Line H. Clemmensen & Sneha Das
DTU Compute

Line H. Clemmensen & Sneha Das, DTU Compute

Exercises 02582
Module 9
Spring 2025

April 2, 2025

Topics: K-means, hierarchical cluster analysis, gap statistic

Exercises (Coding hints at the end of this document):

1 Run K-means clustering using `kmeans_demo` (`Kmeans_demo.m`, `kmeans_demo.R`, or `kmean_demo.py`). The demo fits K-means clustering to tree classes each sampled from a 2D mixture of Gaussian distributions, you can change the values of K in the code.

- Try different numbers of clusters. Which value(s) of K seem reasonable? Why?

2 Perform hierarchical clustering on the zip data (example 3 in ESL). The data consist of 400 samples of handwritten digits 0-9 in 16×16 grayscale images (= 256 features).

- Try different dissimilarity measures. Which work best?
- Where would you cut the dendrogram (ie how many clusters should we choose)?
- You may use `hierarchicalEx.m`, `HierarchicalEx.py`, or `hierarchicalEx.R`.

3 Use the gap-statistic to select K for K-means clustering on the zip data.

- You will need to write the calculations of the within-class dissimilarities and the gap statistic yourself.
- Use `KmeansEx.m`, `KmeansEx.R`, or `kMeans_ex.py`
- Try first `kmeans` and then `kmedoids` (Optional).

4 We have data with four different measures from flowers of three different species (`Fisheriris.csv`). There are 50 observations of each species. See if you can identify

three clusters in data using gaussian mixture modelling. (Two of the species are very similar)

- Plot data using a scatterplot matrix
- Loop over different numbers of clusters
- Plot BIC/AIC for different model orders
- Notice the different extra parameters in the provided Gaussian Mixture function - they might be necessary.

Line H. Clemmensen & Sneha Das
DTU Compute

Line H. Clemmensen & Sneha Das, DTU Compute

Exercises 02582
Module 9
Spring 2025

April 2, 2025

Topics: K-means, hierarchical cluster analysis, gap statistic

Resources for this exercise:

Listing 1: Resources in Matlab

```
Kmeans_demo.m % exercise 1
HierarchicalEx.m % exercise 2
KmeansEx.m % exercise 3
Fisheriris.csv % Fisher's Iris data
zipdata.csv % zip data
ziplabel.csv % labels for the digits in zip data
plotmatrix % plots scatter plots in a matrix
gmdistribution.fit
```

Listing 2: Resources in R

```
kmeans_demo.R # exercise 1
hierarchicalEx.R # exercise 2
KmeansEx.R # exercise 3
Fisheriris.csv # Fisher's Iris data
zipdata.csv # zip data
ziplabel.csv # labels for the digits in zip data
require(MESS)# contain panel.hist for scatter plot matrix
require("mclust")#mixture clustering package
Mclust(X, ...) # gaussian mixture model
```

Listing 3: Resources in Python

```
kmean_demo.py # exercise 1
HierarchicalEx.py # exercise 2
kMeans_ex.py # exercise 3
Fisheriris.csv # Fisher's Iris data
zipdata.csv # zip data
ziplabel.csv # labels for the digits in zip data
import numpy as np
import pandas as pd
from pandas.tools.plotting import scatter_matrix
from sklearn import preprocessing
from sklearn.mixture import GaussianMixture
from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib.pyplot as plt
```

End of exercise