Київський національний університет імені Тараса Шевченка
Кафедра теорії ймовірностей, статистики та актуарної математики

Р. Майборода

Самостійна робота по курсу

"Статистичний аналіз багатовимірних даних"

Для студентів магістратури за напрямом "статистика"

Індивідуальні завдання
та рекомендації по виконанню
Документ в процесі розробки. Версія від 07.10.2018

Київ — 2018

Вступ

Для виконання завдань потрібно встановити R та RStudio на своєму комп'ютері.

Щоб встановити R для Windows зайдіть на сторінку

`http://cran.r-project.org/bin/windows/base/`

і виберіть `Download R 3.4.1 for Windows` (номер версії, скоріше за все, буде вже іншим). Після цього запустіть програму, яка буде завантажена на ваш комп'ютер і відповідайте на її запити.

Якщо вам потрібна версія R для іншої операційної системи, зайдіть на сторінку

`http://www.r-project.org/`

і виберіть там варіант, який вас влаштовує.

Для того, щоб встановити RStudio, зайдіть на сторінку

`www.rstudio.com`

і виберіть там варіант для завантаження. Встановлювати RStudio треба після того, як буде встановлено R.

Книжку [3], присвячену статистичному аналізу даних за допомогою R, можна отримати за адресою:

`http://probability.univ.kiev.ua/userfiles/mre/compsta.pdf`

Завдання 1.
Кластеризація з відомою кількістю кластерів.
Частина 1.

1. У архіві

`http://probability.univ.kiev.ua/userfiles/mre/mult6task.rar`

знайдіть файл з іменем `mult<N>.txt`, де `<N>` — номер Вашого варіанту. Запишіть його на Ваш комп'ютер у зручному для читання місці.

2. У файлі міститься таблиця модельованих даних. У першому її рядочку знаходяться назви змінних, а кожен наступний рядочок відповідає одному спостереженню. Кожен стовпчик таблиці відповідає одній змінній.

Прочитайте цю таблицю у `R` за допомогою функції `read.table()` і проведіть кластерний аналіз цих даних використоуючи методи центроїдів та медоїдів. У методі медоїдів використайте звичайну евклідову відстань між спостереженнями.

Спробуйте різні кількості кластерів, почианючи від 2-х до 20-ти.

3. Виберіть оптимальну кількість кластерів, використовуючи внутрішньокластерну суму квадратів для методу центроїдів та графік середніх силуетів як у методі центроїдів так і у методі медоїдів.

Відповідні графіки наведіть у звіті.

4. Для двох-трьох варіантів кластеризації, які Ви визнаєте найкращими, відобразіть результати на діаграмах розсіювання, використовуючи метод головних компонент та метод канонічних компонент.

Діаграми наведіть у звіті.

5. Порівняйте обрані Вами варіанти кластеризації з п. 4 використовуючи індекс Ренда та поліпшений індекс Ренда.

За потреби, наведіть таблицю спряженості для розглянутих кластеризацій.

Виберіть остаточний варіант кластеризації, який Ви вважаєте оптимальним. Зробіть висновки (який з використаних методів виявився вдалим, який — ні, які методи дали однакові результати і т.п.).

Рекомендації по виконанню.

Частина перша. Прочитаємо дані, використовуючи фунцію verb|read.table()|:

```
>   samp<-read.table("c:\\rem\\mult6\\mult0.txt")
```

Тепер дані знаходяться у фреймі даних `samp`.

Провести кластеризацію центроїдним методом можна, використовуючи функцію `kmeans()`, їй потрібно передати дані і вказати кількість кластерів. Крім того, можна задати опцію `nstart`, що вказує, скільки варіантів початкових наборів центрів буде випробуватись (ці набори вибираються випадково, якщо не задати їх явно у опції `centers`).

Значення номерів кластерів до яких відносяться об'єкти, функція вміщує у атрибуті `$cluster` свого результату.

Можна вивести діаграму розсіювання, у якій точки з різних кластерів пофарбовані у різні кольори. У наступному прикладі це зроблено для випадку чотирьох кластерів.

```
>  km.res <- kmeans(samp, 4, nstart = 25) # кластеризація
> km.res$tot.withinss # внутрішньокластерна сума квадратів

[1] 34148.6

> km.res$betweenss    # міжкластерна сума квадратів

[1] 266704.5

> km.res$betweenss/km.res$tot.withinss

[1] 7.810114

> pal<-c("black","red","green","blue")    # палітра кольорів
> plot(samp[,1],samp[,2],cex=0.2,col=pal[km.res$cluster])
```

Внутрішньокластерна сума квадратів у 7.8 разів менша ніж міжкластерна, отже можна сподіватись, що ця кластеризація виявляє певну структуру даних.

На діаграмі (рис.1) візуально виділяються приблизно 8 кластерів, але результати кластеризації їм явно не відповідають. (На діаграмах розсіювання інших змінних може бути видно щось зовсім інше, спробуйте.) Варто продовжити спроби.

Побудуємо діаграму внутрішньокластерних сум квадратів в залежності від кількості кластерів — рис. 2:

```
>   library(factoextra)
> fviz_nbclust(samp, kmeans, method = "wss",k.max = 20)
```

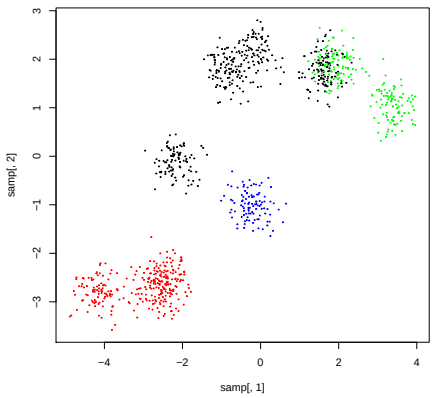

Рис. 1: Діаграма розсіювання перших двох змінних. Кластеризація методом центроїдів.

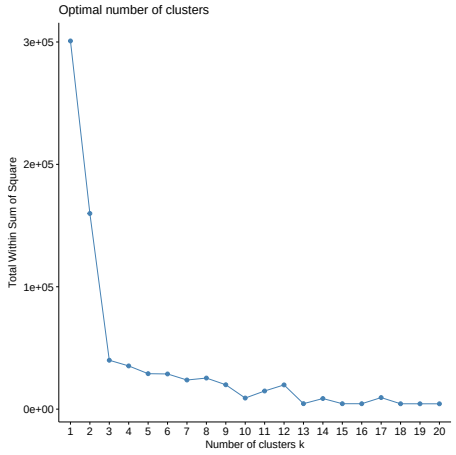

Рис. 2: Внутрішньогрупові суми квадратів. Кластеризація методом центроїдів.

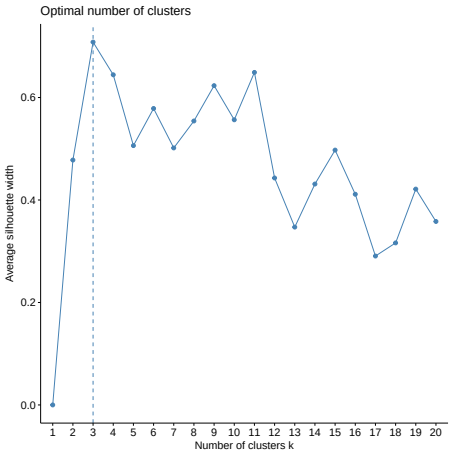

Рис. 3: Середні силуети. Кластеризація методом центроїдів.

На цій діаграмі видно "злам" при $k = 3$ і невеликий "провал" при $k = 10$. Ці числа є претендентами на оптимальну кількість кластерів.

Побудуємо діаграму середніх силуетів — рис. 3

```
>  fviz_nbclust(samp, kmeans, method = "silhouette",k.max = 20)
```

На цій діаграмі помітні максимуми при $k = 3$, $k = 11$, $k = 15$.

Розглянемо випадок $k = 10$. (У роботі Ви перевірите всіх претендентів)

Проведемо кластеризацію і подивимось на діаграму розсіювання даних у просторі перших двох головних компонент — рис. 4:

```
> km.res <- kmeans(samp, 10, nstart = 25)
> pal<-c("black","red","blue","green","magenta","chocolate",
+        "darkblue","darkred","aquamarine","grey")
> plot(princomp(samp)$scores[,1:2],col=pal[km.res$cluster],cex=0.2)
```

На рисунку помітні дев'ять кластерів з десяти (один виявився "прихованим" за іншими), вони розбиваються на три групи. Не зовсім очевидно, наскільки запропоноване розбиття дійсно відображає структуру даних, наприклад, у верхній групі — тут кластери не виглядають відділеними один від одного.

(Тут можна іще подивитись інші пари головних компонент, наприклад, другу і третю, або покрутити дані у тривимірній графіці).

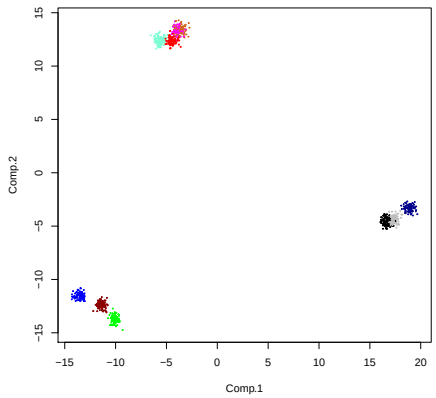

Рис. 4: Діаграма розсіювання перших двох головних компонент, Кластеризація методом центроїдів.

Відобразимо діаграму розсіювання даних у просторі третьої і четвертої канонічних компонент:

```
> require(CCA)
> cl<-km.res$cluster
> k<-length(levels(as.factor(cl)))
> n<-nrow(samp)
> C<-matrix(data=as.numeric(rep(cl,k)==rep(1:k,each=n)),ncol=k,nrow=n)
> cc_res<-rcc(samp,C,0.1,0.1)
> plot(cc_res$scores$xscores[,3:4],col=pal[km.res$cluster],cex=0.2)
```

На цій діаграмі структура “трьох груп” кластерів не помітна, але видно всі 10 кластерів і вони розташовані окремо один від одного. Ті кластери, що розмістились поруч на цій діаграмі, розташовані у різних місцях на діаграмі головних копонент.

Отже можна зробити висновок, що у даних виділяються 10 кластерів, що розбиваються на три групи.

Аналогічне дослідження слід провести, використовуючи метод медоїдів. Для цього можна використати функцію `pam()` з бібліотеки `cluster`:

```
> library(cluster)
> pam.res<-pam(samp,10)
```

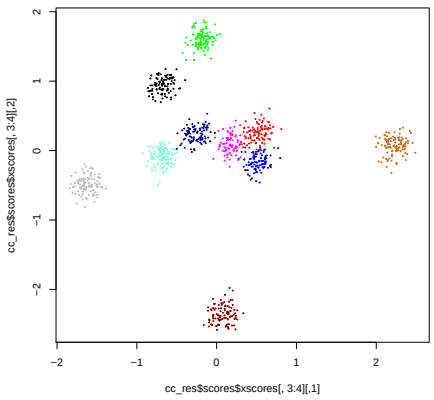

Рис. 5: Діаграма розсіювання перших двох канонічних компонент, Кластеризація методом центроїдів.

Цей метод може дати результати, що відрізняються від результатів методу центроїдів. Щоб порівняти, наскільки різними вийшли кластеризації, можна підрахувати індекс Ренда між ними функцією `rand.index()` з бібліотеки `fossil`:

```
> library(fossil)
> rand.index(pam.res$clustering,km.res$cluster)

[1] 1
```

Індекс виявився рівним 1, тобто ці кластеризації повністю однакові.

Це підтверджує і таблиця спряженості:

```
> library(MASS)
> table(pam.res$clustering,km.res$cluster)
```

	1	2	3	4	5	6	7	8	9	10
1	0	96	0	0	0	0	0	0	0	0
2	0	0	0	0	77	0	0	0	0	0
3	0	0	0	104	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	114	0
5	0	0	98	0	0	0	0	0	0	0

6	99	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	103	0	0
8	0	0	0	0	0	0	0	0	0	107
9	0	0	0	0	0	107	0	0	0	0
10	0	0	0	0	0	0	95	0	0	0

Частина 2.

Підберіть самі для аналізу дані, за якими було б цікаво провести кластеризацію. Застосуйте до них алгоритми, що були використані у частині 1. (Перш ніж виконувати аналіз даних, бажано узгодити їх використання та план роботи з викладачем). Дані потрібно описати у звіті (навести числові значення, якщо їх не дуже багато, або вказати, де їх можна знайти).

Якщо у Вас немає ідеї щодо вибору даних, можна самостійно заповнити рольову гратку і провести її обробку.[1]

Для цього виберіть який-небудь добре знайомий вам твір (книжку, фільм, телесеріал, комп'ютерну гру...) в якому є не менше семи дійових осіб з чітко окресленими особистостями, що відрізняються одна від одної.[2]

Наприклад, візьмемо книгу А. Мілна про Вінні-Пуха та його друзів. З неї можна вибрати Вінні, Крістофера Робіна (КР), Паця, Кролика, Тигру, Сову, Кенгу, Іа.

Запишіть імена обраних дійових осіб у стовпчик. Виберіть випадковим чином трьох з них (можна для цього скористатись генератором випадкових вибірок — функцією `sample()` у `R`, але не обов'язково). Подумайте, яку спільну якість особистості мають два з них, за которою вони відрізняються від третього. Наприклад, взявши Кролика, Тигру і Сову, можна помітити, що Кролик і Сова — розсудливі, а Тигра — спонтанний (вискакучий). Такі якості у методі рольових граток називають *конструктами*. У конструкта є два *полюси*, у нашому прикладі це розумність (позитивний полюс[3]) і спонтанність — негативний.

Спробуйте розмістити всіх обраних вами дійових осіб (будемо називати їх об'єктами нашої гратки) вздовж осі, що з'єднує позитивний та негативний полюси конструкта: той об'єкт, що найкраще відповідає позитивному полюсу, отримує найбільший ранг (КР — 8). Той, що найб-

[1]Про методи психологічного аналізу особистості за допомогою рольових граток (граток Келлі) можна прочитати у [5].

[2]Ви можете попросити заповнити гратку когось із своїх друзів. Тоді твір має бути знайомим для нього/неї, а не обов'язково для вас. **Зауваження.** Недоцільно заповнювати гратку гуртом, це має робити хтось один. Твір, за яким гратка заповнюється має бути реально існуючим, а не, скажімо, намітками сценарію фільму, який ви збираєтесь коли-небудь зняти.

[3]Позитивність і негативність тут не є моральними чи якимось іншими оцінками, а задаються довільно для зручності опису та порівняння з іншими конструктами.

лижчий до негативного полюса — найменший (Тигра — 1).

Якщо обраний вами конструкт неможливо застосувати бо багатьох об'єктів (наприклад, він придатний до опису тільки осіб однієї статі, а у вашій гратці представлені обидві) спробуйте узагальнити його так, щоб він став застосовним до всіх, або хоча б майже всіх. Тим об'єктам, що не отримали рангу за даним конструктом, привласніть значення `NA` (пропущене значення).

Після цього задайте наступний конструкт за тією ж схемою: випадково виберіть набір з трьох елементів і т.д. Намагайтесь обирати нові конструкти так, щоб вони не повторювали вже вибраних. При бажанні можна внести до складу конструктів будь-які риси, що, на вашу думку, потрібні для характеризації об'єктів і були пропущені раніше.

Остаточна кількість виділених конструктів повинна бути не менше 7-ми.

Запишіть отримані вами дані у вигляді таблиці, рядочки яких відповідають об'єктам (дійовим особам), а стовпчики - конструктам. У таблицю занесіть відповідні ранги. Об'єктам та конструктам дайте скорочені імена (не більше 2-х символів). Ці імена будуть використовуватись при візуалізації та при описі результатів кластеризації. У звіті наведіть цю таблицю (гратку), назву твору, за яким складена гратка та пояснення скорочених імен.

Спробуйте обробити цю гратку, використовуючи методи з частини першої. При цьому можна на роль об'єктів використовувати дійових осіб. Тоді конструкти слід трактувати, як змінні, що описують об'єкти. (Кластеризація рядочків гратки).

Можна, навпаки, розглядати конструкти як об'єкти для кластеризації, а значення рейтингів дійових осіб за цими конструктами розглядати, як змінні, що описують конструкти. (Кластеризація стовпчиків).

Проведіть окремо кластеризацію рядочків, окремо - кластеризацію стовпчиків.

Подивіться, чи можна виявити зв'язок між цими кластеризаціями?

Завдання 2.
Класичне багатовимірне шкалування.

Частина 1.

Застосуйте класичне багатовимірне шкалування для візуалізації даних з частин 1 і 2 завдання 1.

При цьому для шкалування спочатку перетворіть індивідуальні дані у таблицю відмінностей, використавши для цього три варіанти відстаней: евклідову, манхаттанську і максимальну. Для кожного варіанту виведіть двовимірну діаграму розсіювання для результатів. Точки на діаграмах розфарбуйте з урахуванням кластеризацій, отриманих у завданні 1.

Порівняйте отримані візуалізації з результатами завдання 1, зробіть висновки.

Рекомендації по виконанню.

Для обчислення відстаней можна скористатись функцією

`dist(x, method)`.

Параметр `x` — фрейм даних (або матриця індивідуальних змінних), за яким розраховуються відстані між об'єктами (рядочками).

`method` — опція, що задає тип відстаней. Може бути `"euclidean"` (за умовчанням), `"maximum"`, `"manhattan"` та ін.

Для багатовимірного шкалування можна використати функцію

`cmdscale(d, k=2, eig=FALSE,...)`

де `d` — матриця відстаней для шкалування,

`k` — вимірність простору, в якому підшукується відповідна конфігурація точок,

`eig` — логічний параметр, що вказує, чи потрібно підраховувати всі власні числа матриці коваріацій для конфігурації у просторі великої вимірності.

Приклад:

```
>         samp<-read.table("c:\\rem\\mult6\\mult0.txt")
> km.res <- kmeans(samp, 10, nstart = 25)
> pal<-rainbow(10)
> d <- dist(samp,method="maximum") # minimum distances between the rows
> fit <- cmdscale(d,eig=TRUE, k=2) # k is the number of dimensions
> x <- fit$points[,1]
> y <- fit$points[,2]
> plot(x, y,col=pal[km.res$cluster],cex=0.2)
```

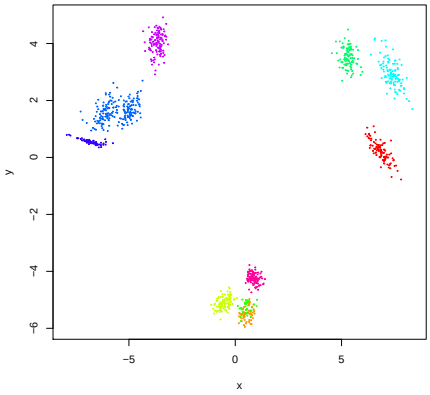

Рис. 6: Результат класичного багатовимірного шкалування.

Результат — на рис 6.

Частина 2.

Розробіть скрипт для застосування техніки проекції на канонічні компоненти до конфігурації точок отриманих методом класичного багатовимірного шкалування у просторі великої вимірності.

Використайте цей скрипт для візуалізації даних із завдання 1 з урахуванням їх кластеризації.

Литература

[1] Карташов М.В. "Імовірність, процеси, статистика". Київ, Видавничо-поліграфічний центр “Київський університет”, 2007, 494 с.

[2] Майборода Р.Є. Регресія: Лінійні моделі.- К. ВПЦ "Київський університет 2007, 296с.

[3] Майборода Р. Комп’ютерна статистика: професійний старт.— 2017

[4] Майборода Р.Є., Сугакова О.В. "Аналіз даних за допомогою пакета R". , 2015. 65 с.

[5] Франселла Ф., Баннистер Д. Новый метод исследования личности. Москва, Прогресс, 1987. 236с.

[6] James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R.— Springer NY 2013.— 440p.

[7] Kassambara Alboukadel Practical Guide to Cluster Analysis in R.— STHDA, 2017.— 187p.