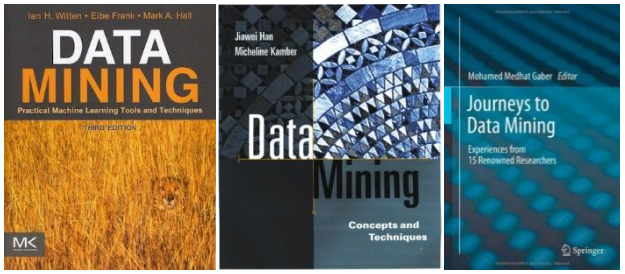
Математические методы в биологии

Блок 4. Многомерный анализ данных

Лекция 9

Козлова Ольга Сергеевна
89276755130, olga-sphinx@yandex.ru

Что такое data mining?



- Это процесс нетривиального извлечения новой, полезной и экстраполируемой информации из большого массива многомерных данных.
- Другими словами, это поиск структуры в данных.
- Исходные данные – совокупность численных векторов (измерений)

Пример. Набор данных iris – 150 наблюдений, представляющих три вида ирисов (50 наблюдений для каждого). Каждый ирис – это вектор вида (Длина_чашелистика, Ширина_чашелистика, Длина_лепе… Ширина_лепестка). Каждый ирис – точка в четырёхмерном …е.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

$\text{Sepal.Length} \in [4.3; 7.9]$ $\text{Sepal.Width} \in [2.0; 4.4]$
$\text{Petal.Length} \in [1.0; 6.9]$ $\text{Petal.Width} \in [0.1; 2.5]$

setosa

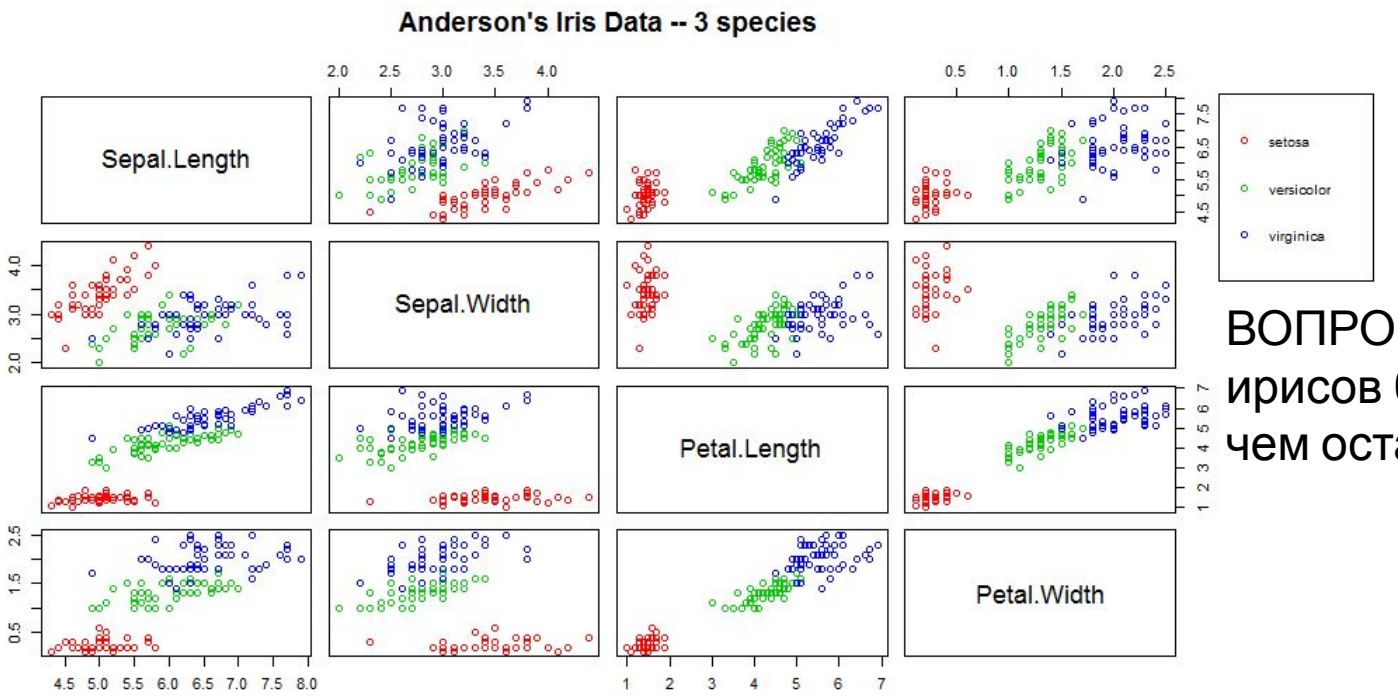
virginica

versicolor

Классификация многомерных методов

- Визуализация
  - Визуализация «сырых» данных (данные как они есть)
  - Методы понижения размерности
    - Анализ главных компонент
    - Кластеризация
- Классификация
  - Деревья принятия решений
  - …

Простая визуализация «сырых» данных:



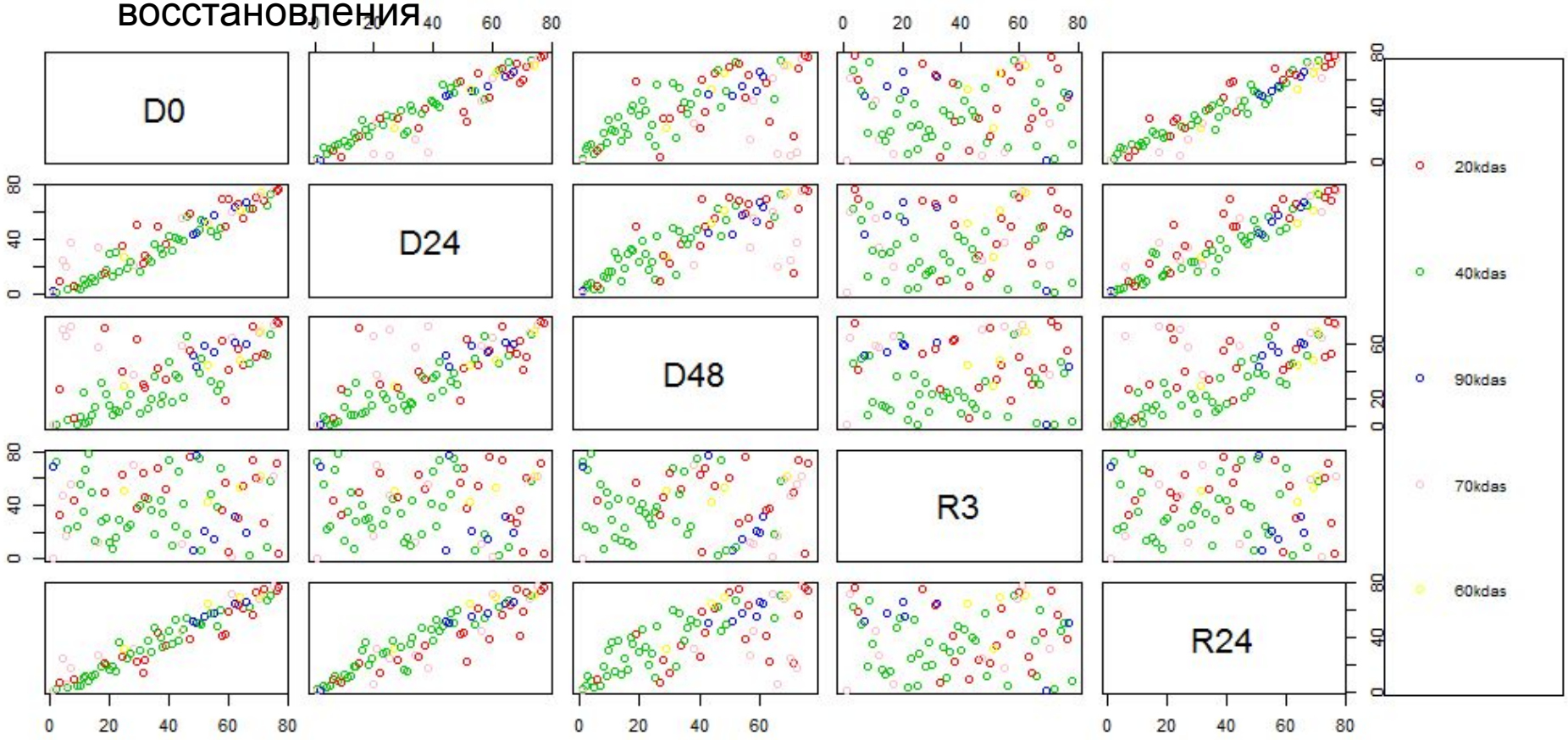
ВОПРОС: какой из видов ирисов более «другой», чем остальные?

Ещё один пример «парной» визуализации:

Белки теплового шока ангидробиотической хирономиды

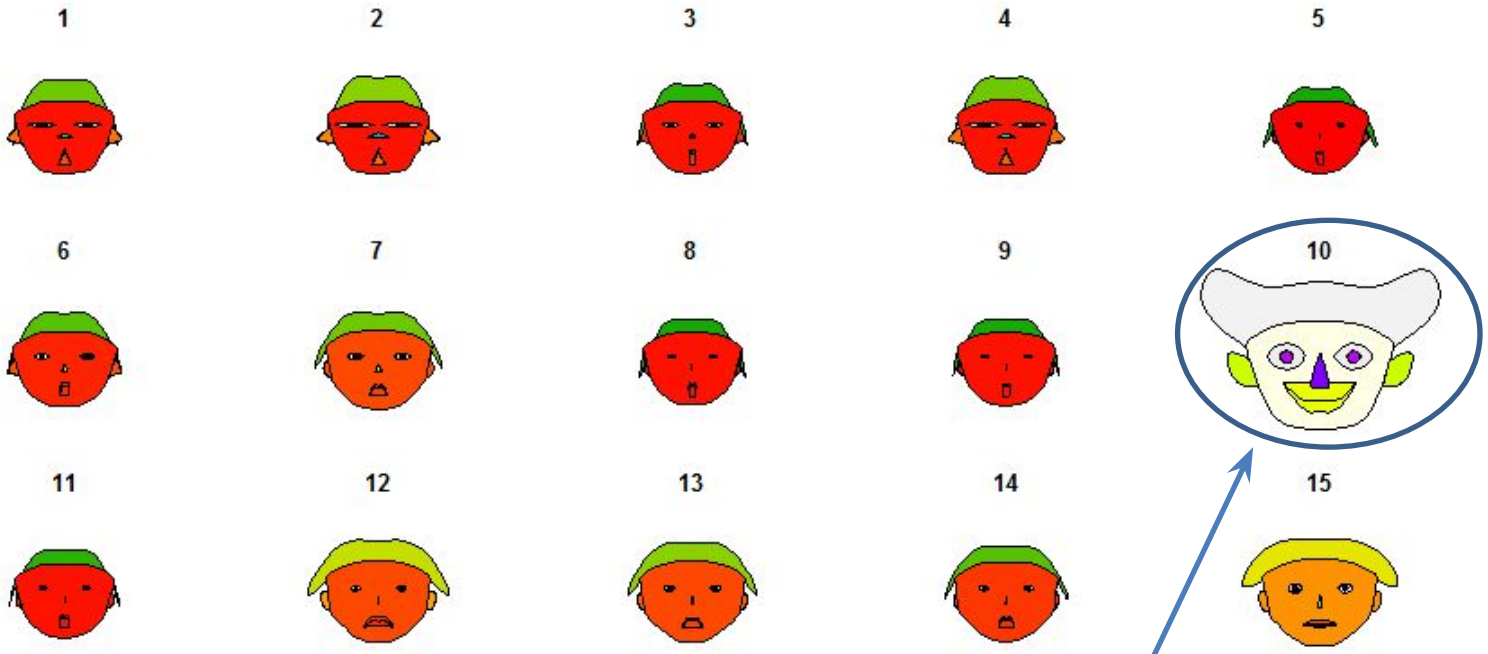
Polypedilum vanderplanki (Африка)

- Ответ на 48-часовое обезвоживание
- Возврат к исходному уровню за 24 часа восстановления



Пиктограммы – весёлый и лёгкий способ находить похожие объекты

- Лица Чернова



Набор из 15 HSP *P. vanderplanki*

D0: высота лица, тип волос, улыбка

D24: высота глаз, ширина лица, высота носа

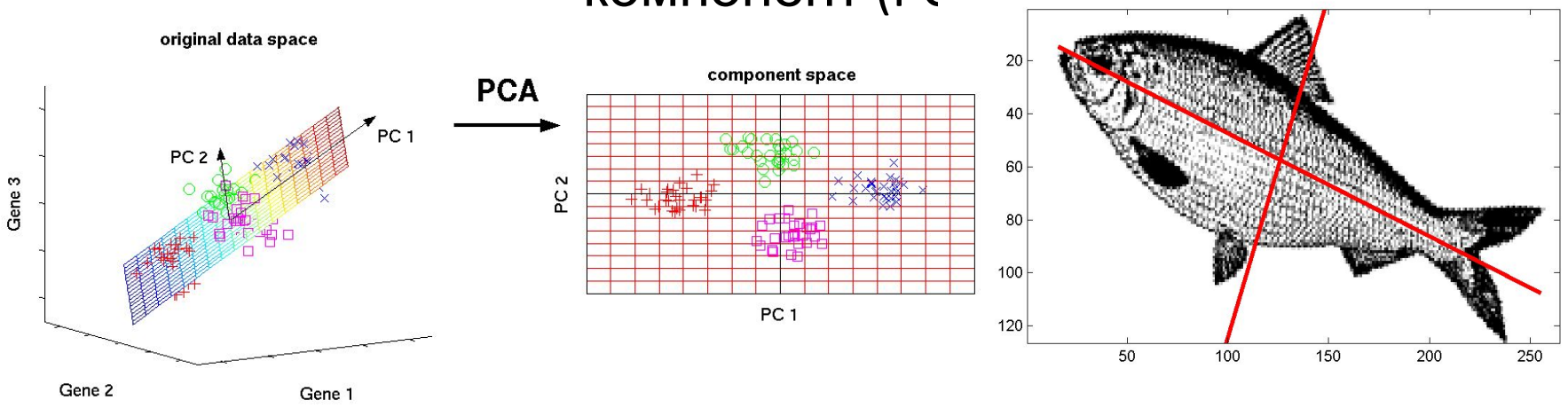
D48: ширина глаз, тип лица, ширина носа

R3: ширина уха, высота рта, высота волос

R24: ширина рта, ширина волос, высота уха

Как вы думаете, «кто» это?

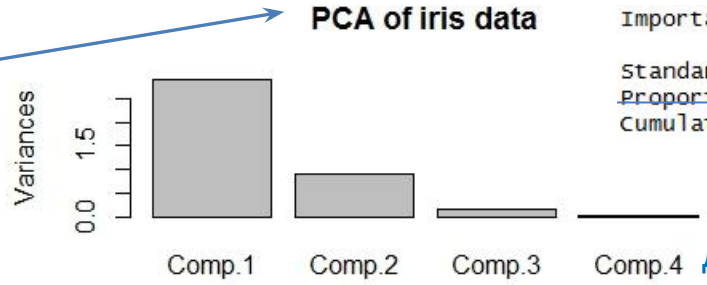
Методы понижения размерности: анализ главных компонент (PCA)



Идея. Каждый объект – точка в n-мерном Евклидовом пространстве, весь массив данных – облако точек. Требуется найти новые оси, которые будут наилучшим образом объяснять изменчивость.

1я главная компонента – прямая, секущая облако в направлении его максимальной изменчивости (линия регрессии, по сути). 2я главная компонента перпендикулярна 1й в наиболее «широком» месте.

Служебный график осыпи (scree plot)




```
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     1.7026571 0.9528572 0.38180950 0.143445939
Proportion of Variance 0.7296245 0.2285076 0.03668922 0.005178709
Cumulative Proportion  0.7296245 0.9581321 0.99482129 1.000000000
```

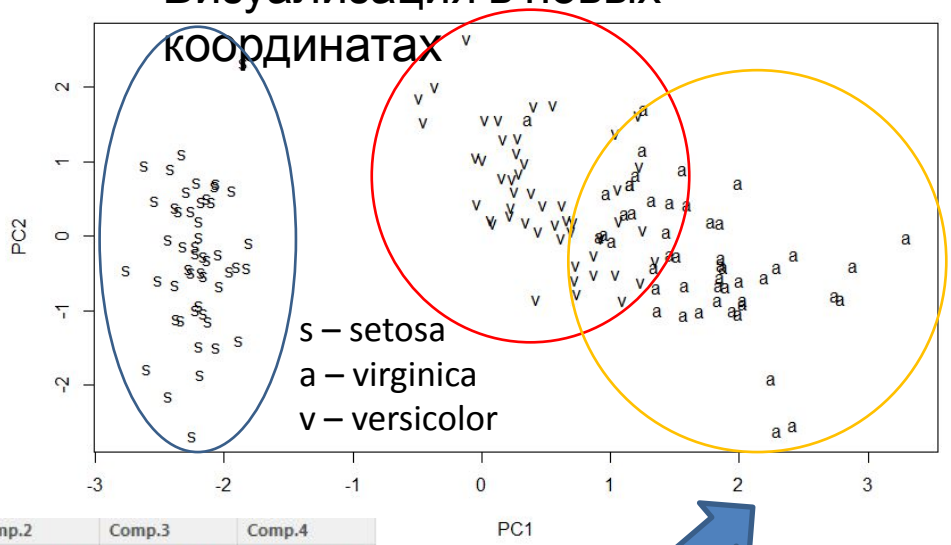
Доля объяснённой дисперсии. Первые 2 гл. компоненты объясняют почти 96% дисперсии!

Как преобразовать 4х-мерное пространство к 2х-мерному?

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa

Исходные данные

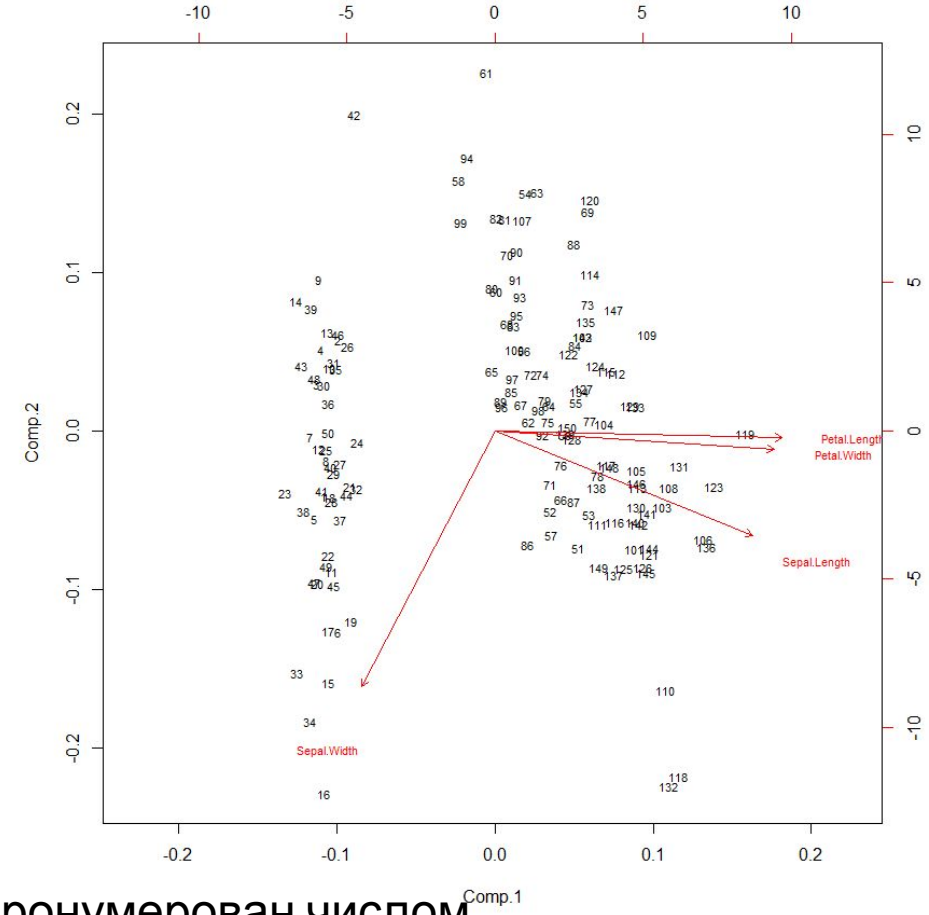
Визуализация в новых координатах



	Comp.1	Comp.2	Comp.3	Comp.4
1	-2.25714118	-0.478423832	-0.127279624	0.024087508
2	-2.07401302	0.671882687	-0.233825517	0.102662845
3	-2.35633511	0.340766425	0.044053900	0.028282305
4	-2.29170679	0.595399863	0.090985297	-0.065735340
5	-2.38186270	-0.644675659	0.015685647	-0.035802870
6	-2.06870061	-1.484205297	0.026878250	0.006586116
7	-2.43586845	-0.047485118	0.334350297	-0.036652767
8	-2.22539189	-0.222403002	-0.088399352	-0.024529919
9	-2.32684533	1.111603700	0.144592465	-0.026769540
10	-2.17703491	0.467447569	-0.252918268	-0.039766068
11	-2.15907699	-1.040205867	-0.267784001	0.016675503
12	-2.31836413	-0.132633999	0.093446191	-0.133037725
13	-2.21104370	0.726243183	-0.230140246	0.002416941
14	-2.62430902	0.958296347	0.180192423	-0.019151375
15	-2.19139921	-1.853846555	-0.471322025	0.194081578
16	-2.25466121	-2.677315230	0.030424684	0.050365010
17	-2.20021676	-1.478655729	-0.005326251	0.188186988
18	-2.18303613	-0.487206131	-0.044067686	0.092779618
19	-1.89223284	-1.400327567	-0.373093377	0.060891973
20	-2.33554476	-1.124083597	0.132187626	-0.037630354
21	-1.90793125	-0.407490576	-0.419885937	0.010884821
22	-2.19964383	-0.921035871	0.159331502	0.059398340
23	-2.76508142	-0.456813301	0.331069982	0.019582826
24	-1.81259716	-0.085272854	0.034373442	0.150636353

Данные в новых координатах

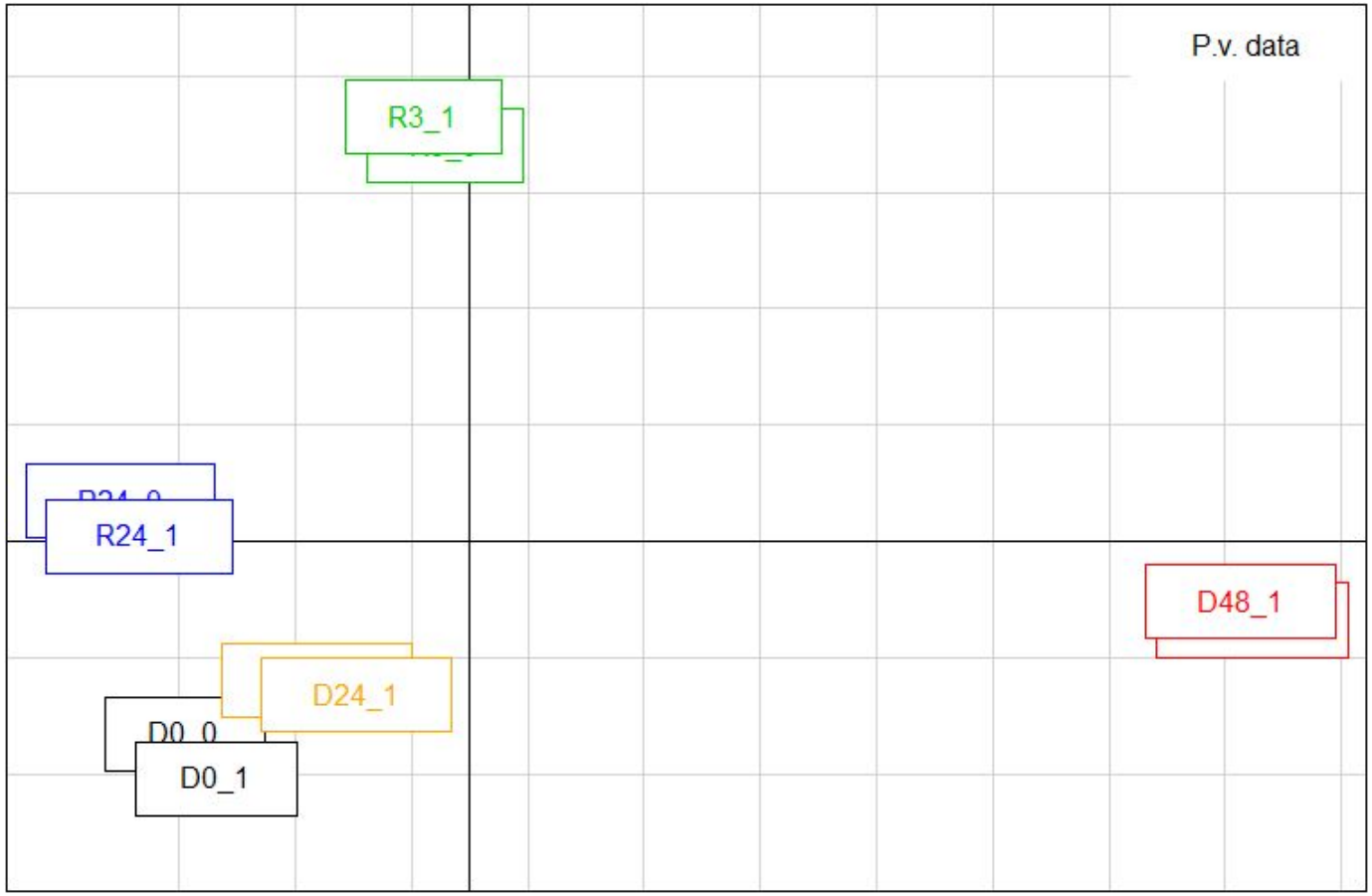
График biplot графически увязывает старые и новые координаты



- Каждый ирис пронумерован числом
- Чем меньше угол – тем больше корреляция
- Чем вектор параллельней новой оси – тем больше вклад

Применение метода главных компонент для анализа дифференциальной экспрессии

- Проверка самосогласованности реплик (повторностей)

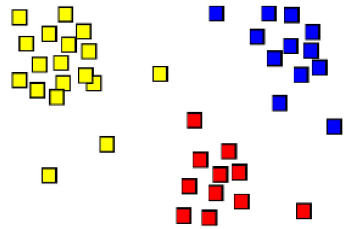


Каждый объект – вектор из нескольких десятков чисел (уровни экспрессии всех HSP *P. vanderplanki*)

Две повторности в каждом эксперименте (и контроле)

Реплики кластеризуются вместе + видно, какие образцы близки друг другу, а какие – нет.

Методы понижения размерности: кластеризаци



- Кластеризация – разбиение большого набора объектов на более мелкие наборы (кластеры)
- Основная идея: **объекты внутри кластера должны быть более «похожи» между собой, нежели объекты из разных кластеров.**
- Для того чтобы формировать кластеры, мы должны научиться измерять расстояния (метрики) между объектами

Основные метрики:

- Расстояние Евклида (1)
- Квадрат расстояния Евклида (2)
- Расстояние Чебышева (3)
- Манхэттенское расстояние (4)

(1) $$\rho(x,x') = \sqrt{\sum_i^n (x_i - x_i')^2}$$

(2) $$\rho(x,x') = \sum_i^n (x_i - x_i')^2$$

(3) $$\rho(x,x') = \max(|x_i - x_i'|)$$

(4) $$\rho(x,x') = \sum_i^n |x_i - x_i'|$$

Классификация методов кластеризации

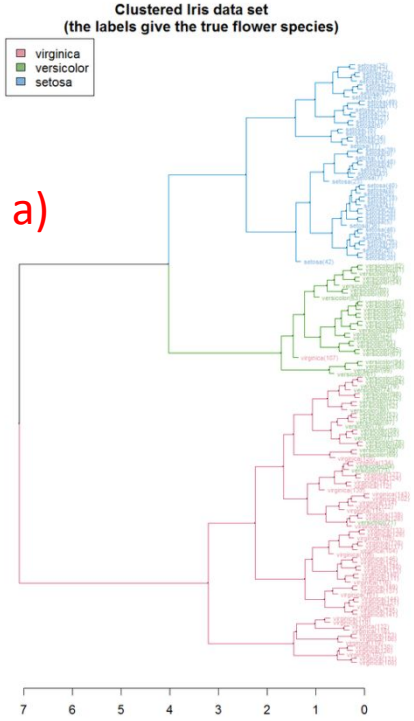
- Иерархическая / плоская

Комплексная древоподобная система разбиений a) / одно и только одно разбиение на кластеры одного и того же уровня b)

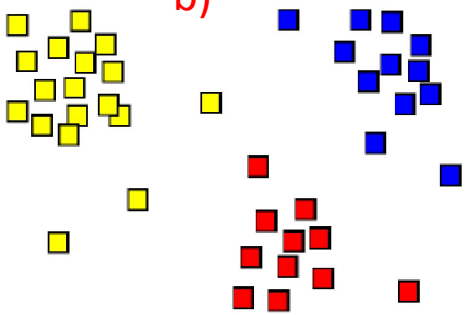
- Точная / неточная

Каждый объект принадлежит только одному кластеру c) / каждый объект может принадлежать разным кластерам со своими вероятностями d)

a)



b)



c)

Объекты	O1	O2	O3	O4	O5
Кластеры	C1	C3	C3	C2	C6

d)

Объекты	O1	O2	O3
Вектор вероятностей	{0,0,0.45,0.55}	{1,0,0,0}	{0.3,0.3,0.4,0}

Кластеризация методом k-средних (k-means)

Основные «правила игры»:

1. k – число кластеров – выбирается заранее
2. Начальные координаты центров кластеров выбираются случайным образом (рис.1)
3. Основная идея – минимизировать целевую функцию $\sum_{i=1}^{n} {d_i}^2$, где n – число объектов в кластере, а d_i – расстояние между i-ым объектом и центром кластера (рис.2)
4. На каждой итерации d – центр кластера – сдвигается в центр масс (точку, каждая координата которой – среднее соответствующих координат объектов кластера) (рис.3)

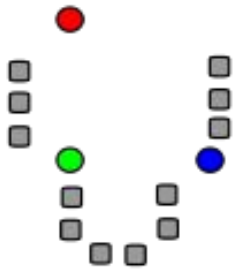


Рис. 1

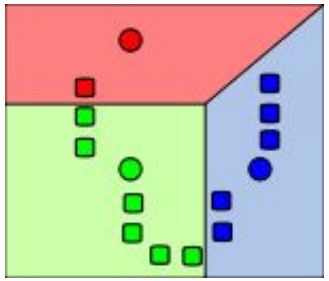


Рис. 2

(Фактически, каждая точка окрашивается в цвет того центра, к которому она ближе

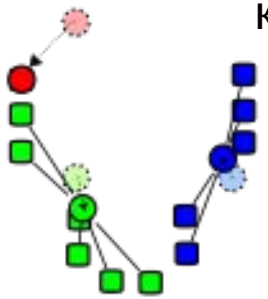


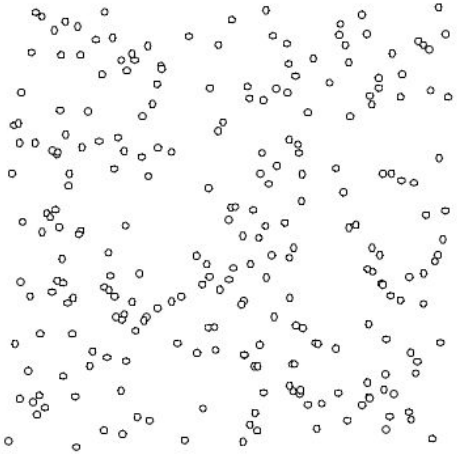
Рис. 3

итоговая кластеризация

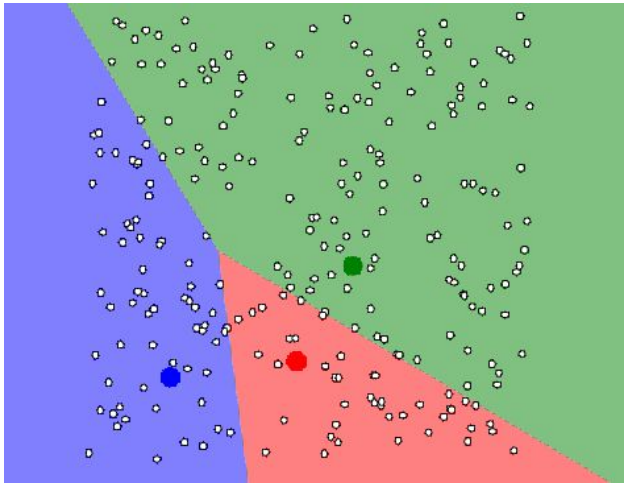
Замечательная визуализация!

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

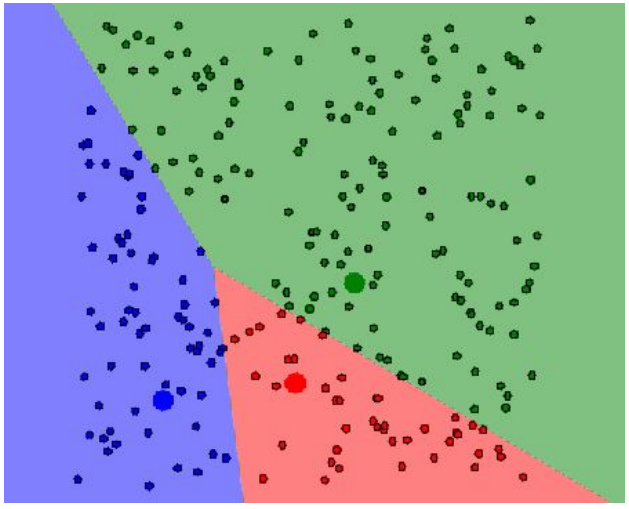
Шаг 0. Начальное положение точек кластеров



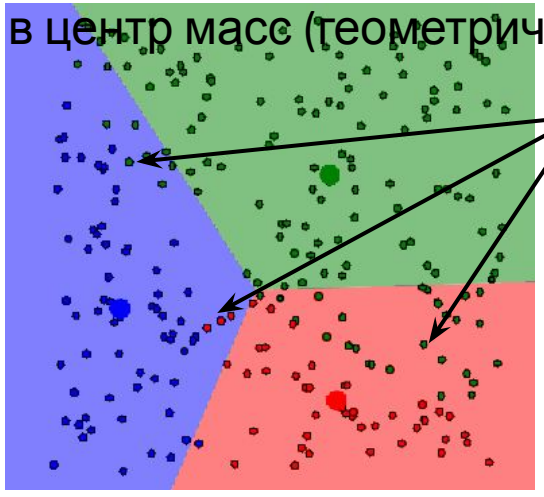
Шаг 1. Бросаем начальные центры



Шаг 2. «Раскрашиваем» точки по принципу ближайшего центра

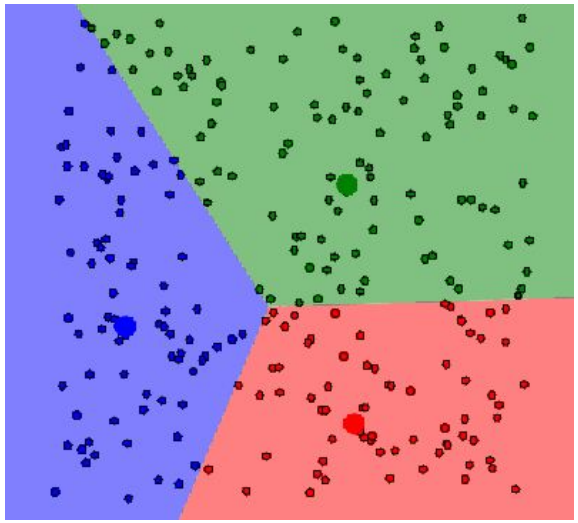


Шаг 3. Переставляем центры кластеров в центр масс (геометрический центр)

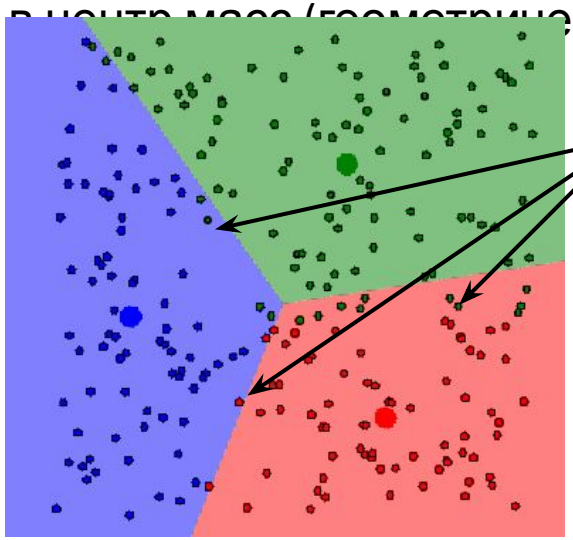


Точки, которые находятся «на чужой территории»

Шаг 4. «Перекрашиваем» точки, которые находятся «на чужой территории»

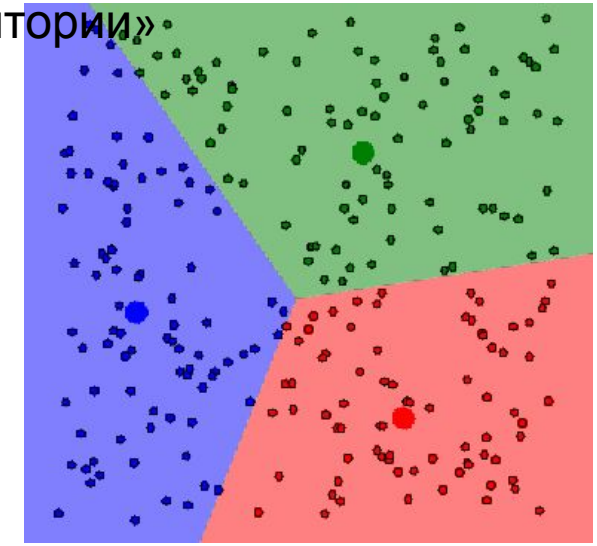


Шаг 5. Переставляем центры кластеров в центр масс (геометрический центр)

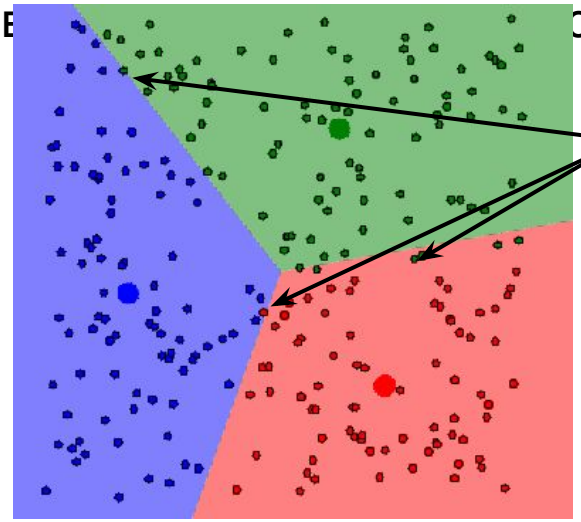


Точки, которые находятся «на чужой территории» (стало меньше!)

Шаг 6. «Перекрашиваем» точки, которые находятся «на чужой территории»

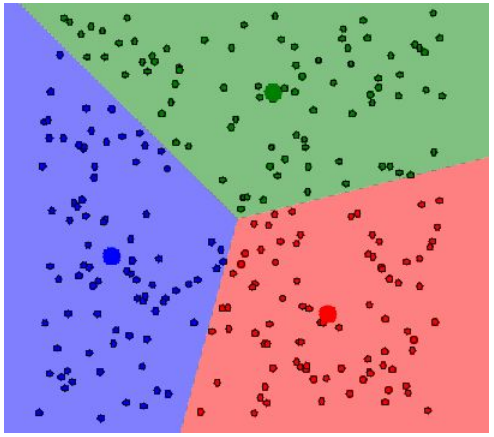


Шаг 7. Переставляем центры кластеров в центр масс (геометрический центр)



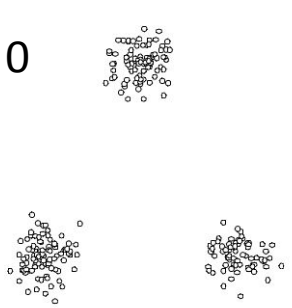
Точки, которые находятся «на чужой территории» (ещё меньше!)

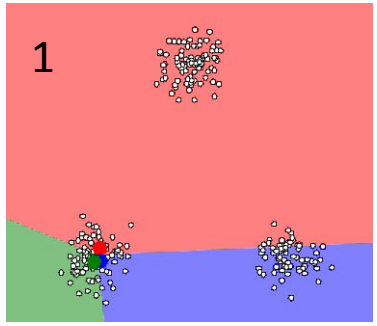
И так до тех пор, пока есть что «перекрашивать»!

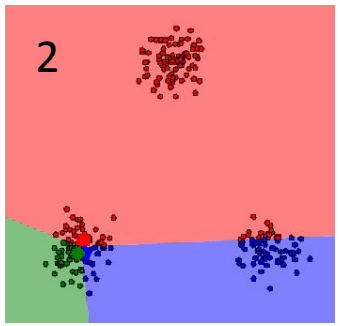


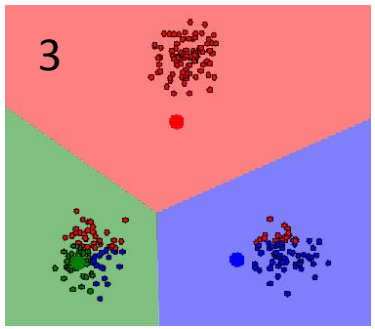
Финальная «раскраска» – после очередного перемещения центров кластеров ни одна из точек не оказалась «на чужой территории»

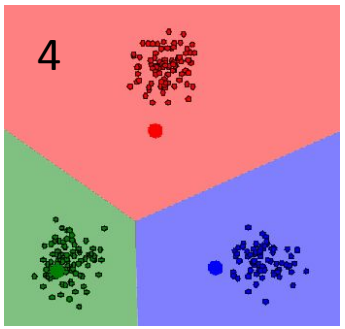
Чем более явные кластеры в данных, тем быстрее сойдётся

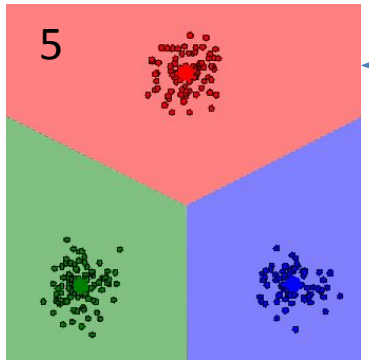












Финальная «раскраска»

Как помочь анализу методом k-средних?

Совет 1. Максимально растаскивать начальные центры кластеров

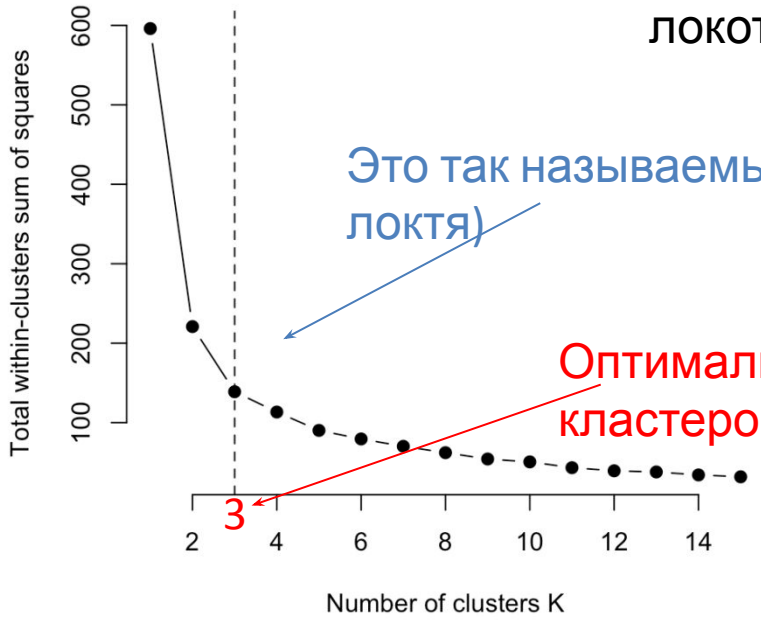
Совет 2. Повторить кластеризацию несколько раз

Совет 3. Разумно выбирать число кластеров

SSW – внутригрупповая сумма квадратов расстояний точек от центра (наша целевая функция $\sum_{i=1}^{n} {d_i}^2$, по сути)

Можно нарисовать график зависимости $\sum_{\text{по всем кластерам}} SSW$ как функции от числа кластеров:

Чем более явные кластеры в данных, тем круче локоть!



Это так называемый elbow-plot (график локтя)

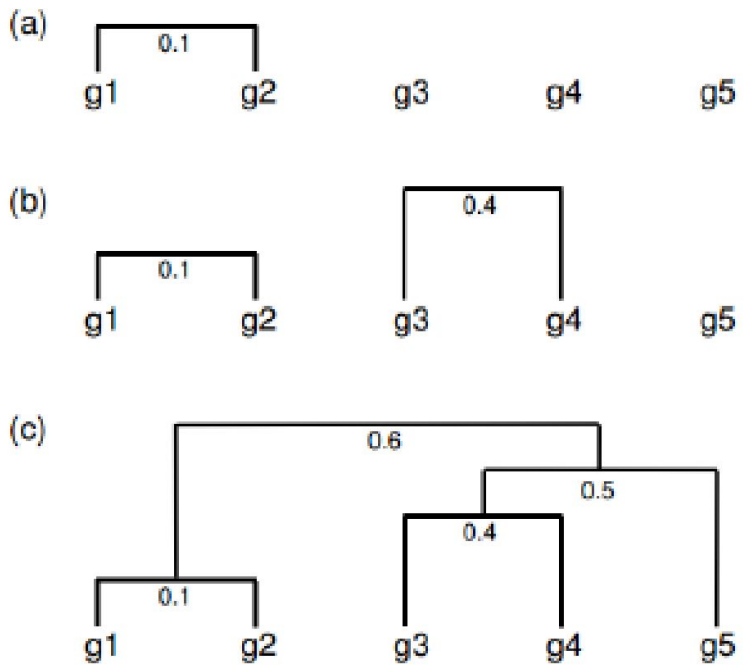
Оптимальное число кластеров

Если добавление ещё одного кластера значительно понижает $\sum SSW$, оно имеет смысл.

Иерархическая кластеризация

Два принципиально разных подхода:

- Снизу-вверх (каждая точка – один кластер, дальше кластеры объединяются в кластеры более высокого порядка)
- Сверху-вниз (всё множество точек – один кластер наивысшего порядка, а затем он делится на множество более мелких)

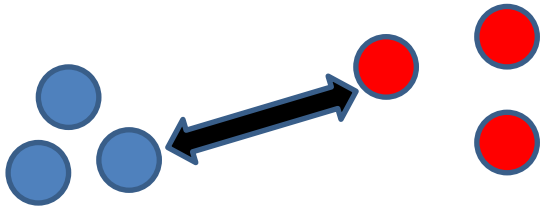


Подход снизу-вверх

Как вычислять расстояния между кластерами?

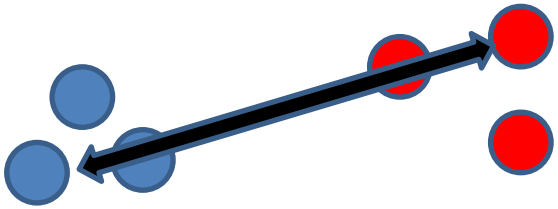
- Метод ближайшего соседа

(метод одиночной связи)



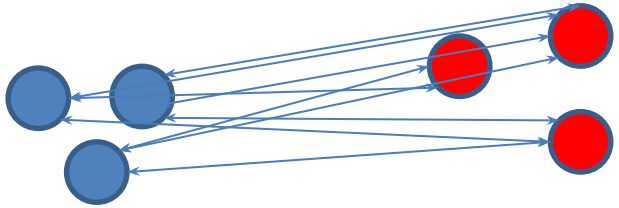
- Метод дальнего соседа

(метод полной связи)

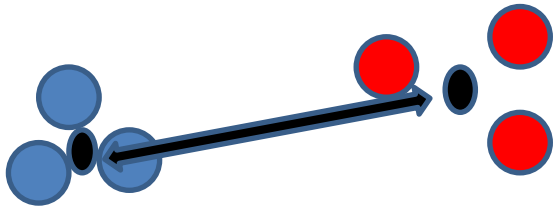


- Метод попарных средних

$$\sum_{\substack{i\in\{1,\dots,N\}\\ j\in\{1,\dots,M\}}} d_{ij}/NM$$



- Центроидный метод



Иерархическая кластеризация 30 ирисов (по 10 каждого вида)

a – virginica, s – setosa, v - versicolor

Задача классификации

- Похожа на кластеризацию, но деление на группы происходит с учётом конкретных признаков объектов

Например, классификация биологических видов

- Классификация – пример обучения с учителем:

Набор исходных данных делится на 2 множества – обучающее и тестовое:

1) Обучающее используется для конструирования модели (≈70% общего объёма данных)
2) Тестовое используется для проверки модели (≈30% общего объёма данных)

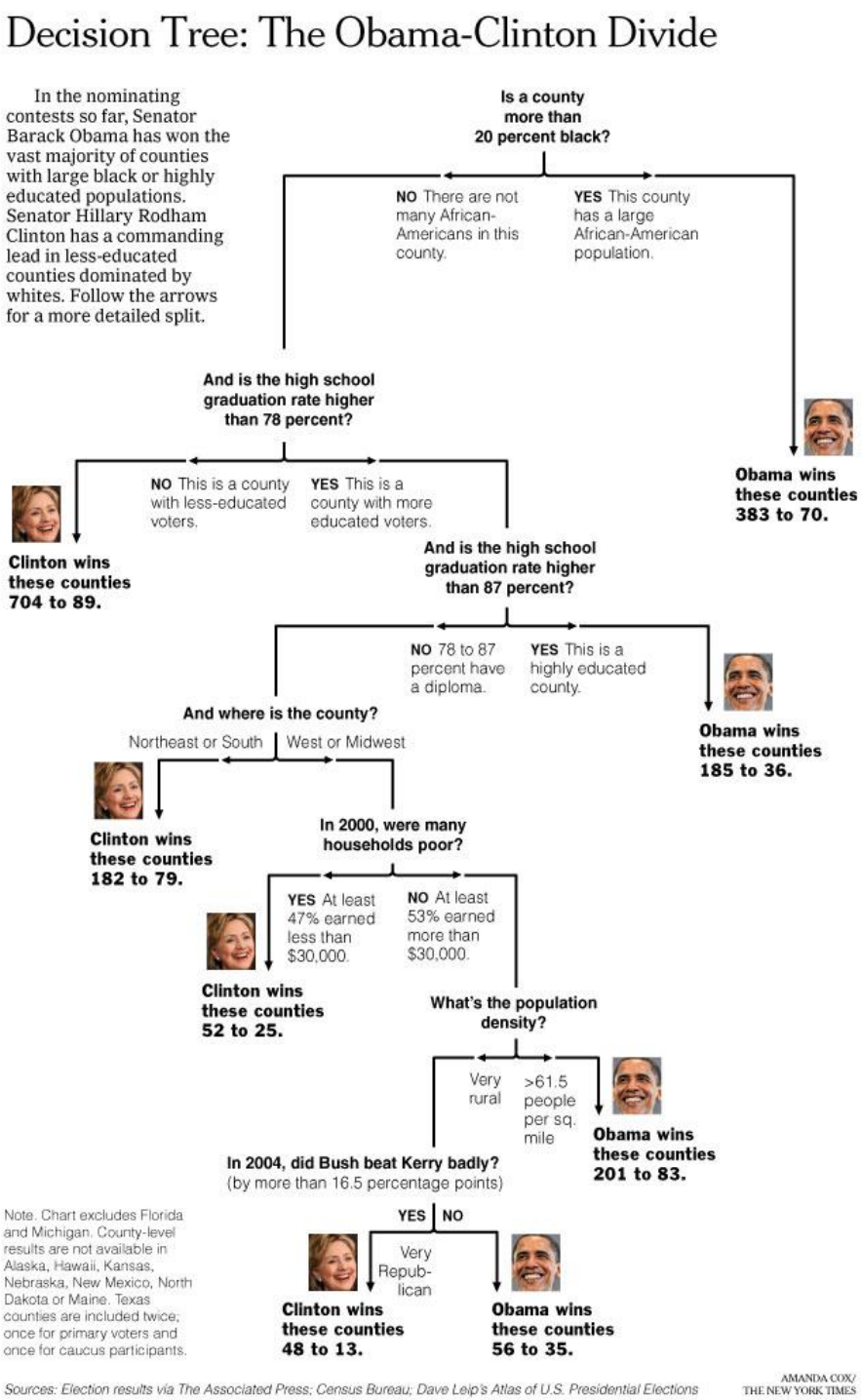
Таким образом, процесс классификации состоит из двух этапов:

конструирования модели и её использования.

Уровень точности (то есть доля верно классифицированных объектов) для тестовой выборки должен соответствовать уровню точности для обучающей!

Базовый алгоритм классификации

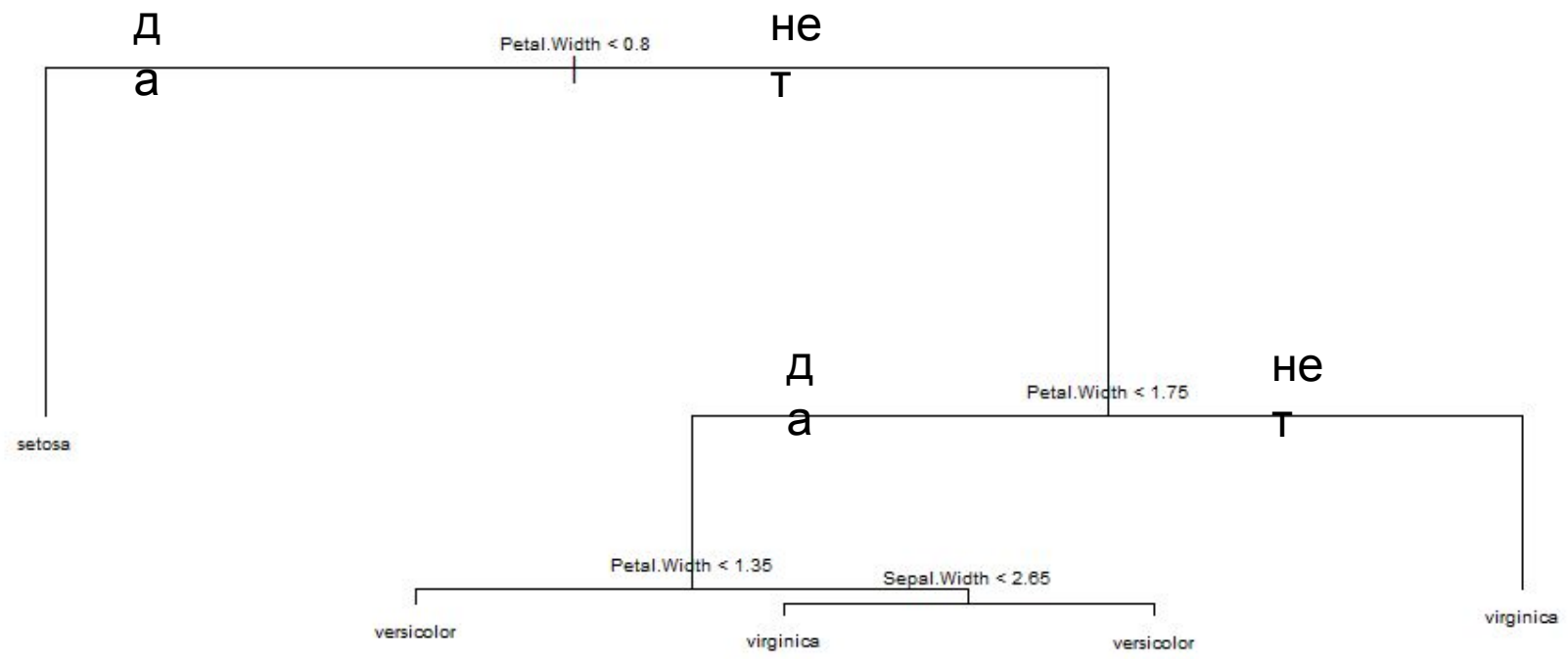
1. Находим параметр, по которому группа разделяется лучше всего
2. Делим данные на 2 группы (листья)
3. Внутри каждой группы снова находим параметр, разделяющий группу лучше всего
4. Продолжаем, пока листья не окажутся достаточно маленькими или «чистыми»



Дерево принятия решений – наиболее популярный, простой и интуитивно понятный метод решения задач классификации

Его результат – древовидная структура, на каждом узле которой задаётся вопрос, и разделение происходит в зависимости от ответа (да/нет).

Дерево принятия решений для ирисов



Спасибо за внимание!

До встречи на практике!