Contents

1 Working with Unlabeled Data – Clustering Analysis **1**
1.1 Grouping objects by similarity using k-means 1
1.1.1 k-means clustering using scikit-learn 1

Chapter 1

Working with Unlabeled Data – Clustering Analysis

1.1 Grouping objects by similarity using k-means

1.1.1 k-means clustering using scikit-learn

Algorithm 1: The k-means algorithm

begin
  Randomly pick k centroids from the examples as initial cluster centers;
  repeat
    Assign each example to the nearest centroid, $\mu^{(i)}, j \in \{1, \ldots, k\}$;
    Move the centroids to the center of the examples that were assigned to it;
  until *the cluster assignments do not change or a user-defined tolerance or maximum number of iterations is reached*;
end

A problem with k-means is that one or more clusters can be empty.

Feature scaling

When we are applying k-means to real-world data using a Euclidean distance metric, we want to make sure that the features are measured on the same scale and apply z-score standardization or min-max scaling if necessary.