Домашнее задание по теме “Кластерный анализ”

Алла Тамбовцева

Дедлайн: 10 мая 2018 г, 23:59

1. Загрузите базу данных, которая содержится в файле cpds.csv и посмотрите на нее:

```
dat <- read.csv(file.choose(), dec = ",")
View(dat)
```

   Не забудьте опцию dec (как в коде выше). Она нужна для того, чтобы R понимал, что в качестве разделителя в дробных числах используется запятая, а не точка. Иначе все столбцы с дробными числами считаются как текст, как факторные переменные!

   В базе сохранены различные данные по 35 странам за 2015 год, файл cpds.csv – сильно сокращенная версия базы данных проекта COMPARATIVE POLITICAL DATA SET (http://www.cpds-data.org/index.php/data).

   Ссылка на codebook по переменным в базе (в файле для работы названия переменных сохранены): http://www.cpds-data.org/images/Update2017/Codebook-CPDS-1960-2015-Update-2017.pdf

2. При необходимости, добавьте в базу новые переменные. Например, дамми-переменные.

   Пример кода для добавленния дамми “посткоммунистическое государство или нет”:

```
dat$postcom <- ifelse(dat$poco == "Post-communist", 1, 0)
View(dat)
```

   Добавляем столбец postcom в базу dat, если переменная poco из базы df принимает значение “Post-communist”, то в столбце ставится 1, иначе – ставится 0.

3. Выберите из базы dat переменные интереса – переменные, по которым вы будете кластеризовать страны в базе и сохраните в новую базу to_clust. Требование: должно быть выбрано не менее 4 переменных. Для удобства используйте функцию select() из библиотеки dplyr. В скобках в select достаточно просто через запятую перечислить названия нужных столбцов.

   Назовите строки в базе to_clust разумным образом: по названию или коду стран.

4. Реализуйте иерархический кластерный анализ на основе базы to_clust. Обоснуйте выбор используемого расстояния и метода агрегирования. Постройте дендрограмму. Если подписи на графике слишком большие, приведите их в порядок, отрегулировав шрифт.

5. Выберите число кластеров на основе полученной дендрограммы. Обоснуйте свой выбор, исходя из содержательных соображений. Наложите на дендрограмму границы получившихся кластеров (rect.hclust). Сохраните полученные метки кластеров в исходную базу df.

6. Проведите проверку качества кластеризации.

   - Выберите строки в базе df, соответствующие каждому полученному кластеру, и прокомментируйте, какие страны входят в каждый кластер. Есть ли какие-то особенности у каждого кластера?
   - Визуализируте распределения выбранных показателей по кластерам любым разумным способом. Прокомментируйте, заметны ли отличия в распределении разных показателей по группам.
   - Выведите какие-нибудь описательные статистики кластеров. Проинтерпретируйте.
   - Используйте подходящий статистический критерий для того, чтобы проверить, отличаются ли средние значения/распределения показателей по кластерам. Проинтерпретируйте полученный результат.

- Реализуйте иерархический кластерный анализ с использованием другого расстояния/метода агрегирования. Сравните результаты.

7. Проверьте, используя “метод согнутого локтя” и “силуэтный метод”, какое число кластеров нужно выбрать, исходя из статистических соображений. Соответствует ли это число выбранному вами числу кластеров? Прокомментируйте.

8. Реализуйте кластерный анализ методом k-средних с выбранным вами окончательным числом кластеров. Сохраните метки кластеров, полученные в результате процедуры k-means, в исходную базу df. Выберите строки из базы df, соответствующие каждому кластеру и предложите окончательную содержательную интерпретацию каждому кластеру.

   (Пример интерпретации: в первом кластере находятся посткоммунистические страны с высокой долей левых партий в правительстве, во втором . . .).