Кластеризация многомерных данных

Общепринято, что какую-либо обработку статистических данных (усреднение, установление связей и т. д.) надо производить только в однородных группах наблюдений.

И. Д. Мандель, «Кластерный анализ»

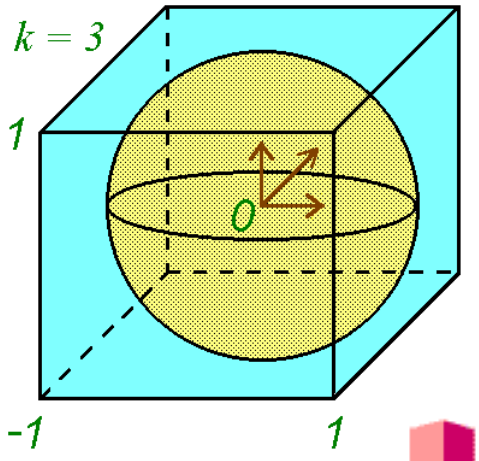
«Проклятие размерности»

При использовании статистических методов для анализа таблицы данных могут возникнуть неожиданности, связанные с тем, что наша обычная геометрическая интуиция способна навести на неверное представление о k-мерных множествах.

Рассмотрим k-мерный шар $\{\boldsymbol{x} : x_1^2 + \ldots + x_k^2 \leqslant 1\}$, вписанный в k-мерный куб $\{\boldsymbol{x} : |x_j| \leqslant 1, \quad j = 1, \ldots, k\}$.

Вероятность p_k, что выбранная случайно в кубе точка окажется внутри шара, равна отношению объема k-мерного шара радиуса $r = 1$ к объему k-мерного куба со стороной 2. Очевидно, $p_2 = \pi r^2/2^2 = \pi/4 \approx 0{,}785$; $p_3 = \frac{4}{3}\pi r^3/2^3 = \pi/6 \approx 0{,}524$.

Вопрос. *Что подсказывает Вам интуиция о порядке этой вероятности при* $k = 10$*?*



Вероятность $p_{10} \approx 2{,}5 \cdot 10^{-3}$. В результате в среднем только одна из 400 случайных точек попадает внутрь вписанного десятимерного шара.

Практическое задание 1

1) Напишите программу на языке R для моделирования n-кратного выбора точек наудачу из k-мерного куба $[-1, 1]^k$ и подсчёта частоты попаданий этих точек в k-мерный шар радиуса 1 с центром в начале координат (используйте цикл for, функцию runif, команду if (условие) {…})

Чтобы набрать код программы, щёлкните по кнопке с белым крестиком внутри зелёного кружка, находящейся в левом верхнем углу экрана и выберите в меню пункт R Script Ctrl+Shift+N. В появившемся окне введите код программы. Для выполнения программы выделите весь код и нажмите кнопку Run с зелёной стрелкой, находящуюся над окном с кодом

2) Задайте n = 1000 и k = 10

3) Запустите код несколько раз, обращая внимание на значение счётчика количества попаданий в шар

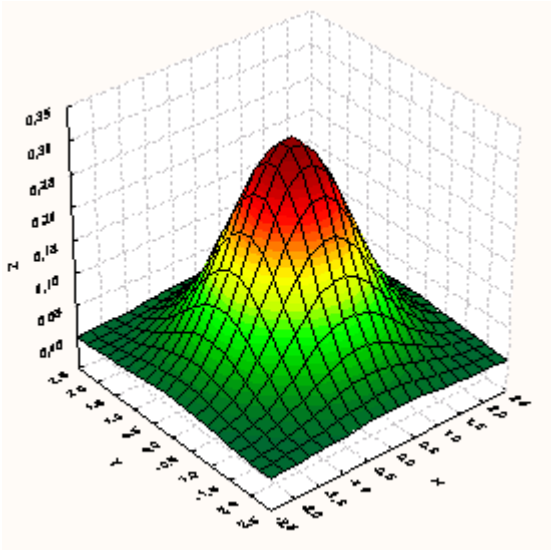
---

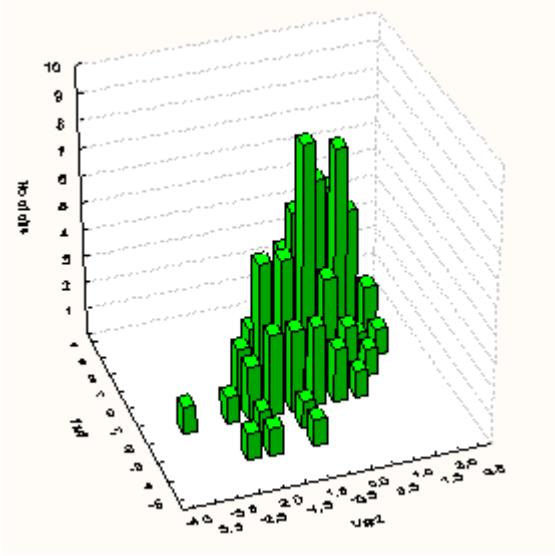
4) Вычислите приближённо вероятность, что количество точек, попавших в шар, окажется больше 8 (используйте формулу $p_k = \pi^i 2^{-k} / i!$, где $k = 2i$, pi для числа π, символ «^» для возведения в степень и функцию factorial)

5) Вычислите приближённо эту вероятность на основе теоремы Пуассона

6) Повторите пункты 3-5 для $n = 10^6$ и k = 16

Проблема проверки многомерной нормальности распределения

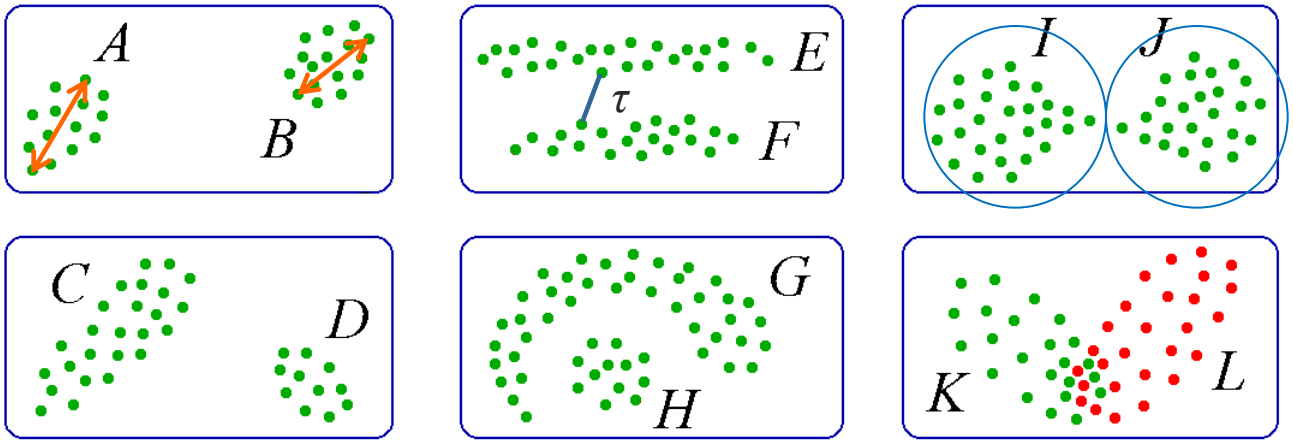




Допустим, что мы хотим проверить с помощью критерия хи-квадрат гипотезу согласия данных с некоторым многомерным законом распределения (например, k-мерным нормальным законом).

Предположим для простоты, что плотность распределения сосредоточена внутри k-мерного единичного куба. Возьмем $h = 0{,}2$ в качестве длины ребра ячейки группировки. Желание иметь в каждой из ячеек не менее 5 наблюдений влечет неравенство $nh^k \geqslant 5$. Из него при $k = 2$ находим, что $n \geqslant 5 \times 5^2 = 125$. Но при $k = 10$ получаем, что $n \geqslant 5 \times 5^{10} \approx$ 50 миллионов! На практике количество наблюдений, имеющихся у исследователя, редко достигает нескольких тысяч.

Различные типы кластеров



- **Кластер типа ядра:** ***A, B*** («чужие» — дальше диаметра)
- **Сгущение в среднем:** ***C, D*** (среднее межточечное расстояние)
- **Кластер типа ленты:** ***E, F, G, H*** («прыжок» длины τ)
- **Кластер с центром:** ***I, J*** (внутри шара заданного радиуса)
- **Накладывающиеся разноцветные «облака» точек:** ***K, L***

(такие множества точек будут рассматриваться позже при обсуждении темы «Классификация с обучением»)

Сгущённость и изолированность

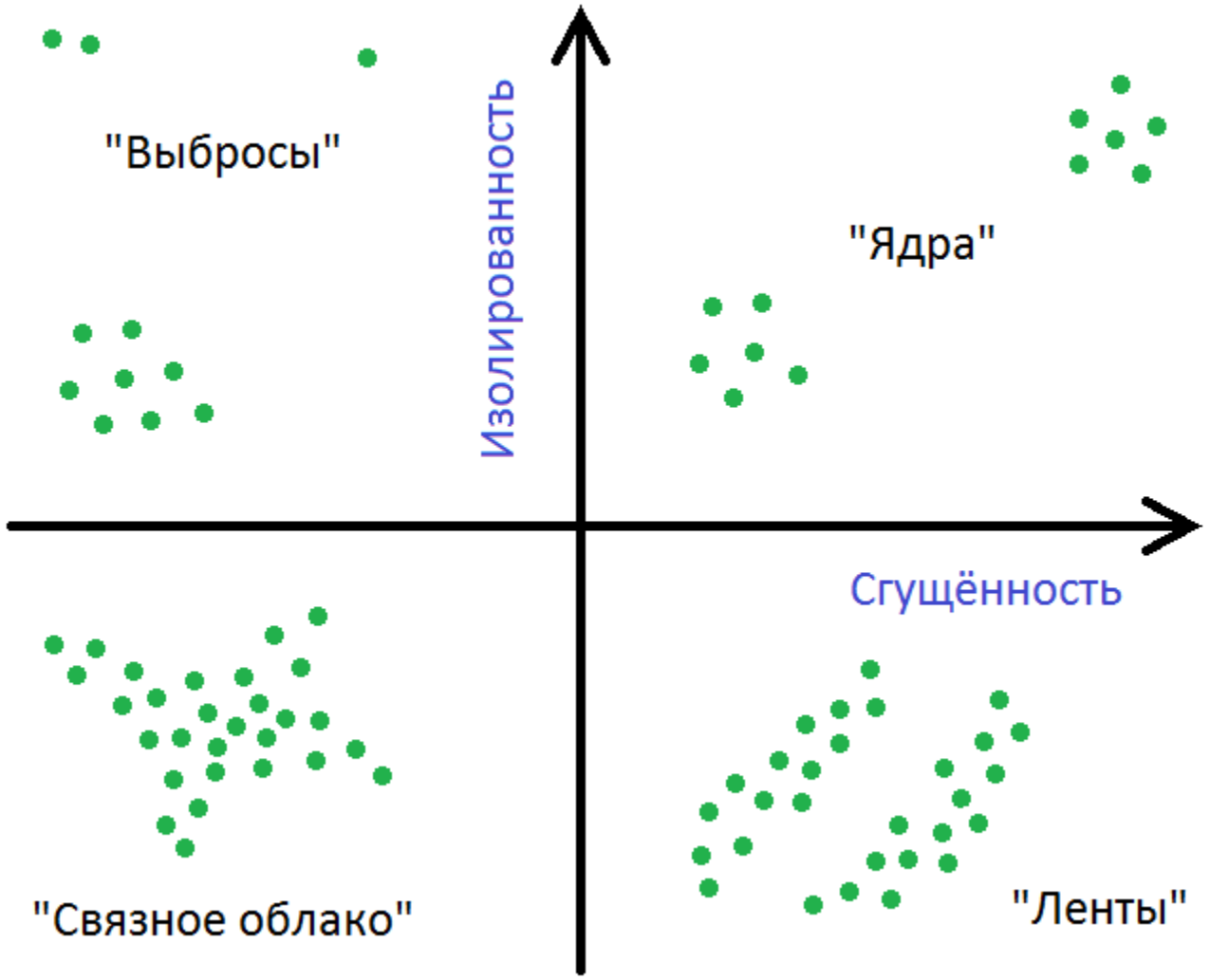
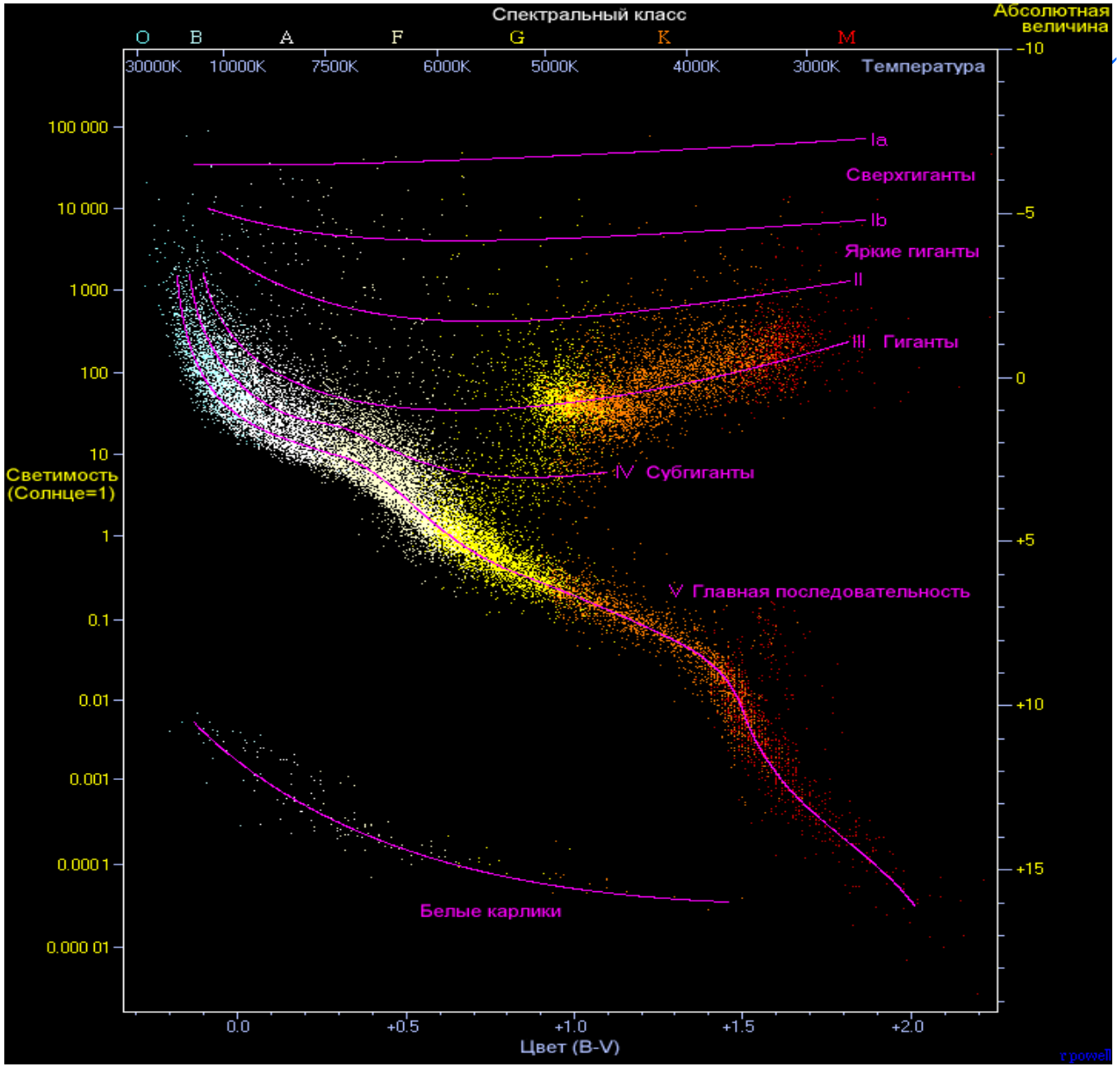
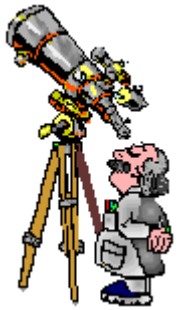


Диаграмма «спектр – светимость»





Ядерная оценка плотности (Розенблатта – Парзена)

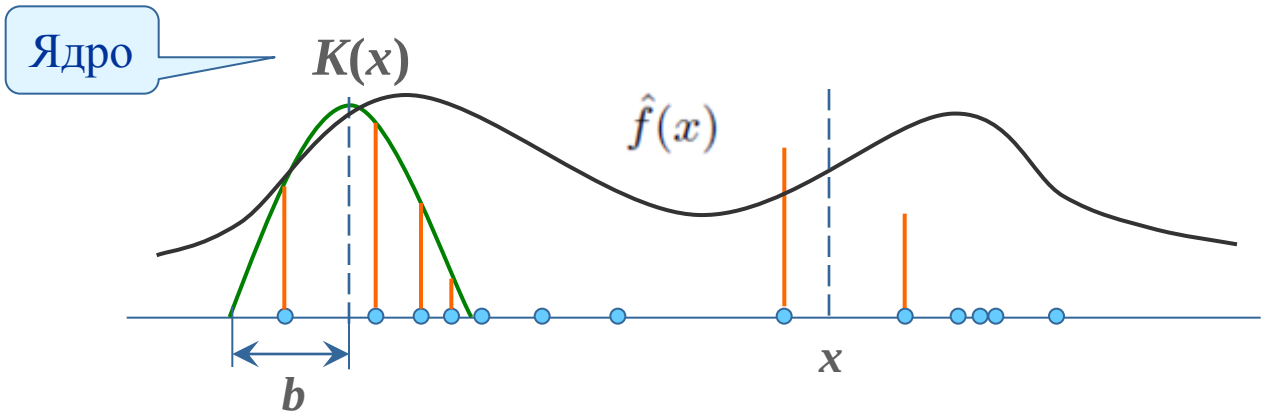
This idea of an *average shifted histogram* or *ASH* density estimate is a useful motivation and is discussed in detail in Scott (1992). However, the commonest from of density estimation is a *kernel density estimate* of the form

$$\hat{f}(x) = \frac{1}{nb} \sum_{j=1}^{n} K\left(\frac{x - x_j}{b}\right)$$

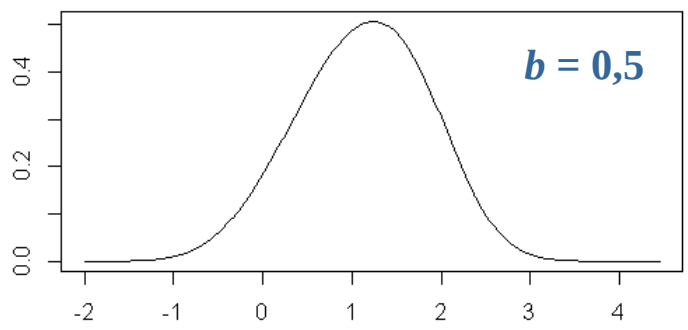
Ширина окна сглаживания

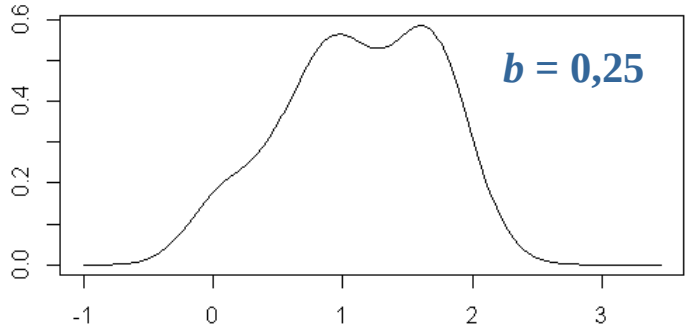
for a sample $x_1, \ldots, x_n$, a fixed kernel $K()$ and a bandwidth b; the kernel is normally chosen to be a probability density function.

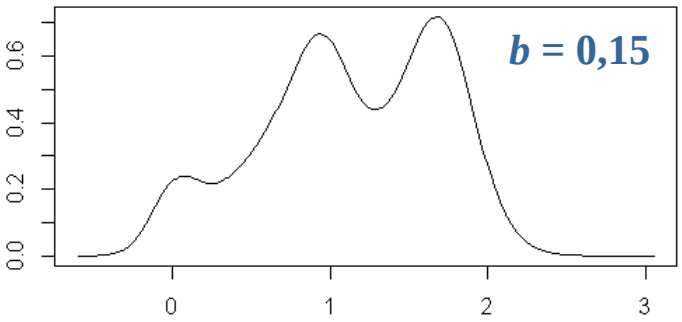
S-PLUS has a function `density`. The default kernel is the normal (argument `window="g"` for Gaussian), with alternatives `"rectangular"`, `"triangular"`.

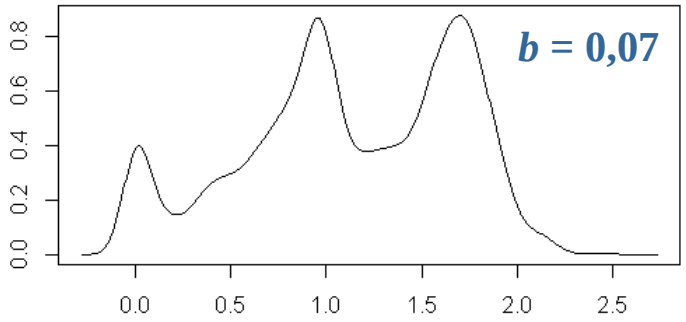


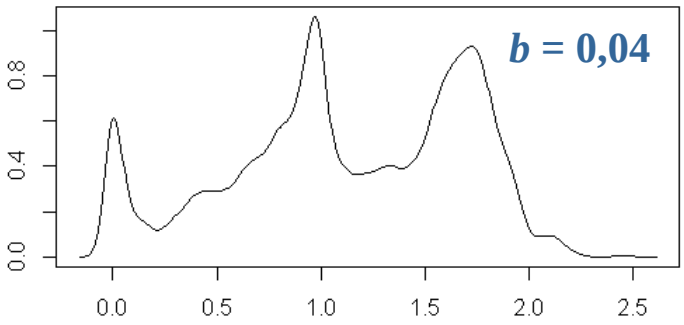
Влияние ширины окна сглаживания на ядерную оценку

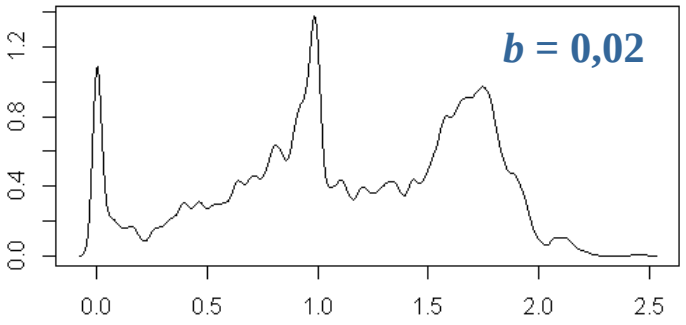




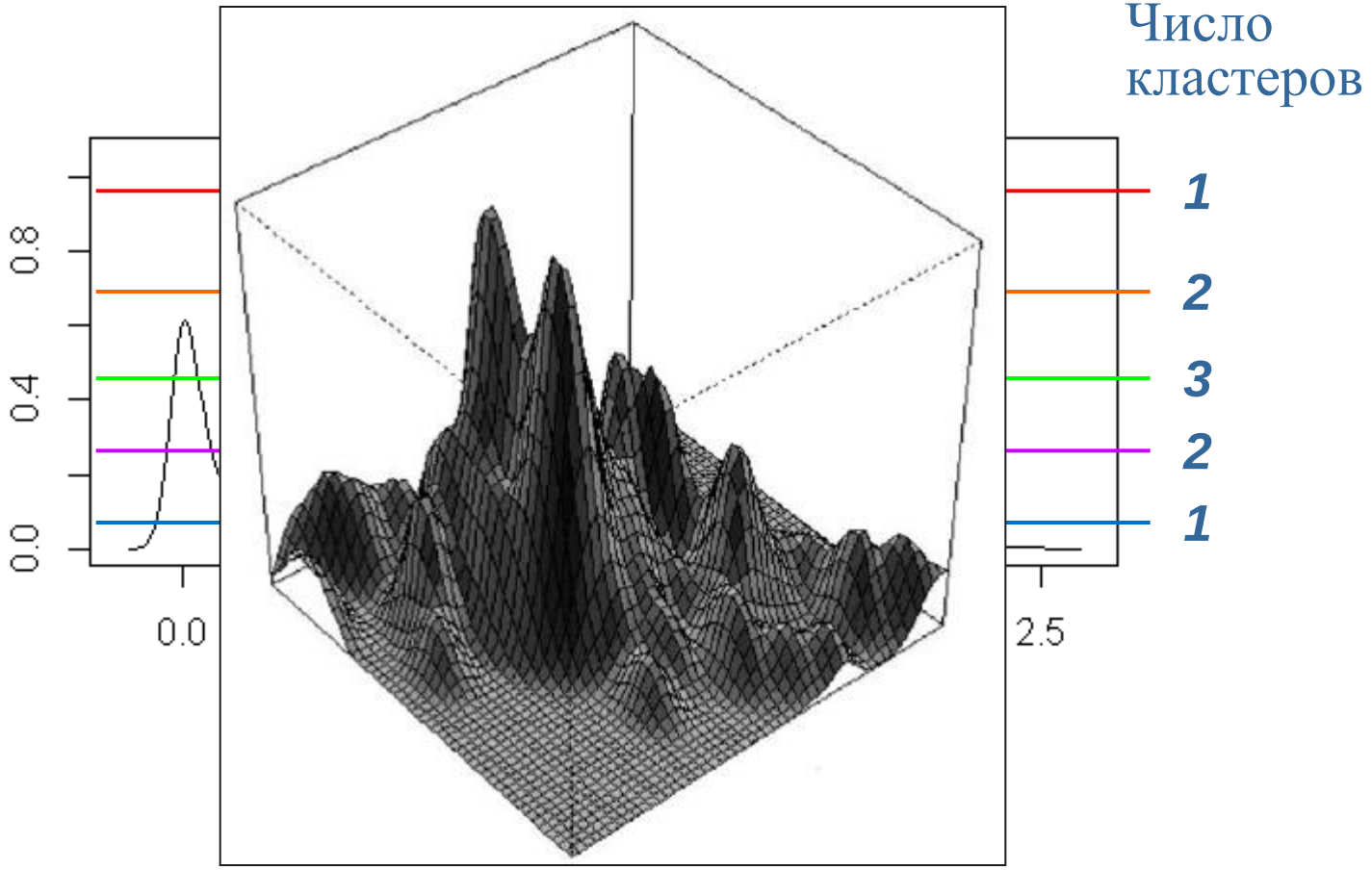






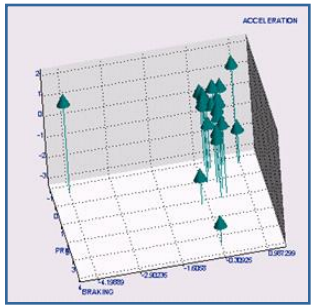


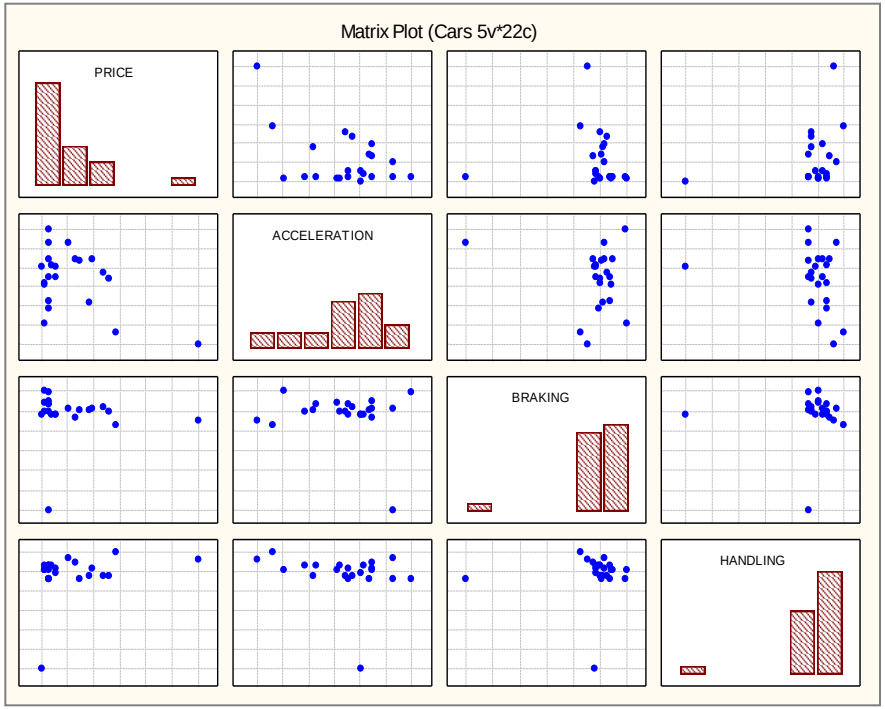
Зависимость числа кластеров от выбора уровня сечения ядерной оценки плотности

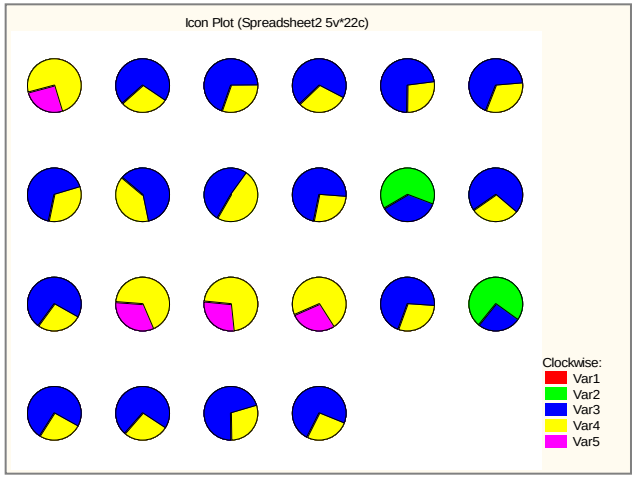


Разведочный анализ данных (Data Mining)

- **Matrix Plot — матричная диаграмма рассеяния**
- **3D XYZ Graph — трёхмерная диаграмма рассеяния**
- **Icon Plot — диаграмма профилей**







Типы нормировки

N1) $Z_i' = (Z_i - Z_{(1)})/(Z_{(n)} - Z_{(1)})$, где $Z_{(1)} = \min\{Z_1, \ldots, Z_n\}$, $Z_{(n)} = \max\{Z_1, \ldots, Z_n\}$.
Заметим, что все величины Z_i' принадлежат отрезку $[0, 1]$.

N2) $Z_i' = (Z_i - \overline{Z})/S$, где $\overline{Z} = \frac{1}{n}\sum Z_i$, $S^2 = \frac{1}{n-1}\sum (Z_i - \overline{Z})^2$.
Это преобразование обычно называют *стандартизацией*.

N3) $Z_i' = (Z_i - MED)/MAD$, где

N3 наименее подвержена влиянию «выбросов»

***MED* — выборочная медиана,**

***MAD* — (нормированная) медиана абсолютных отклонений от *MED*:**

Median of Absolute Deviations

$$MAD = \frac{1}{\Phi^{-1}(3/4)}\, MED\{|Z_i - MED|,\ i = 1, \ldots, n\},$$

Этот множитель (≈ 1,483) обеспечивает для выборки из распределения $N(\mu, \sigma^2)$ сходимость оценки *MAD* к параметру σ

где $\Phi^{-1}(x)$ — функция, обратная к функции распределения закона $\mathcal{N}(0, 1)$.

Для выполнения нормировок **N1-N3** на языке R используйте функции min, max, mean, std, median, mad

Кластеризация автомобильных марок

В файле CarChar.txt содержатся данные о 5 характеристиках автомобилей разных марок:

1 – *примерная цена* (Price),

2 – *ускорение* (Acceler, число секунд на разгон до 60 миль/час),

3 – *тормозное расстояние* (Braking, при скорости 80 миль/час до остановки автомобиля),

4 – *индекс способности удерживать дорогу* (Handling),

5 – *расход топлива* (Mileage, количество миль на галлон).

Практическое задание 2

1) Импортируйте файл CarChar.txt (не перепутайте его с CarSales.txt) под именем d и подсчитайте число столбцов в таблице (функция ncol)

2) Несмотря на то, что признаки уже были нормированы с помощью стандартизации N2, примените к столбцам таблицы d нормировку N3, используя цикл, функции median и mad. Запишите результат в таблицу z

3) Выясните, как изменилось наибольшее значения признака PRICE

---

4) Найдите «выбросы» для всех нормированных признаков, построив boxplot для таблицы z без первого столбца, замените их на пропуски NA Какие марки автомобилей выделяются и по каким признакам?

5) Постройте для таблицы z без первого столбца матричную диаграмму рассеяния (plot) (нажмите кнопку Zoom). Есть ли двумерные «выбросы»?

6) Постройте отдельную диаграмму рассеяния для признаков ACCELER и BRAKING, затем выполните команду text, задав аргумент labels=z[,1]

---

7) Примените к столбцам таблицы d нормировку N1, исключая пропуски при вычислении min и max (используйте функцию na.omit)

8) Постройте для марки «Buick» (строка 4) столбиковую диаграмму (barplot) и выясните, по каким именно признакам эта марка отличается от других (уберите названия марок, транспонируйте матрицу командой t)

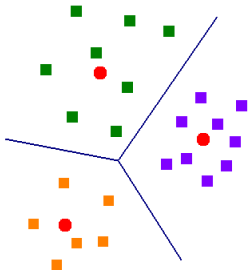
Эвристические алгоритмы

Подавляющая часть классификаций на практике проводится именно эвристические методами. Это объясняется:

1) относительной простотой и содержательной ясностью таких алгоритмов;
2) возможностью вмешательства в их работу путем изменения одного или нескольких параметров, смысл которых обычно понятен;
3) невысокой трудоемкостью алгоритмов.

На практике наиболее часто используются метод «*K*-средних» и иерархические процедуры.

Метод «*K*-средних»



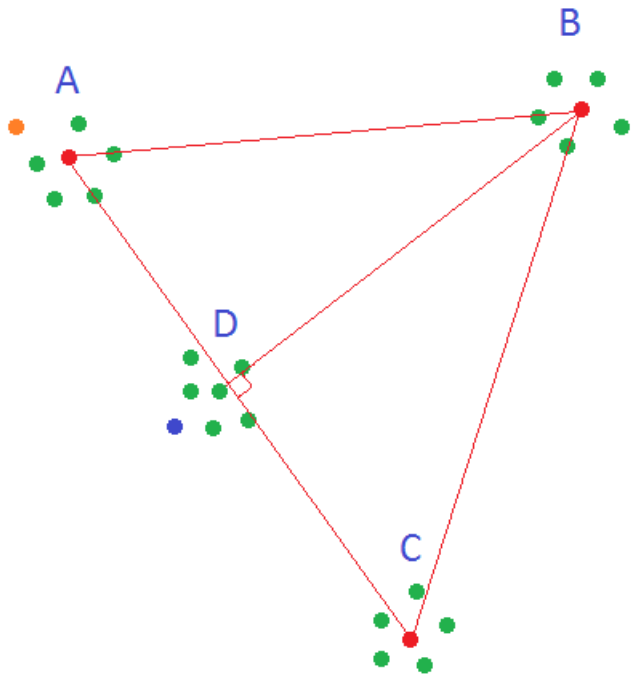
Число классов k задаётся исследователем

Метод “*K*-средних” предназначен для выделения классов типа “класс с центром”. Это название, ставшее популярным, введено Дж. Мак-Кином. Опишем один из вариантов метода.

Случайно выбираются k объектов — так называемых *эталонов*. Затем каждый объект присоединяется к ближайшему эталону, тем самым образуются k классов. В качестве новых эталонов принимаются центры масс классов (считается, что каждому объекту приписана масса 1). Другими словами, координаты новых эталонов определяются как средние арифметические координат объектов, входящих в соответствующий класс. После пересчета объекты снова распределяются по ближайшим эталонам и т. д. Критерием окончания алгоритма служит стабилизация центров масс всех классов.

Вместо случайно выбираемых эталонов, как правило, лучше использовать k наиболее удаленных объектов: сначала отыскиваются два самых удаленных друг от друга объекта, затем следующий эталон определяется как наиболее удаленный в среднем от уже имеющихся.

Наиболее удалённые начальные «эталоны» — не всегда оптимальный вариант выбора «эталонов»



Наиболее удалённой от трёх уже выбранных эталонов (красных точек) является не синяя точка из класса D, а оранжевая точка из класса A.

Медиана Кемени

Для одномерной выборки легко убедиться, что минимум суммы расстояний до элементов выборки достигается на выборочной медиане *MED*. Это свойство можно обобщить на разбиения, если некоторым образом определить расстояние между ними.

Разбиение $\mathbf{A}$ можно описать матрицей из 0 и 1: $a(i, j) = 1$, если и только если объекты i и j входят в один класс. В 1959 году американский статистик Джон Кемени предложил использовать следующий способ вычисления расстояния между разбиениями:

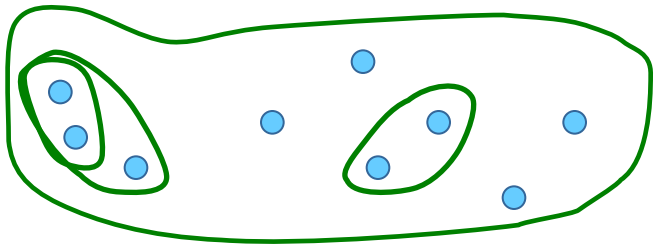
$$d(\mathbf{A},\mathbf{B}) = \sum_{i<j} |a(i,j) - b(i,j)|$$

Иначе его можно описать как количество пар объектов входящих в один класс для разбиения $\mathbf{A}$ ($\mathbf{B}$) и входящих в разные классы для разбиения $\mathbf{B}$ ($\mathbf{A}$).

Алгоритм на языке R реализован в пакете **ConsRank**

Кемени установил, что данное расстояние характеризуется рядом аксиоматических свойств (см. Орлов А. И. «*Нечисловая статистика*», 2004, с. 144). Разбиение, на котором достигается минимум суммы расстояний до других разбиений, называется их *медианой Кемени*.

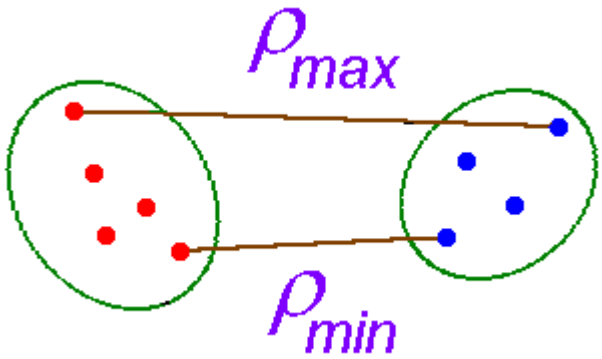
Иерархические процедуры

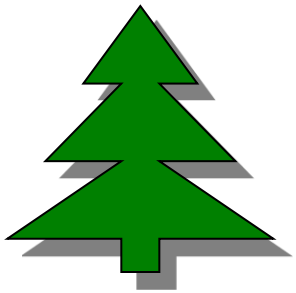


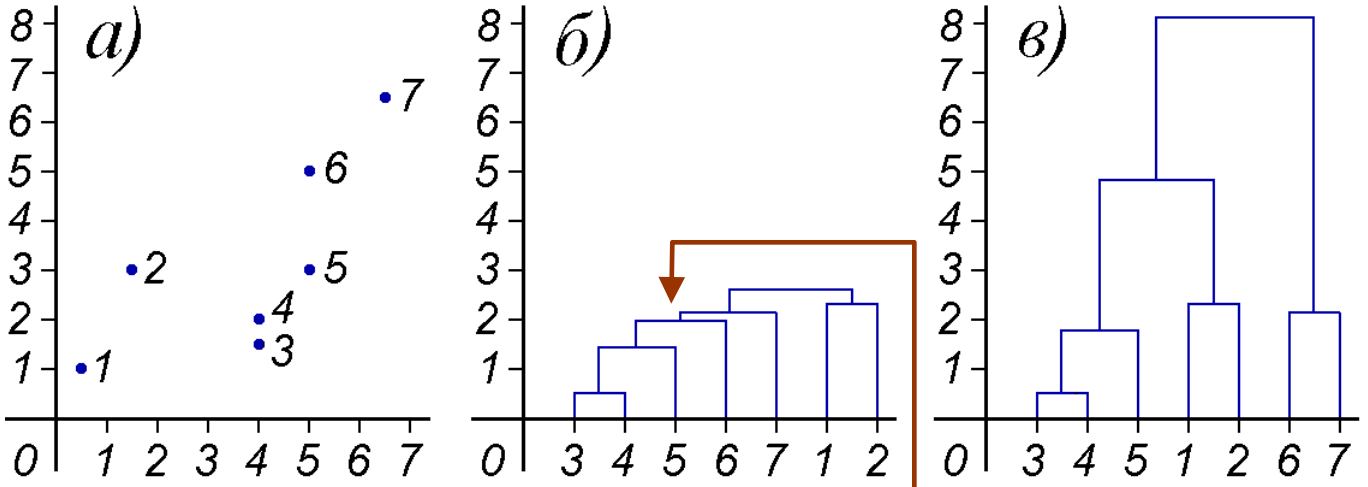
Опишем общую схему этих процедур.
Сначала каждый объект считается отдельным классом.
На первом шаге объединяются два ближайших объекта, тем самым создавая новый класс.
Для нового класса вычисляются *меры отдалённости* от него до всех остальных классов.
Шаги процедуры повторяются до тех пор, пока все объекты не объединятся в один класс.
Если прервать процедуру объединения за несколько шагов до полного объединения, то получим разбиение объектов на несколько классов. Наилучший момент прерывания можно определить визуально с помощью *дендрограммы* (от déndron (*греч.*) — дерево).

Примеры дендрограмм см. на следующем слайде

Процедуры «ближнего соседа» и «дальнего соседа». Дендрограммы.







Метод «ближнего соседа» имеет недостаток — так называемый «цепочечный эффект»

Типы иерархических процедур

P1) Метод «ближнего соседа»: $\rho_{min} = \min\limits_{\boldsymbol{x}_i \in S_k,\, \boldsymbol{x}_j \in S_l} d_{ij}$,

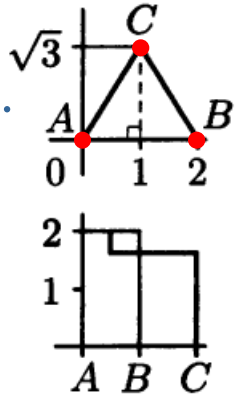
P2) Метод «дальнего соседа»: $\rho_{max} = \max\limits_{\boldsymbol{x}_i \in S_k,\, \boldsymbol{x}_j \in S_l} d_{ij}$.

P3) Метод средней связи: $\rho_{ave} = \dfrac{1}{n_k n_l} \sum\limits_{\boldsymbol{x}_i \in S_k} \sum\limits_{\boldsymbol{x}_j \in S_l} d_{ij}$

(здесь n_k и n_l — количества объектов в кластерах S_k и S_l).

P4) Метод центров масс: $\rho_{center} = |\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}_l|^2$,
где $\overline{\boldsymbol{x}}_k$ и $\overline{\boldsymbol{x}}_l$ обозначают центры масс k-го и l-го классов.

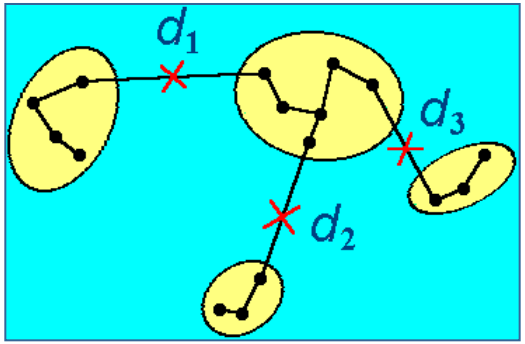
Недостатком этой процедуры является возможность появления *инверсий* — нарушений монотонности увеличения уровня при построении дендрограммы.

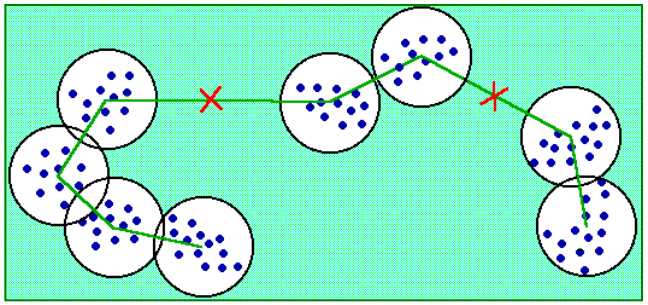


P5) Метод Уорда: $\rho_W = \dfrac{n_k n_l}{n_k + n_l} |\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}_l|^2$.

Для этой процедуры инверсии невозможны.

Связные компоненты и алгоритм «Форель»





Для процедуры «ближнего соседа» разрезание дендрограммы на некотором уровне равносильно отбрасыванию наиболее длинных рёбер в графе связности объектов. В результате чего граф распадается на связные компоненты.

Предварительная очистка от «шумящих» объектов обычно позволяет существенно улучшить классификацию, получаемую методом «ближнего соседа».

Для большого числа объектов применяется также алгоритм «Форель», в котором близкие объекты предварительно объединяются в «шаровидные» группы.

Уменьшение размера выборки

Для решения задачи классификации очень большого числа объектов (когда таблица содержит миллионы строк) на практике обычно применяется следующий подход:

1) Выбрать подмножество строк по некоторым условиям на признаки (например, в корпусе можно отбирать слова по частоте встречаемости, по части речи и т. п.)
2) Отобрать случайное подмножество строк (скажем, 10%)
3) Объединить похожие объекты в группы, применив первый этап алгоритма «Форель» (это часто позволяет уменьшить размер выборки в 10-100 раз)
4) Выделить объекты, входящие в области высокой сгущённости (плотности), т. е. выделить «ядра» классов (уменьшение размера выборки примерно вдвое)
5) Определить для «ядер» число классов с помощью иерархических процедур (ближний сосед, метод Уорда).

Главное в теме

- Перед кластеризацией строк таблицу данных, как правило, стандартизуют по столбцам (для очищенных от «выбросов» признаков применяют нормировку N1, для других — N3)
- Полезно визуально изучить данные с помощью построения диаграмм рассеяния всех пар признаков
- Перед кластеризацией иногда бывает полезно удалить «шумящие» объекты, т. е. находящиеся в «хвостовых» областях и в областях «разреженности» объектов
- Если число классов неизвестно, то обычно используют иерархические процедуры: методы «дальнего соседа» (complete) и метод Уорда (ward.D2) хорошо выделяют кластеры типа «ядра» или «сгущения в среднем», а метод «ближнего соседа» (single) — кластеры «типа ленты»
- Если число кластеров известно (нередко оно меньше 10), то применяют метод *K*-средних, который ориентирован на поиск кластеров типа «кластер с центром». Для таких кластеров важным показателем качества кластеризации служит большая доля межклассовой инерции
- Модификации метода K-средних (pam) и иерархических процедур (agnes), а также другие методы кластеризации реализованы в пакете cluster, который сопровождает книгу Kaufman L., Rousseeuw P. *"Finding Groups in Data"*

**Научное исследование — это искусство,
а правила в искусстве, если они слишком жёстки,
приносят больше вреда, чем пользы.**

Дж. Томсон, «Дух науки»

Домашнее задание

1) Импортируйте файла CarSales.txt поместите столбцы Price и Engine_s в таблицу данных d, вычислите число строк n в таблице d функцией nrow

2) Стандартизуйте столбцы d с помощью устойчивой нормировки **N3**

3) Напишите программу на языке R для выявления изолированных ("шумовых«) объектов, которые имеют сумму расстояний до ближайших к ним k объектов больше, чем выборочная (1 – α)-квантиль в векторе x, содержащем такие суммы расстояний для каждого объекта (строки d) (Используйте вложенные циклы, команду if, функцию sort для сортировки вектора t, содержащего k текущих ближайших квадратов расстояний, командой rep инициируйте компоненты вектора t символом Inf, обозначающим машинную бесконечность. *Идея алгоритма*: если квадрат расстояния до очередной точки меньше, чем t[k], то максимум заменяется на этот квадрат расстояния и вектор сортируется.)

4) Положите k = 10, α = 0,1 и запустите программу

5) Вычислите границу b, отделяющую "шумовые" объекты от остальных, затем командой plot постройте диаграмму рассеяния для таблицы d, раскрасив красным цветом (col=2) точки, соответствующие изолированным («шумовым») объектам

Подсчитайте числа красных точек

6) Измените значение параметра α на 0,05 и повторите пункт 5