Clustering Methods for Incomplete Datasets

Author: Connor J. Charlton
Supervisor: Dr Tahani Coolen-Maturi

Durham University

Cluster Analysis

Cluster analysis is the study of methods for segmenting data into a specified number of clusters or for matching data drawn from a mixture distribution to the appropriate component distributions. [1]

Principal Component Analysis (PCA)

Principal component analysis is a field of unsupervised learning aimed at reducing the dimension (say from dimension p to dimension q) of a dataset while losing as little information as possible from said dataset. That corresponds to constructing a set of q ordered linear combinations of the dataset's p variables, such that variables with higher variances are given heavier (greater in absolute value) weightings and variables which are highly correlated with other variables, that already have a heavy weighting assigned to them, are given a lighter weighting. This set of linear combinations can be represented by a matrix $V_{p\times q}$ (under the constraint - to ensure uniqueness of the matrix - that each column be a unit vector), the weightings of the columns for each data point reconstruction can be represented by the q-entry vector $\boldsymbol{\lambda}$ and the mean of all the data points can be represented by the p-entry vector $\boldsymbol{\mu}$. The problem of calculating the principal components, for an $n \times p$ dataset, is then reduced to solving the optimisation problem, [2]

$$\underset{\boldsymbol{\mu},V_{p\times q},\boldsymbol{\lambda}}{\arg\min} \sum_{i=1}^{n} \|\mathbf{x_i} - \boldsymbol{\mu} - V_{p\times q}\boldsymbol{\lambda}\|^2 \tag{1}$$

We can partially optimise $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ analytically:

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}} \tag{2}$$

$$\hat{\boldsymbol{\lambda}} = V_{p\times q}\,(\mathbf{x_i} - \overline{\mathbf{x}}) \tag{3}$$

However, $V_{p\times q}$ must be optimised numerically.
Principal component analysis is particularly useful for clustering, because it can often enable visual identification of clusters in high-dimensional datasets.

k-Means Clustering

However, for some datasets, drawn from mixture distributions, known to have some number k component distributions, the k clusters are not separately greatly enough to be visually identified even under PCA and thus clustering algorithms are required. k-means clustering is an approach that involves solving the optimisation problem, [2]

$$C^* = \underset{C,\{\mathbf{m_\kappa}\}_1^k}{\arg\min} \sum_{\kappa=1}^{k} N_\kappa \sum_{C(i)=\kappa} \|\mathbf{x_i} - \mathbf{m_\kappa}\|^2 \tag{4}$$

To do so, we first note that, for any set of observations $S = \{\mathbf{x_1}, ..., \mathbf{x_n}\}$,

$$\bar{\mathbf{x}} = \underset{\mathbf{m}}{\arg\min} \sum_{i=1}^{n} \|\mathbf{x_i} - \mathbf{m}\|^2 \tag{5}$$

And thus deduce that,

$$C^* = \underset{C}{\arg\min} \sum_{\kappa=1}^{k} N_\kappa \sum_{C(i)=\kappa} \|\mathbf{x_i} - \bar{\mathbf{x}}_{\boldsymbol{\kappa}}\|^2 \tag{6}$$

where $\bar{\mathbf{x}}_{\boldsymbol{\kappa}}$ denotes the mean of the κ^{th} cluster.
Alternating implementation of formulae (5) and (6) thus enables us to converge towards a clustering assignment that increasingly minimises (4). This is the logical basis of the k-means clustering algorithm.

k-Means Algorithm

1. A random set $\{\mathbf{m_1}, ..., \mathbf{m_k}\}$ of cluster means are selected from the space $\mathcal{X}$ of possible $\mathbf{x}$ values.
2. Given the current set $\{\mathbf{m_1}, ..., \mathbf{m_k}\}$ of cluster means, (4) is minimised by assigning each observation to the closest current cluster mean. That is, $C(i) = \arg\min_{\kappa\in\{1,...,k\}} \|\mathbf{x_i} - \mathbf{m_\kappa}\|^2$.
3. Given the current cluster assignment $C : \{1, ..., n\} \to \{1, ..., k\}$, (4) is minimised with respect to the cluster means $\{\mathbf{m_1}, ..., \mathbf{m_k}\}$ by taking $\{\mathbf{m_1}, ..., \mathbf{m_k}\} = \{\bar{\mathbf{x}}_\mathbf{1}, ..., \bar{\mathbf{x}}_\mathbf{k}\}$.
4. Steps 2. and 3. are iterated until the cluster assignment does not change.
5. The final cluster assignment is returned.

Note that the initial selection of cluster means could potentially effect the returned cluster assignment because the above algorithm only minimises (4) locally not necessarily globally; thus, in practice, the above algorithm is usually run several times with different initialisations and if multiple distinct assignments are returned, then whichever assignment results in the least total cluster variance (i.e. minimises (4)) is estimated to be the global minimum.

k-Means Clustering on Simulated Dataset

Here we illustrate k-means clustering on a simulated dataset, containing instances known to be drawn from a mixture distribution with three distinct component distributions. Using a simulated dataset is preferable because it enables us to calculate what proportion of instances are assigned to their true clusters.

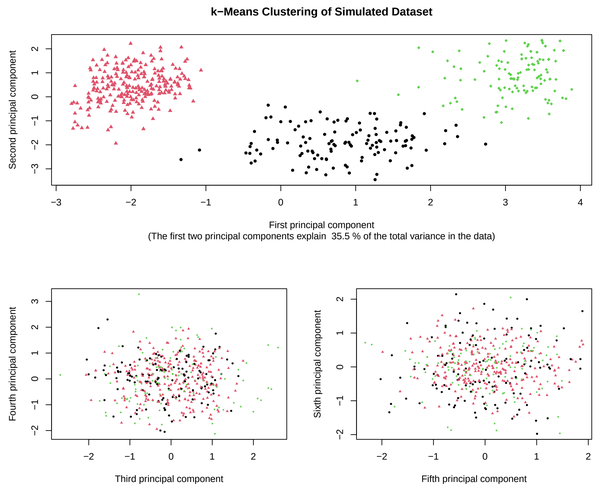

Fig. 1: Graphs illustrating k-means clustering on the simulated dataset

We have graphed this 10-dimensional dataset in three 2-dimensional planes, showing the first six consecutive principal components. The first two principal components collectively explain more than a third of the variance in the data and the graph of these components enables us to visually identify the three clusters. The next four principal components explain far less (per component) of the variance in comparison and, as can be seen, do not enable visual identification of the three clusters and so may as well be omitted in future clustering illustrations. The colours of the instances indicate the clusters to which the k-means algorithm has assigned them.

Multiple Imputation by Chained Equations (MICE)

Many real world datasets are incomplete and thus we now consider a method (MICE) for imputing the missing values and assess its effectiveness as a prerequisite for clustering of incomplete datasets. Consider a dataset with n observations of p variables, with missing values for the q variables $\{p_{\sigma_1}, ..., p_{\sigma_q}\}$. The missing values of these variables can be imputed via the MICE method as follows: [3]

1. Impute all missing values using a column-wise single imputation method.
2. Reset the missing values for the $\sigma_1{}^{\text{th}}$ variable to NA (or missing).
3. Build a regression/classification model for the $\sigma_1{}^{\text{th}}$ variable, using the other $p-1$ variables as predictor variables, to predict and impute the missing values for the $\sigma_1{}^{\text{th}}$ variable.
4. Repeat steps 2. and 3. for the $\sigma_2{}^{\text{th}}$, $\sigma_3{}^{\text{th}}$, ..., $\sigma_{q-1}{}^{\text{th}}$ and $\sigma_q{}^{\text{th}}$ variables.
5. Continue to cycle through this omission-regression cycle until the values assigned in place of the initially missing values stabilise.

k-Means Clustering of the Simulated Dataset after MICE

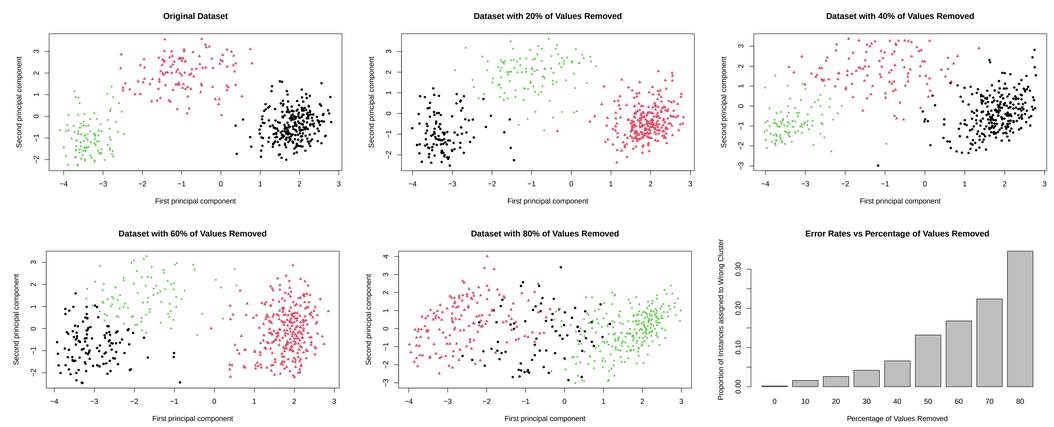

Fig. 2: Graphs illustrating the effectiveness of implementing MICE on randomly sparsified versions of the simulated dataset prior to k-means clustering of the dataset

The above figure shows the results of k-means clustering on versions of the simulated dataset with different proportions of data removed. In each case, the missing data has been imputed by MICE. Note that the colours of each cluster vary from graph to graph, because the colour coding is purely for the purpose of distinguishing clusters. Likewise, principal components' signs are not unique and thus vertical or horizontal reflections of the above graphs does not change their meaning. From these graphs we can see that when two fifths of the data values are removed, MICE still imputes those values accurately enough that graphing the first two principal components against each other still enables visual identification of three distinct clusters. The bottom right graph shows the error rates for each proportion of data values removed and imputed and as can be seen, even with four fifths of the data removed, MICE still imputes the missing values accurately enough for the k-means clustering algorithm to correctly assign to their clusters almost two thirds of the instances - double what would occur by chance.

References

[1] B. Everitt, S. Landau, M. Leese, and D. Stahl, "Cluster Analysis," tech. rep., 2011.
[2] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning Data Mining, Inference, and Prediction," tech. rep., 2017.
[3] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?," *International Journal of Methods in Psychiatric Research*, vol. 20, pp. 40–49, 3 2011.