Лабораторная работа 1. Кластеризация данных в R

Цель занятия — научиться выполнять k-means и иерархическую кластеризацию данных в R, анализировать и представлять полученные результаты в виде графиков.

Вспомогательный материал: Лекции 6–8.

Общие указания:

1. В качестве отчета о выполнении практического занятия студент представляет преподавателю программный код (***lr1.R***).
2. Для импорта, экспорта и обработки данных нельзя использовать возможности интерфейса RStudio — можно только писать программный код.
3. Задания должны быть выполнены наиболее оптимальным образом (автоматизация, минимизация строк кода, универсальность и др.).
4. Программный код должен быть оформлен в соответствии с Google's R Style Guide.
5. Смысловые блоки программного кода необходимо сопровождать комментариями.

Задание 1

—> *wine.RData*

<— *kmeans.txt*

Загрузите data set **wine.RData**.

В data set собраны данные по различным видам вин. Дан набор наблюдений (x_1, x_2, ..., x_n). Цель k-means кластеризации — разделить n наблюдений на k≤n, т.о. чтобы минимизировать внутрикластерную сумму квадратов. Кластерный анализ можно выполнять с помощью функции *kmeans* в R.

Далее необходимо выполнять представленный код, изучая используемые функции и поучаемые результаты.

```
wine.stand <- scale(wine[-1])  # To standardize the variables

# K-Means
k.means.fit <- kmeans(wine.stand, 3) # k = 3
```

Сохраните k.means.fit в файл **kmeans.txt**.

В k.means.fit содержатся все атрибуты результата кластеризации:

```
attributes(k.means.fit)

# Centroids:
k.means.fit$centers
```

```
# Clusters:
k.means.fit$cluster

# Cluster size:
k.means.fit$size
```

Фундаментальный вопрос заключается в том, чтобы определить значение параметра k. Если смотреть на процент дисперсии как зависимость от количества кластеров: необходимо выбрать количество кластеров таким образом, что добавление еще одного кластера не будет давать намного лучшее моделирование данных.

Если построить график зависимости внутрикластерной суммы квадратов от количества кластеров, первые кластеры будут добавлять много информации (объяснять большую долю дисперсии). Но в некоторой точке предельная выгода (усиление модели) начнет снижаться, что отразится в появлении точки перегиба на графике, — так называемый “elbow criterion”. Число кластеров выбирается в этой точке.

```
wssplot <- function(data, nc = 15, seed = 1234) {
  wss <- (nrow(data) - 1) * sum(apply(data, 2, var))
  for (i in 2:nc) {
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)
    }
  plot(1:nc, wss, type = "b", xlab = "Number of Clusters",
       ylab = "Within groups sum of squares")
  }

wssplot(wine.stand, nc = 6)
```

Какое число кластеров будет оптимальным в этом случае?

Задание 2

<— clusplot.png

Библиотека **clusters** позволяет визуализировать (с помощью PCA) результат кластеризации в двух измерениях:

```
library(cluster)
clusplot(wine.stand, k.means.fit$cluster, main = '2D representation
of the Cluster solution',
         color = TRUE, shade = TRUE,
         labels = 2, lines = 0)
```

Сохраните полученный график под именем **clusplot.png**.

Для оценки эффективности кластеризации строим матрицу несоответствий (confusion matrix):

```
table(wine[, 1], k.means.fit$cluster)
```

Задание 3

<— dendrogram.png

Методы иерархической кластеризации используют матрицу расстояний в качестве входных данных для алгоритма кластеризации. Выбор подходящей метрики будет влиять на форму кластеров, так как некоторые элементы могут быть близки друг к другу, в соответствии с одним расстоянием и дальше в соответствии с другим.

```
d <- dist(wine.stand, method = "euclidean") # Euclidean distance
matrix.


H.fit <- hclust(d, method = "ward.D2")
```

Результат кластеризации может быть представлен в виде дендрограммы.

```
plot(H.fit) # displays dendrogram

groups <- cutree(H.fit, k = 3) # cut tree into 3 clusters

# Draws dendogram with red borders around the 3 clusters
rect.hclust(H.fit, k = 3, border = "red")
```

Сохраните полученный график под именем **dendrogram.png**.

Для оценки эффективности кластеризации строим матрицу несоответствий:

```
table(wine[, 1], groups)
```

Дополнительные задания

—> Cluster Analysis with R.pdf

—> protein.csv

—> clustering-vanilla.xls

—> snsdata.csv

Выполните Study cases I, II, III из **Cluster Analysis with R.pdf**.

Литература:
Martos, G. Cluster Analysis with R [Electronic resource]. – Mode of access: https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html. – Date of access: 20.02.2017.