1. *Основні поняття кластеризації та постановка задачі*

Кластеризація, або кластерний аналіз — це статистична процедура, задача якої полягає в розбитті вибірки об'єктів на підмножини, що не перетинаються і називаються кластерами. Кожен кластер має складається зі схожих об'єктів, а об'єкти різних кластерів мають істотно відрізнятися один від одного. Задача кластеризації відноситься до статистичної обробки, а також до широкого класу задач навчання без вчителя. Ще її можна описати через задачу класифікації.

Задача кластеризації — це по факту задача класифікації, бо в обох випадках ми ділимо об'єкти на основі їх подібності між собою, але у випадку кластеризації приналежність навчальних об'єктів будь-яким класам не задається. Така задача — загальна, тому для її розв'язання використовуються різні підходи. Алгоритми побудови кластерів можуть дуже відрізнятися у підходах до того, що відносити в один кластер і як їх взагалі ефективніше шукати. Кластери можна утворювати ґрунтуючись на відстані між ними, на щільності ділянок у просторі даних, інтервалах або на конкретних статистичних розподілах. Це все залежать від конкретного набору даних та мети використання результатів. Кластерний аналіз не є автоматизованим, це скоріше ітераційний процес, тому що часто доводиться змінювати метод опрацювання даних та параметри моделі, поки не буде отримано з результат з заданими властивостями.

Розв'язок неоднозначний, і на це є кілька причин. По-перше, не існує найкращого критерію якості кластеризації. Відомий цілий ряд досить ефективних критеріїв, а також ряд алгоритмів, які не мають чітко вираженого критерію, але все одно здійснюють досить якісну кластеризацію по побудові. Всі вони можуть давати різні результати. По-друге, число кластерів, як правило, не відомо заздалегідь і встановлюється відповідно до деякого суб'єктивного критерія. По-третє,

результат кластеризації істотно залежить від метрики ρ, вибір якої, як правило, також суб'єктивний і визначається спеціалістом.

Щодо формальної постановки задачі:

Маємо вибірку $X^\ell = \{x_1, \ldots, x_\ell\} \subset X$ і функцію відстані між об'єктами $\rho(x, x')$. Треба розбити вибірку на підмножини, які не будуть перетинатися і щоб кожен кластер складався з об'єктів, близьких по метриці ρ, а об'єкти різних кластерів істотно отличались. При цьому кожному об'єкту $x_i \in X^\ell$ приписується мітка (номер) кластера y_i.

Алгоритм кластеризації — це функція $a: X \to Y$, що будь-якому об'єкту $x \in X$ ставить у відповідність мітку кластера $y \in Y$. Множина міток Y в деяких випадках відома заздалегідь, однак частіше завдання полягає в тому, щоб визначити оптимальне число кластерів з точки зору того чи іншого критерію якості кластеризації.

Задача групування набору об'єктів полягає в тому, що об'єкти в одному кластері більш схожі один на одного, ніж об'єкти в інших кластерах. Подібність — це буквально кількість, яка собою відображає міцність взаємозв'язку між двома об'єктами. Кластеризація використовується в основному для видобутка даних, а також в інших областях, таких як машинне навчання, розпізнавання образів, аналіз зображень, пошук інформації, біоінформатика, стиснення даних і комп'ютерна графіка.

Існує два типа кластеризації: жорстка та м'яка. У жорсткій кластеризації кожен об'єкт даних або повністю належить кластеру чи взагалі не належить. В м'якій кластеризації точка чи об'єкт даних може з певною ймовірністю належати більш ніж одному кластеру.

Типи вхідних даних та проблема міри відстані:

- Кожен об'єкт описується набором своїх характеристик, які називаються ознаками. Ознаки можуть бути числовими або нечисловими.

- Матриця відстаней між об'єктами. Кожен об'єкт описується відстанями до всіх інших об'єктів навчальної вибірки.

Дані, що використовуються в кластерному аналізі, можуть мати інтервальний, порядковий або категоріальний тип. Однак наявність суміші різних типів змінної зробить аналіз більш складним. Це пояснюється тим, що в кластерному аналізі необхідно мати певний спосіб вимірювання відстані між спостереженнями, тобто тип використовуваної міри залежить від типу даних.

Для вимірювання відстані для бінарних та категоріальних даних було запропоновано ряд різних варіантів. Наприклад, для інтервальних даних найчастіше використовуваною мірою відстані є евклідова відстань.

Кластеризація — це досить суб'єктивна задача, для розв'язання якої може існувати більше одного правильного алгоритму. Кожен алгоритм слідує своєму набору правил для визначення «подібності» між об'єктами даних. Найбільш відповідний алгоритм кластеризації для конкретної проблеми часто потрібно вибирати експериментально, якщо немає математичної причини віддати перевагу одному алгоритму кластеризації над іншим. Алгоритм може добре працювати на певному наборі даних, але не працювати для іншого.

Алгоритми кластеризації можна класифікувати на основі кластерної моделі, яка ґрунтується на тому, як саме вони формують кластери або групи. Наприклад, такі як:

а) кластеризація на основі зв'язку: основна ідея кластеризації полягає в тому, що точки даних, які знаходяться ближче в просторі даних, є більш спорідненими (подібними). Кластери формуються шляхом з'єднання точок даних відповідно до їх відстані. На різних відстанях будуть формуватися різні кластери, які можуть бути представлені за допомогою дендрограми (це також називають «ієрархічною кластеризацією»). Ці методи створюють саме ієрархію, а не унікальне розбиття набору даних, з якої

користувачі все ще повинні обрати відповідні кластери, тобто рівень, на якому вони хочуть кластерувати. Вони також не дуже стійкі до викидів, які можуть з'являтися у вигляді додаткових кластерів або навіть призводити до злиття інших кластерів.

б) кластеризація на основі центроїда: у цьому типі кластери представлені центральним вектором або центроїдом. Цей центроїд не обов'язково має бути в наборі даних. Це ітеративні алгоритми кластеризації, в яких поняття подібності виводиться з того, наскільки близька точка даних до центроїда кластера. k-середніх, наприклад, є кластеризацією на основі центроїда.

2. *Методи кластеризації*

Методи, які ми розглянемо в рамках цієї роботи, є одними з найбільш популярних методів кластеризації — метод k-середніх та ієрархічний метод.

Метод k-середніх

Метод k-середніх є широко використовуваним методом кластеризації, який прагне мінімізувати середню квадратичну відстань між точками в одному кластері. Незважаючи на те, що він не гарантує абсолютної точності, його простота і швидкість це компенсують. Доповнюючи метод k-середніх простою рандомізованою технікою висіву, отримуємо алгоритм, який є O($log\ k$)-конкурентним з оптимальною кластеризацією. Експерименти показали, що таке доповнення досить сильно збільшує швидкість і точність k-середніх.

Перевагою методу k-середніх є те, що він більш зручний для кластеризації великої кількості спостережень, ніж, наприклад, метод

ієрархічного кластерного аналізу, де дендограми швидко стають перевантаженими і мають тенденцію втрачати наочність.

Із недоліків методу можна відзначити те, що він чутливий до викидів, які можуть викривити середнє значення, та той факт, що кількість кластерів (що і являє тут собою k) має бути визначена спеціалістом заздалегідь.

У загальному сценарії кластеризація використовується у випадку, коли треба дізнатися більше про представлений набір даних. Таким чином є сенс запустити процес кластеризації не один раз, дослідити цікаві кластери і записати отримані дослідження. Кластеризація — це скоріше інструмент, який допомагає краще дослідити набір даних, але який не завжди повинен використовуватися як автоматичний метод класифікації даних. Вони бувають досить ненадійні, і одна кластеризація сама по собі не зможе дати всю інформацію, яку можна отримати з набору даних.

Основна ідея методу k-середніх полягає у визначенні кластерів таким чином, щоб мінімальна сумарна варіація між кластерами (тобто загальна варіація в межах кластера) була мінімізована. Існує декілька алгоритмів для цього методу, однак стандартний алгоритм визначає загальну варіацію в межах кластера як суму квадратичних евклідових відстаней між елементами і відповідним центроїдом:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Де x_i — точка даних, що належить до кластера C_k, а μ_k — середнє значення точок, що присвоєні кластеру C_k.

Кожне спостереження x_i присвоюється заданому кластеру таким чином, щоб сума квадратів відстані спостереження до їх призначених кластерних центрів μ_k мінімізувалася.

Алгоритм виглядає так:

1. Визначити k — кількість кластерів, які будуть створені.

2. Вибрати випадкові k об'єктів з набору даних у ролі початкових центрів кластерів.

3. Призначити кожне спостереження найближчому центроїду на основі евклідової відстані між об'єктом і центроїдом.

4. Для кожного з k кластерів перерахувати центроїд кластера шляхом обчислення нового середнього значення усіх точок даних у кластері.

5. Ітеративно мінімізувати загальну суму в межах площі. Повторити кроки 3 і 4, поки центроїди не зміняться або не буде досягнуто максимальної кількості ітерацій (R за замовчуванням використовує 10 як максимальну кількість ітерацій).

Загальна сума, або загальна варіація в межах кластера визначається таким чином:

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Ієрархічна кластеризація

Ієрархічні алгоритми кластеризації, або алгоритми таксономії, будують не одне розбиття вибірки на непересічні класи, а систему вкладених розбиттів. Результат таксономії зазвичай представляється у вигляді таксономічного дерева — дендрограми. Класичним прикладом такого дерева є ієрархічна класифікація тварин і рослин.

Дендограми дозволяє уявити кластерну структуру у вигляді плаского графіка незалежно від того, яка розмірність початкового простору. Існують і інші способи візуалізації багатовимірних даних, такі як багатовимірне шкалювання або карти Кохонена, але вони привносять в картину штучні спотворення, вплив яких досить важко оцінити.

Є два типи методів:

1. Агломератні методи: нові кластери утворюються шляхом об'єднання дрібніших кластерів, і таким чином дерево створюється від листя до стовбура.
2. Дивізійні методи: нові кластери створюються шляхом ділення більших кластерів на більш дрібні, і таким чином дерево створюється від стовбура до листя.

Подібність кластерів часто розраховується через «неподібність», наприклад, евклідова відстань між двома кластерами. Отже, чим більше відстань між двома кластерами, тим краще.

Ключовою операцією в ієрархічній агломераційній кластеризації є неодноразове об'єднання двох найближчих кластерів у один кластер, але дуже важливо спочатку відповісти на три питання: як ви представити кластер з більш ніж однією точкою, як визначити «близькість» кластерів та коли перестати поєднувати кластери.

Злиття кластерів припиняється в залежності від доступної інформації про дані, які ми маємо. Якщо групувати футболістів на полі на основі їхніх позицій на полі, яке представлятиме їх координати для розрахунку відстані між гравцями, очевидно, що треба зупинитися на лише двох кластерах, оскільки можуть бути тільки дві команди, які грають у футбольний матч.

Алгоритм методу виглядає таким чином:

1. Обчислити матрицю близькості, що містить відстань між кожною парою шаблонів. Розглядати кожен зразок як окремий кластер.

2. Знайти найбільш схожу пару кластерів за допомогою матриці близькості. Об'єднати ці два кластера в один більший кластер. Оновити матрицю близькості, щоб відобразити цю операцію злиття.

3. Якщо всі шаблони знаходяться в одному кластері, зупинитися. В іншому випадку перейти до кроку 2.

Ієрархічний алгоритм дає дендограму, що представляє собою складене групування шаблонів і рівні схожості, на яких змінюються самі групування.

Більшість ієрархічних алгоритмів кластеризації є варіантами однозв'язного і повнозв'язного алгоритму, а також алгоритму мінімальної дисперсії. Найбільш популярними з них є однозв'язний та повнозв'язний алгоритми. Вони відрізняються тим, як вони характеризують подібність між парою кластерів.

У методі з однозв'язним алгоритмом відстань між двома кластерами — це мінімум відстаней між усіма парами шаблонів, взятих з двох кластерів (один з першого кластера, другий — з другого).

У повнозв'язному алгоритмі відстань між двома кластерами — це максимум усіх попарних відстаней між шаблонами в двох кластерах. В обох випадках два кластери об'єднуються для формування більшого кластера на основі мінімальних критеріїв відстані. Повнозв'язний алгоритм створює щільні та компактні кластери, а однозв'язний алгоритм, навпаки, страждає від ланцюгового. Він має тенденцію виробляти кластери, які є занадто громіздкими або подовженими.

3. *Дослідження набору даних методами кластеризації*

Спробуємо дослідити набір даних розглянутими вище методами кластеризації. Використаємо набір даних під назвою «Іриси Фішера», на прикладі якого Рональд Фішер продемонстрував роботу розробленого ним методу дискримінантного аналізу. Також називають ірисами Андерсона, так як дані були зібрані американським ботаніком Едгаром Андерсоном. Цей набір даних вважається класичним і часто використовується в літературі для ілюстрації роботи різних статистичних алгоритмів.

Для дослідження будемо використовувати середу розробки R. Поглянемо на дані:

```
library(datasets)
data(iris)
str(iris)
summary(iris)
```

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                  
```

Цей відомий набір даних про іриси дає вимірювання в сантиметрах по довжині та ширині чашолистка та довжині та ширині пелюстки відповідно для 50 квітів кожного з 3 видів ірису. Видами є *iris setosa*, *versicolor* та *virginica*.

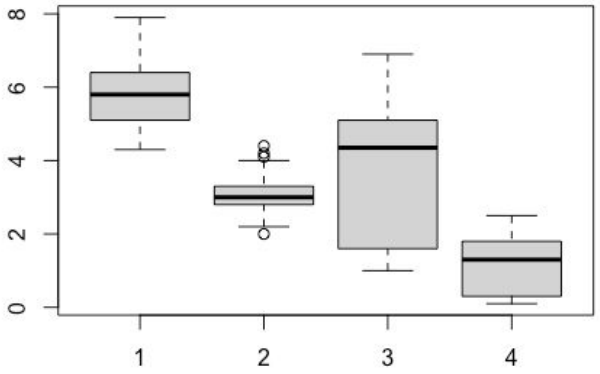
Ще трохи поглянемо на набір даних. Подивимося, які наявні кореляції між змінними.

```
> cor(iris[1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Бачимо, що кореляції присутні і досить значні. Наприклад, 87% кореляції між довжинами пелюсток і чашолистків та 96% між шириною та довжиною пелюстків.

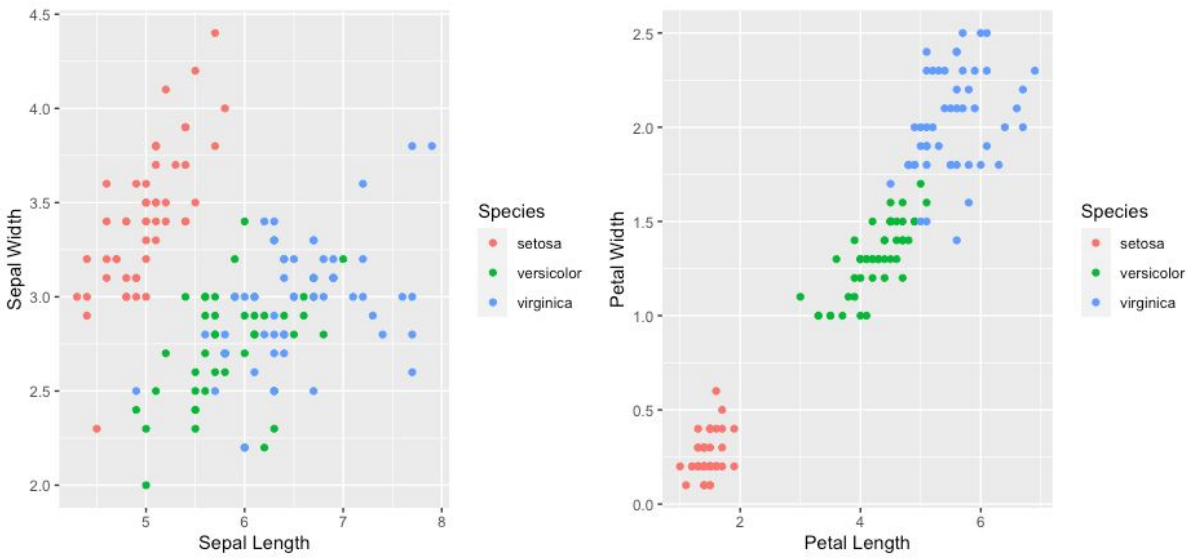
Також подивимося на наявність очевидних викидів в даних. Бачимо, що нічого критичного, особливих викидів немає, лише трохи в змінній довжині чашолистка.

```
boxplot(iris$Sepal.Length, iris$Sepal.Width,
iris$Petal.Length, iris$Petal.Width)
```



Графіки довжини і ширини пелюсток та чашолистків:

```
g1 <- ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width,
color = Species)) +
        geom_point() +
        labs(x = "Sepal Length", y = "Sepal Width")
g2 <- ggplot(iris, aes(x = Petal.Length, y = Petal.Width,
color = Species)) +
        geom_point() +
        labs(x = "Petal Length", y = "Petal Width")
grid.arrange(g1, g2, nrow = 1)
```



Перейдемо до методу k-середніх.

```
set.seed(99)
irisCl <- kmeans(iris[,1:4], center = 3, nstart = 20)
irisCl
clusplot(iris, irisCl$cluster, color=T, labels=0, lines=0)
```

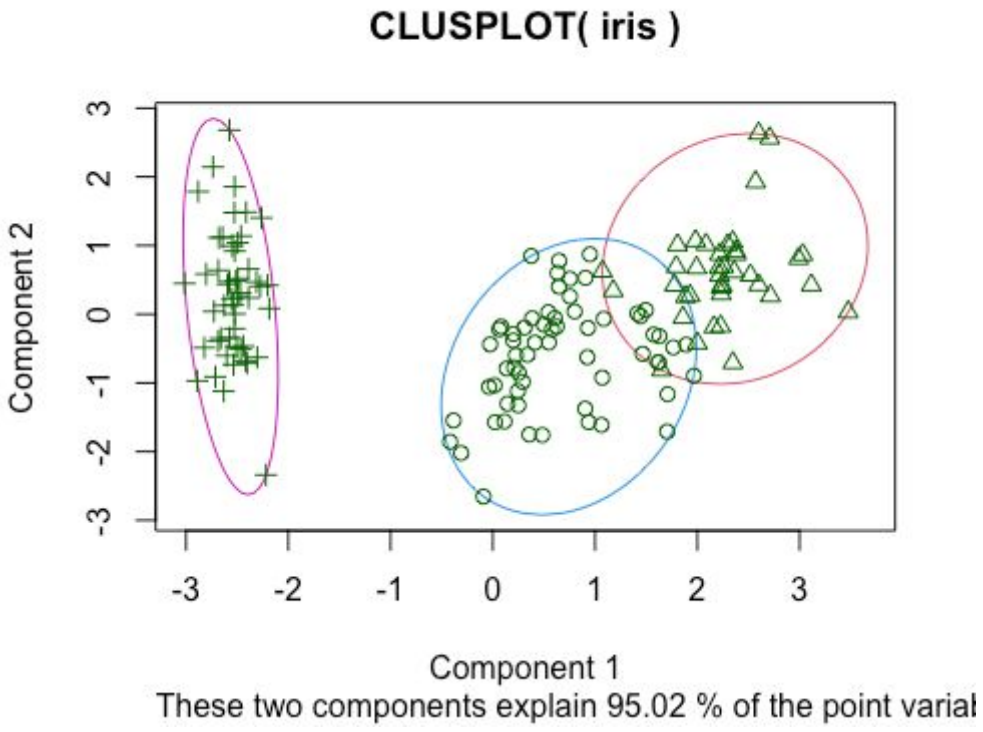
```
> irisCl
K-means clustering with 3 clusters of sizes 62, 38, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.901613    2.748387     4.393548    1.433871
2     6.850000    3.073684     5.742105    2.071053
3     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [45] 3 3 3 3 3 3 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1
 [89] 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 2
[133] 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1

Within cluster sum of squares by cluster:
[1] 39.82097 23.87947 15.15100
 (between_SS / total_SS =  88.4 %)
```

Кластерів, ми знаємо, три, тому *center = 3*, маємо три кластери з такою кількістю об'єктів: 62, 38, 50. Графік:

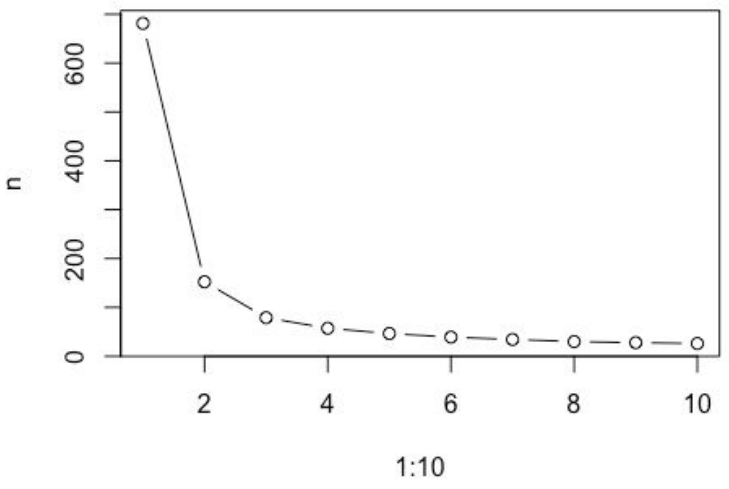


Видно, що кластеризація дозволила пояснити 95.02% даних, *setosa* повністю чітко відображена, а між *virginica* та *versicolor* є деякий шум. Власне, можемо побачити таблицю з правильністю:

	setosa	versicolor	virginica
1	0	48	14
2	0	2	36
3	50	0	0

Ще розглянемо варіант, коли нам невідома кількість кластерів і нам треба самим оцінити оптимальну кількість, що ми зробимо «ліктьовим методом».

```
n <- vector(mode = "character", length = 10)
for (i in 1:10){
irisCluster <- kmeans(iris[,1:4], center = i, nstart = 20)
n[i] <- irisCluster$tot.withinss}
plot(1:10, n, type = "b", pch = 1)
```



Три кластери і є оптимальною кількістю.

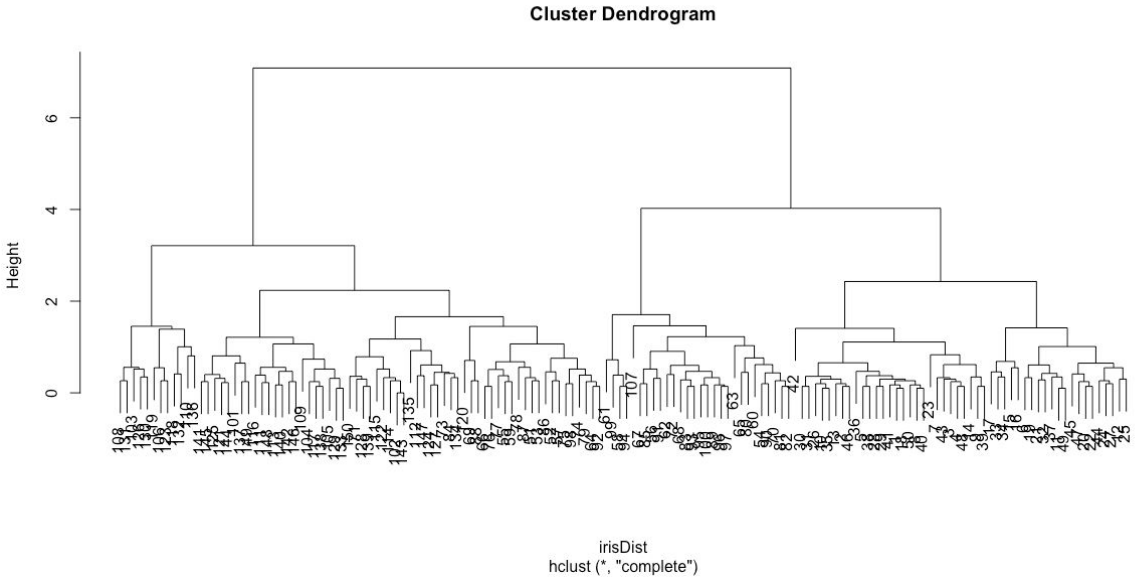
Щодо ієрархічного методу:

```
set.seed(100)
species <- as.list(iris$Species)
species <- unlist(species)
irisDist <- dist(iris[1:4], method = "euclidean")
model <- hclust(irisDist, method = "complete")
modelcut <- cutree(model, 3)
```

```
table(species, modelcut)
```

```
> table(species, modelcut)
            modelcut
species       1  2  3
  setosa     50  0  0
  versicolor  0 23 27
  virginica   0 49  1
```

Дендрограма виглядає таким чином:



4. Оцінка результатів і висновки

У цій роботі ми розглянули такі методи кластеризації, як метод k-середніх та ієрархічний метод. Метод k-середніх дуже широко використовується, він прагне мінімізувати середню квадратичну відстань між точками в одному кластері, а ієрархічний метод будує систему вкладених розбиттів. В нашому випадку ми використовували повнозв'язний алгоритм, де відстань між двома кластерами — це максимум усіх попарних відстаней між шаблонами в двох кластерах.

Ми розглянули ці методі на прикладі класичного набору даних — ірисах Фішера. У випадку методу k-середніх ми обрали ліктьовим методом оптимальну кількість кластерів (три), а потім застосували безпосередньо кластеризацію і отримали точність 88,4%, а розбиття на три кластери дозволило пояснити 95% варіації.

При ієрархічному методі, як і при методі k-середніх, вид *setosa* був чітко визначений на відміну від інших двох через те, що вони в деяких місцях досить схожі, але незважаючи на це, ми побачили, що точність була досить висока, хоч найменш точно було розпізнано вид *versicolor*.