Домашнее задание по курсу
«Введение в многомерный статистический анализ»
по теме «Кластерный анализ»

Формат сдачи

- Выполненную работу необходимо отправить на почту **allatambov@gmail.com** с темой *Кластерный анализ: домашнее задание* до **1 февраля 23:59**.
- Выполнение теоретической части не предполагает использования R, его разрешается использовать только для самопроверки. Решения теоретической части должны быть сданы в виде файла PDF. Выполнение работы в LaTeX и других текстовых редакторах приветствуется, но не является обязательным. Можно аккуратно написать решения от руки, отсканировать/сфотографировать, а затем все картинки связать в один файл PDF. Также теоретическую часть можно сдать в бумажном виде на семинаре не позднее дедлайна.
- Решения практической части предпочтительнее сдавать в виде файла PDF/Word, полученного из Rmd-файла в RStudio или Sweave. Если вы не работаете с Rmd-файлами или Sweave, можно просто вставить код R и графики в обычный текстовый файл.

Теоретическая часть: 10 баллов

Дан небольшой двумерный массив с данными по пяти наблюдениям:

id	X	Y
1	0	8
2	5	1
3	8	6
4	2	1
5	1	7

a. (1 балл) Сколько различных ненулевых расстояний необходимо посчитать для построения матрицы расстояний между наблюдениями? Приведите ответ и объясните его.

b. (3 балла) Запишите матрицу расстояний[1] для предложенного массива, используя в качестве метрики расстояние Чебышёва. Приведите необходимые вычисления.

c. (4 балла) Используя полученную на предыдущем шаге матрицу расстояний и метод дальнего соседа, реализуйте иерархический кластерный анализ и постройте дендрограмму. Приведите все необходимые вычисления.

d. (1 балл) Сколько кластеров нужно выделить, если в каждом кластере должно быть не менее двух наблюдений?

e. (1 балл) Сколько кластеров получится, если для деления на группы мы будем «разрезать» дендрограмму по расстоянию 5.5?

[1] Для упрощения вычислений считайте, что показатели X и Y измерены в одной шкале, шкалировать данные не нужно.

Практическая часть часть: 40 баллов

Поставновка задачи

Необходимо провести мини-исследование и выяснить:

- какие группы районов города Балтимор можно выявить на основе данных о числе преступлений разной степени серьёзности;
- можно ли считать, что деление районов на более опасные и менее опасные связано с их географическим положением (например, что самые опасные преступления совершаются на окраинах) или с другими социально-экономическими характеристиками районов (например, промышленные районы, районы с высокой долей ночных клубов).

Описание данных и подготовка данных

Для работы над этим заданием вам потребуются библиотеки `tidyverse`, `ggplot2` и `geojsonio`, а также некоторые библиотеки для проверки качества кластеризации.

В файле `Neighborhoods.geojson` хранятся географические характеристики районов города Балтимор, необходимые для отрисовки карты города. Файл доступен по ссылке.

Чтобы построить в R карту города с границами районов, выполните следующие действия:

```
library(ggplot2)
library(geojsonio)

# вместо link вставьте ссылку на geojson-файл в кавычках

spdf <- geojson_read(link,  what = "sp")
fortified <- fortify(spdf, region = "name")

ggplot() + geom_polygon(data = fortified, aes(x = long, y = lat, group = group),
               fill = "white", color = "grey") + theme_void() + coord_map()
```

Пояснения к коду:

- функция `geojson_read()` из `geojsonio` считывает файл в формате `.geojson`, который хранится на Github;
- функция `fortify` из `ggplot2` превращает объект типа `SpatialPolygonsDataFrame` в обычный датафрейм, с которым умеет работать `ggplot()`;
- функция `geom_polygon()` отрисовывает карту, задействуя значения широты и долготы по осям `x` и `y` и отмечая границы районов `group`.

В файле `BaltimoreCrimesAgg.csv` хранятся данные по числу преступлений разного вида, совершённых в каждом районе в 2011-2016 годах:

- `ASSAULT`: число нападений;
- `BURGLARY`: число краж со взломом;
- `HOMICIDE`: число убийств;
- `LARCENY`: число хищений разного вида;
- `RAPE`: число изнасилований;
- `ROBBERY`: число ограблений;

Загрузите в R данные из этого файла и сохраните их в датафрейм `crime`.

Задание

Задача 1. Отбор переменных (1 балл)

Выберите из полученного датафрейма все числовые переменные и сохраните их в новый датафрейм `to_clust`. Для удобства используйте функцию `select()` из библиотеки `tidyverse`. В скобках достаточно через запятую перечислить названия нужных столбцов. Например, так будет выглядеть код для выбора столбцов A, B, C:

```
library(tidyverse)
to_clust <- crime %>% select(A, B, C)
```

Так как названия районов не нужны для реализации кластеризации (текстовый столбец всё испортит, мы не посчитаем расстояния из-за него), но нужны на этапе интерпретации, имеет смысл столбец с названиями районов не выбирать, но назвать строки датафрейма по названиям районов:

```
rownames(to_clust) <- crime$Neighborhood
```

Задача 2. Иерархический кластерный анализ (3 балла)

Реализуйте иерархический кластерный анализ на основе датафрейма `to_clust`. Обоснуйте выбор используемого расстояния и метода агрегирования. Постройте дендрограмму. Если подписи на графике слишком большие, приведите их в порядок, отрегулировав шрифт (аргумент `cex`).

Выберите число кластеров на основе полученной дендрограммы (не менее 4). Обоснуйте свой выбор, исходя из содержательных соображений. Сохраните полученные метки кластеров в столбец `cluster` датафрейма `to_clust`.

Задача 3. Оценка качества кластеризации (16 баллов)

Проведите проверку качества кластеризации.

- Выберите строки в `to_clust`, соответствующие каждому полученному кластеру, и прокомментируйте, какие районы входят в каждый кластер. Есть ли какие-то особенности у каждого кластера?

  Пример кода для выбора кластера с меткой 1 на основе столбца `cluster`:

  ```
  cluster01 <- to_clust %>% filter(cluster == 1)
  View(cluster01)
  ```

- Выведите описательные статистики по каждому кластеру. Проинтерпретируйте.

- Визуализируте распределения наболее интересных показателей по кластерам любым разумным способом. Прокомментируйте, заметны ли отличия в распределении разных показателей по группам.

- Используйте подходящий статистический критерий для того, чтобы проверить, отличаются ли средние значения/распределения показателей по кластерам. Проинтерпретируйте полученный результат.

Задача 4. Уточнение числа кластеров (5 баллов)

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 и выберите итоговое количество кластеров.

Задача 5. Кластеры и география (5 баллов)

Проверьте, можно ли объяснить деление на выбранное вами число кластеров географическим расположением районов. Для этого выполните раскраску карты города таким образом, чтобы районы, относящиеся к одному кластеру, были одного цвета.

Объедините датафрейм с географической информацией и датафрейм с данными по преступности (в первом названия районов сохранеы в столбце `id`, во втором – в `Neighborhood`):

```
full <- fortified %>% left_join(. , crime, by=c("id"="Neighborhood"))
```

Сделайте заливку элементов (многоугольников, отвечающих за районы) зависящей от метки кластера в `cluster`:

```
ggplot() + geom_polygon(data = full,
           aes(fill = cluster, x = long, y = lat, group = group)) +
           theme_void() + coord_map()
```

Проинтерпретируйте полученные результаты.

Задача 6. K-means (10 баллов)

Реализуйте кластерный анализ методом k-средних с выбранным вами окончательным числом кластеров. Сохраните метки кластеров, полученные в результате процедуры k-means, в датафрейм `to_clust`. Выберите строки `to_clust`, соответствующие каждому кластеру и предложите окончательную содержательную интерпретацию каждому кластеру.

Пример интерпретации. В первом кластере находятся районы на окраине города, в которых совершаются самые опасные преступления, такие как убийства и...