-3.csv` (**Figure** 5).

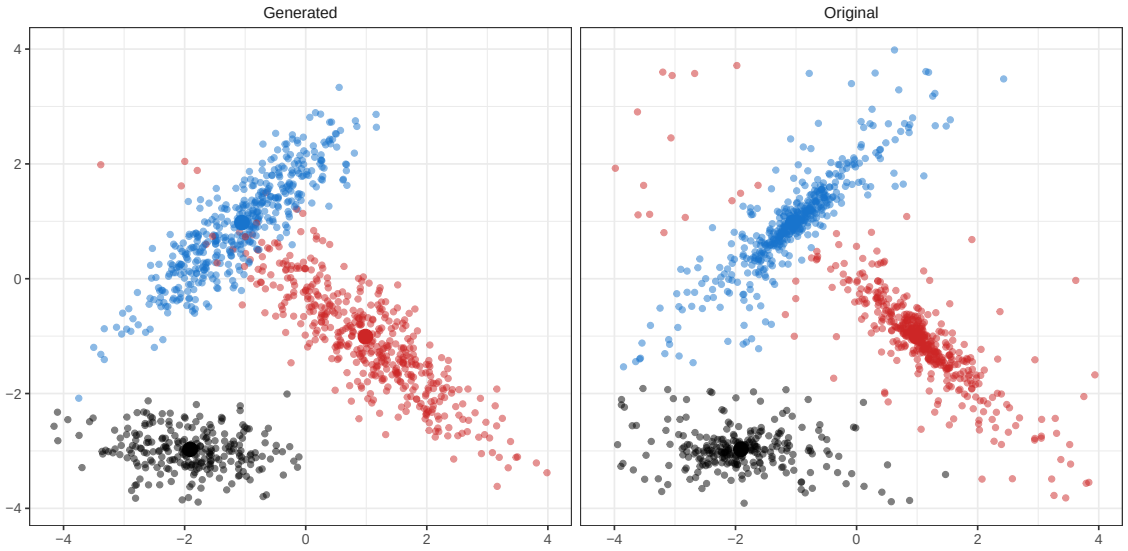

Figure 5: K-means++ with data *challenge-3.csv*. The model sufficiently describes the data, though is still influenced by outliers.