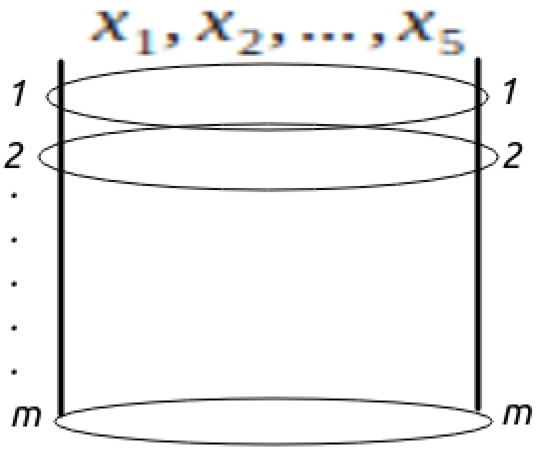
График параллельных координат

Пусть будет 5 признаков:

x_1, x_2, …., x_5

Каждый из них может состоять из скольки угодно элементов



Тут у нас m измерений, чему оно равно нас не волнует

1.Отнормируем каждый признак на интервал [0;1]

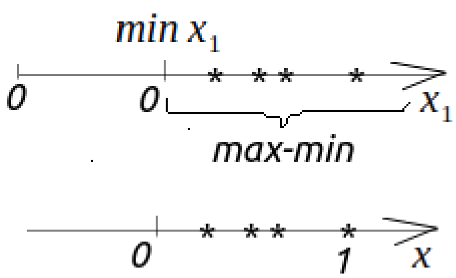
2.Отнимим минимальное значение от каждого и поделим на размах

3.Первое, что мы делаем, это отнимаем минимумы. Минимум становится равным 0, а вот максимум = макс-мин

4.Когда поделим на размах все эти точки (то на прямой появится единица=макс)

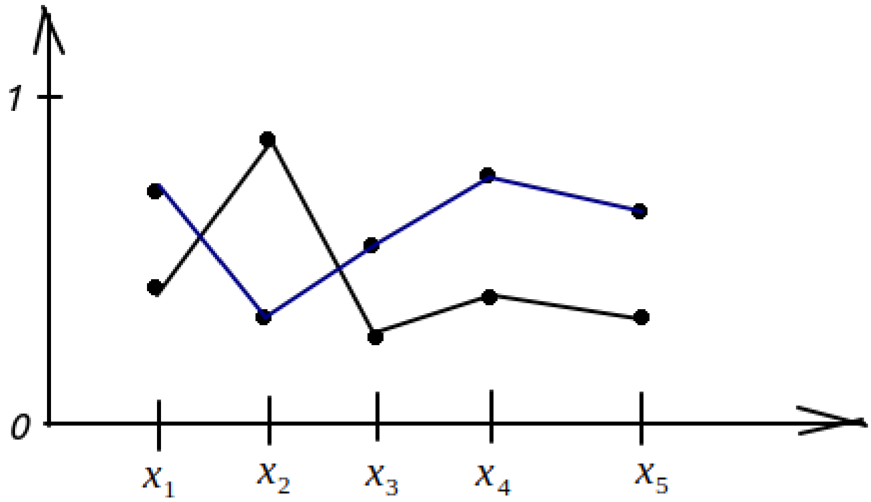
$$x_1' = \frac{x_1 - \min x_1}{\max x_1 - \min x_1}$$

$$x_1' \in [0,1]$$



Как строится график после нормировки?

Очень просто. По оси абсцис - номер или имя признака, по оси ординат - нормированные значения от 0 до 1



И вот мы берем первый объект и у него x*1=0.5, x*2=0.8, x*3=масенький, x*4 и X_5 где-то равны. Соединяем эти точки прямыми.

Переходим ко второму объекту. Снова наносим все точки и соединяем их.

Делаем так, пока все объекты мы не описали.

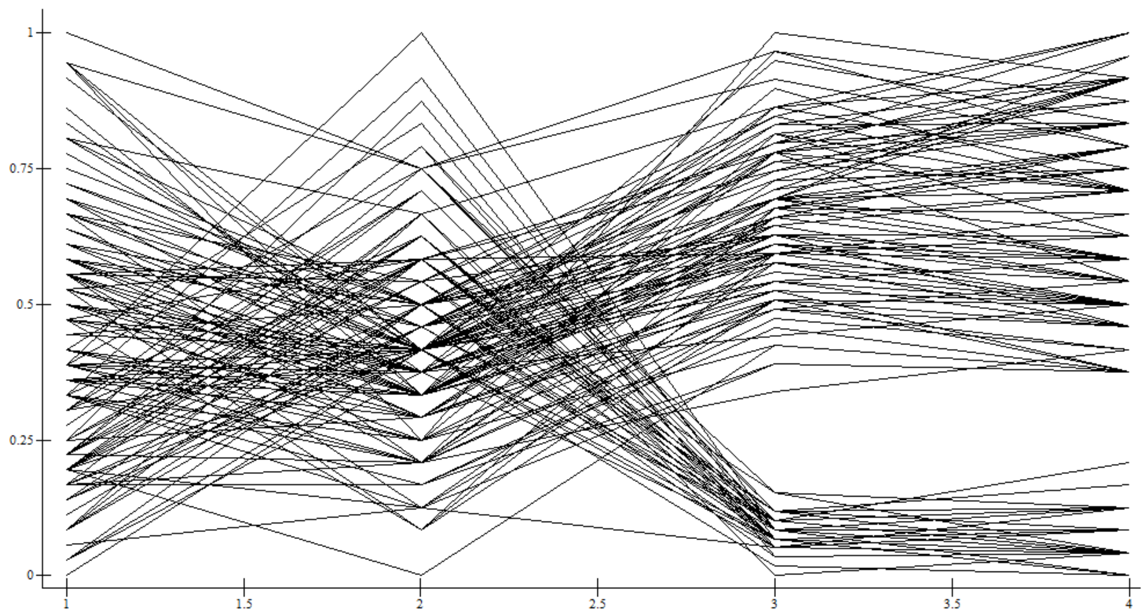
Реализация на APL:

Соединяем все ириски в матрицу ir и проверяем размерность. У нас 150 цветочков и 4 признака

```
      ⍴ir←ir1⍪ir2⍪ir3
150 4
```

Получаем график ломаных

```
      (150/1) parcoor ir
```



Что гляда на график можно сказать? Если наклонить головку вправо и представить как будет выглядеть гистограмма, она будет близка к равномерному распределению. От минимума до максимума мы видим, что все 150 цветочков перемешаны в кучу и нет никакой структуры соответственно. (Это по первому признаку)

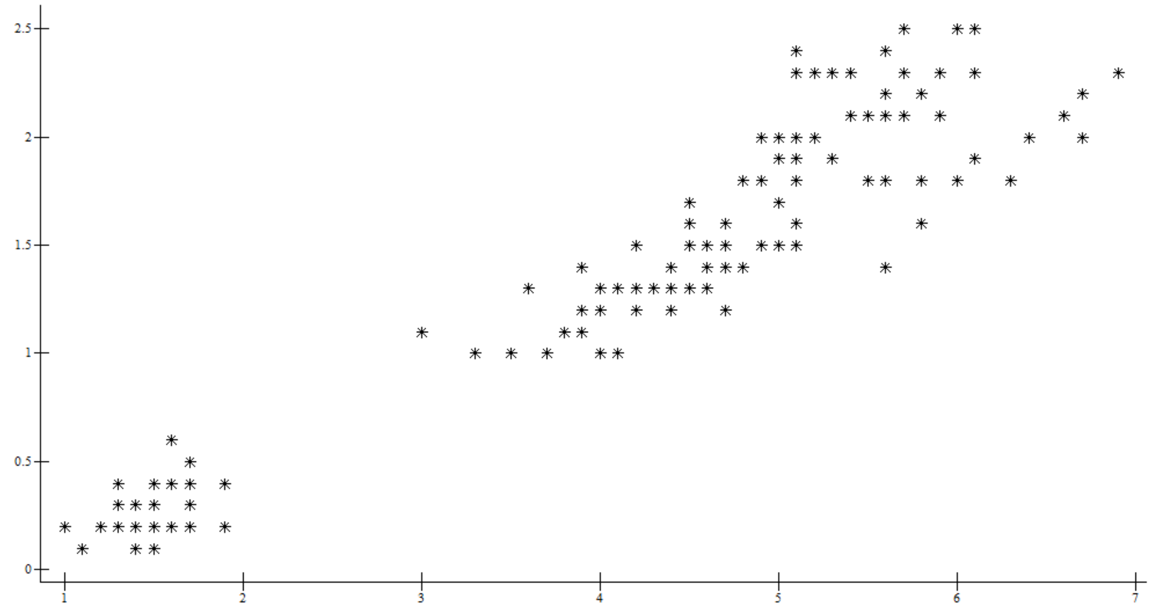
Глядя на второй признак, видим, что все цветки снова перемешаны.

А вот глядя на третий признак. ОПА!!! Видим две группы. Часть цветков имеют малое значение этого признака, а другая часть - большие значения этого признака.

Аналогично по признаку 4.

Выбираем интересные на наш взгляд признаки 3 и 4. Построим для них графичек.

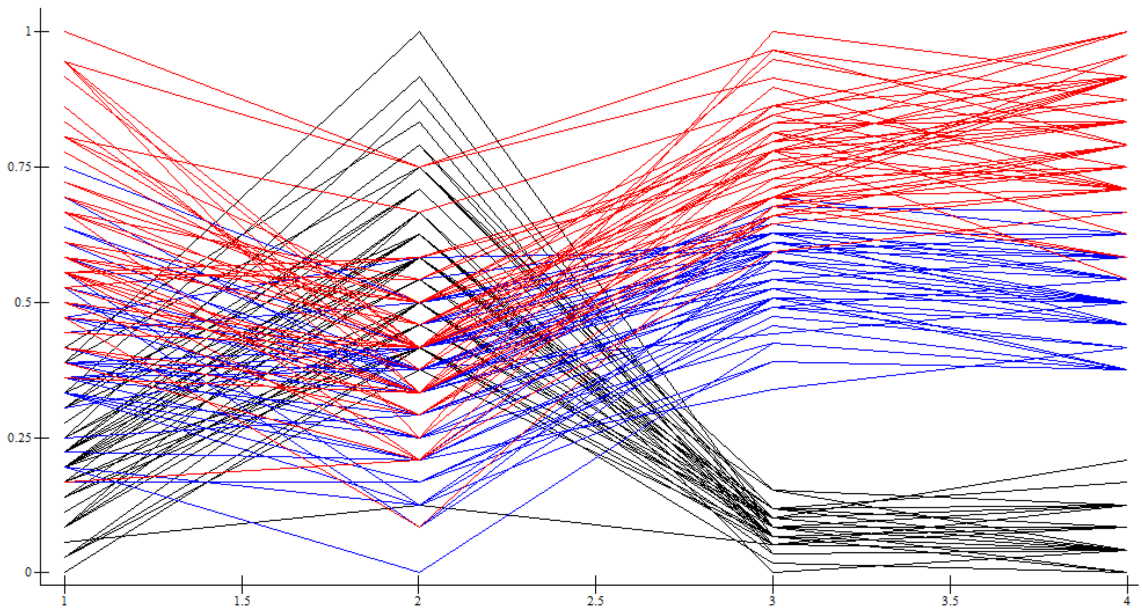
```
plot ir[;3 4]
```



Видем две группы, два кластера. Мы открыли то, чего не ожидали, что все цветки разделены на 2 группы. Мы не ожидали, что все цветки разобьются на две группы.

Теперь мы рассматриваем эти две кучки, Зовем Марью Ивановну и она нам говорит. Что одна группа принадлежит одному сорту, а во второй группе у нас смешаны два сорта. И Марья Ивановна нам их рисует. Но перед этим мы построим покрашенный график параллельных координат.

```
      (50/ι3) parcoor ir
```



По первому признаку черный, красный, синий сорты перемешаны к чертовой матери.

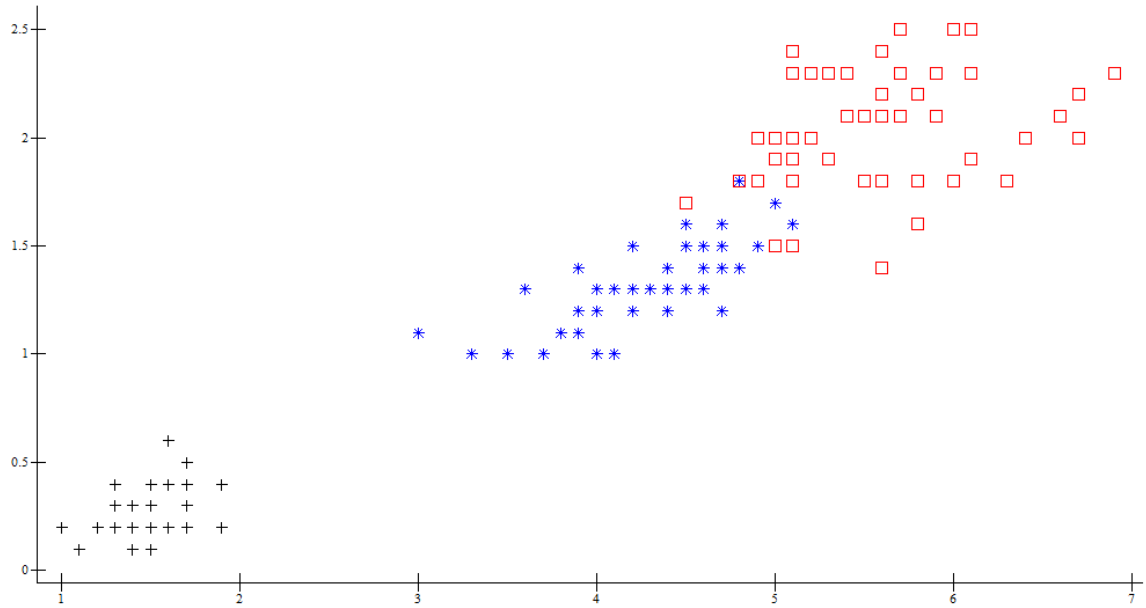
По второму признаку анлогично.

А по третьему и четвертому. Черненький (первый сорт) сильно отстоит от синего и красного сортов. Есть возможность поставить порог.

Мы видим, что не смотря на то, что чуть-чуть они перемешаны (4 красненьких попали к синеньким), т.е. мы не можем безошибочно разделить эти два класса, но можно сказать, что синенькие имеею меньшие значения признаков, чем красные.

Теперь можем постоить плоскость 3 и 4 признака и покрасить.

```
(50/⍳3)plotc ⊂[1]ir[;3 4]
```

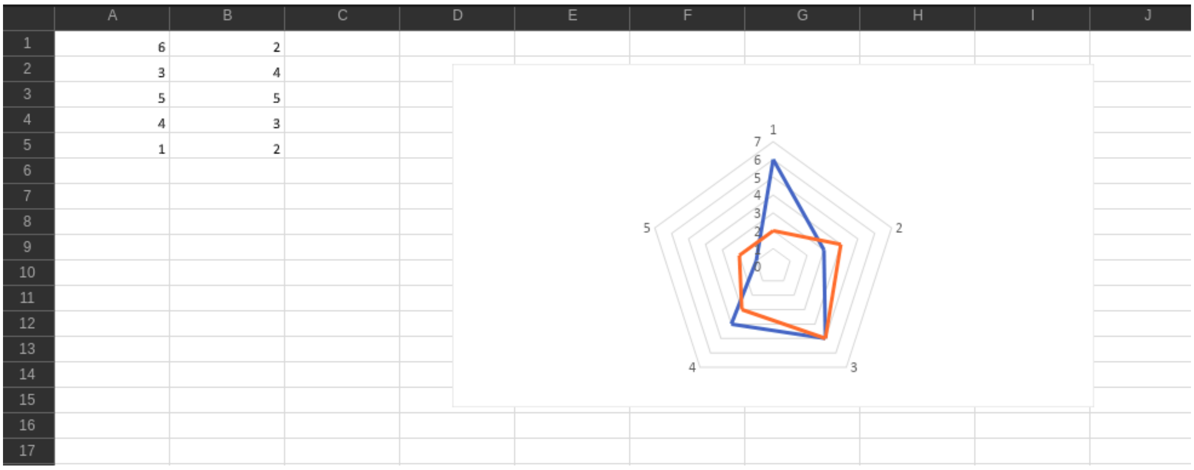


Этот образ отличается от первого, тем, что в принципе этому графику наплевать, сколько у нас признаков 4 или 400. Это очень мощный метод визуализации.

Еще один метод на котором мы подробно не будет останавливаться.

Radar plot (у него много названий)

Продемонстрируем в Экселе. Это тот же самый графичек, но который мы строим в полярных координатах.



	A	B
1	6	2
2	3	4
3	5	5
4	4	3
5	1	2

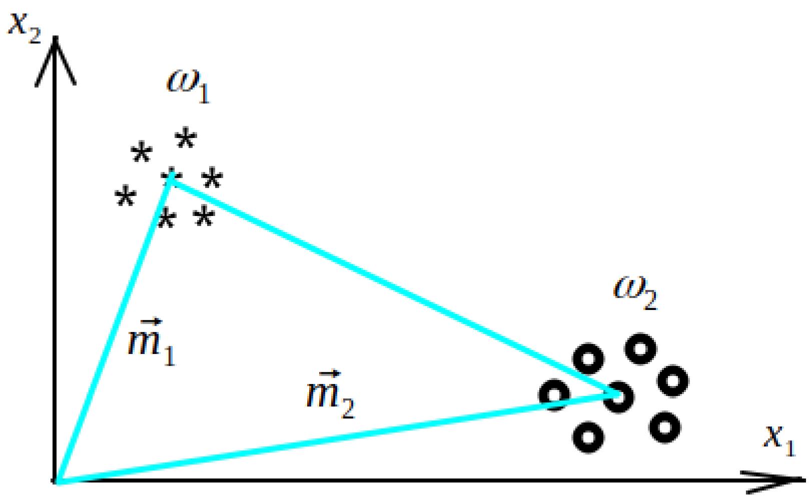
Видим, как у нас отличаются эти два столбца по своим значениям. И если их много, мы можем тоже выявить какие-то разные типы. Часть с острыми уголками, часть без острых уголков.

Вот это мы можем положить картинки и сказать: "Дети, разложите пожалуйста картинки на две кучки" и они прекрасно проведут это кластерный аналих.

Все эти графики позволяют выявить какие-то интересные признаки и понизить размерность пространства, т.е. из большого числа X выбрать небольшое и в небольшой размерности мы можем замечательно смотреть.

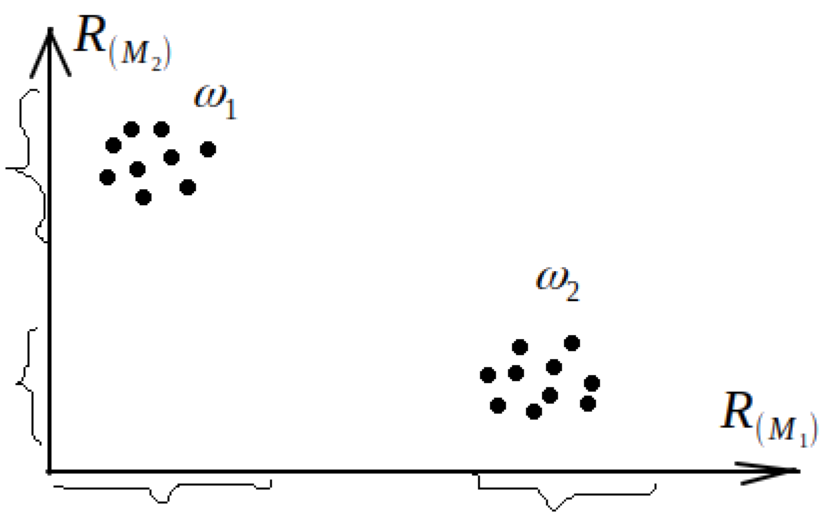
Теперь посмотрим простые два примера понижения размерности пространства:

1. ***Расстояние до двух фиксированных точек.*** это центры классов чаще всего фиксированных точек. Если известно, что есть два класса омега1 и 2, то две фиксированные точки. Так это будет выглядеть



M1-центр первого класса

M2-центр второго класса, строим следующий график



по оси абсцис расстояние до R1, по ординат - до R2

Сначала идут омега 1, потом омега два

У омега1 расстояние до M1 маленькое, а до M2 - большое. Для омега2 наоборот.

У нас изменются деления по осям, а вид графика незначительно изменится.

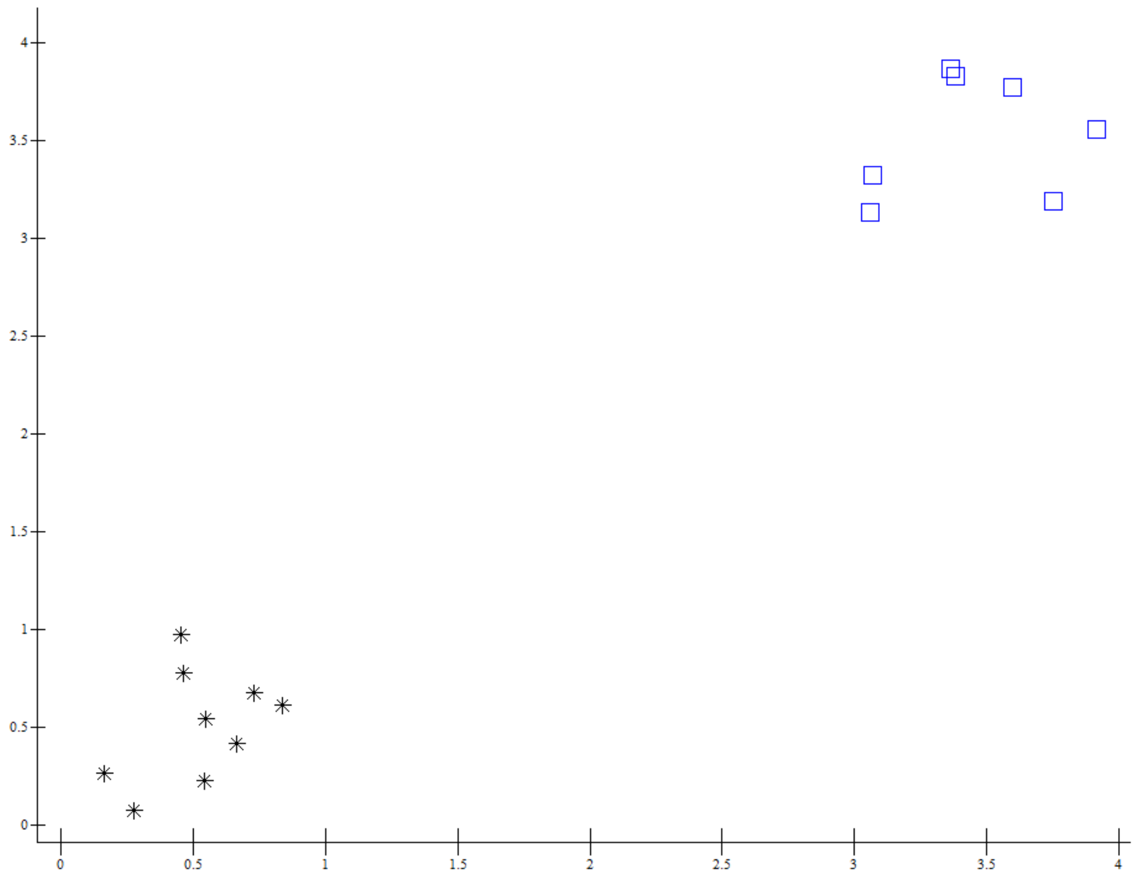
А зачем мы его строим? Его не надо было строить в двумерном пространстве потому что и так все видим, а вот в многомерном пространстве нам его надо построить.

Начнем с двумерного:

```
      a←?9 2⍴0     ⍝ A-первый класс(9 случайных точек от 0 до 1)
      b←3+?7 2⍴0   ⍝ B-второй класс(7 случайных точек от 0 до 1) и мы прибавим 3-
ку
```

Строим:

```
      plot (⊂[1]a)(⊂[1]b)
```



Если вычислим M1. Напишем ф-цию ave и она будет суммировать матрицу по первому измерению и делить на число строк.

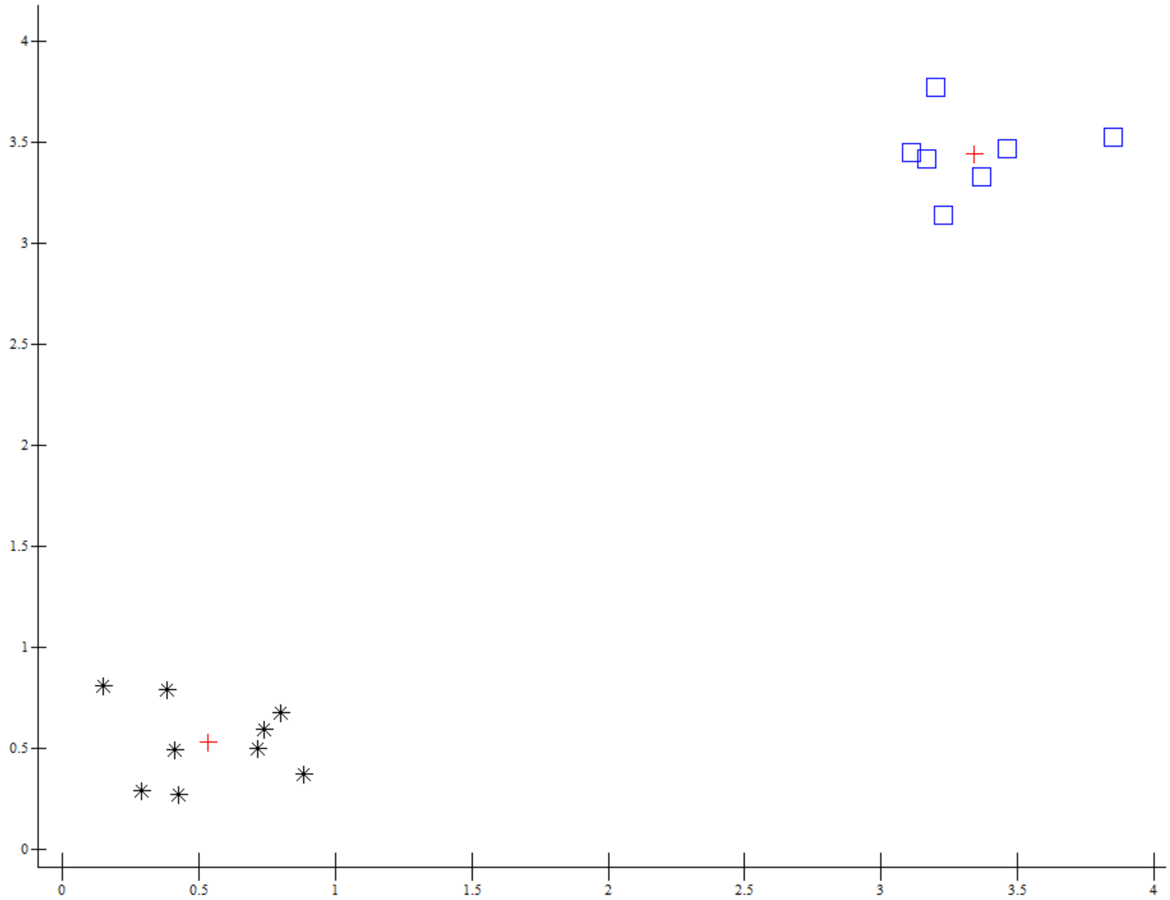
```
      ave←{(+⌿⍵)÷≢⍵}
```

Посчитали центры

```
      m1←ave a
      m2←ave b
```

Теперь мы можем их пометить на графике

```
      color 'red'
      marker m1  ⍝ центр первого класса
      marker m2  ⍝ центр второго класса
```

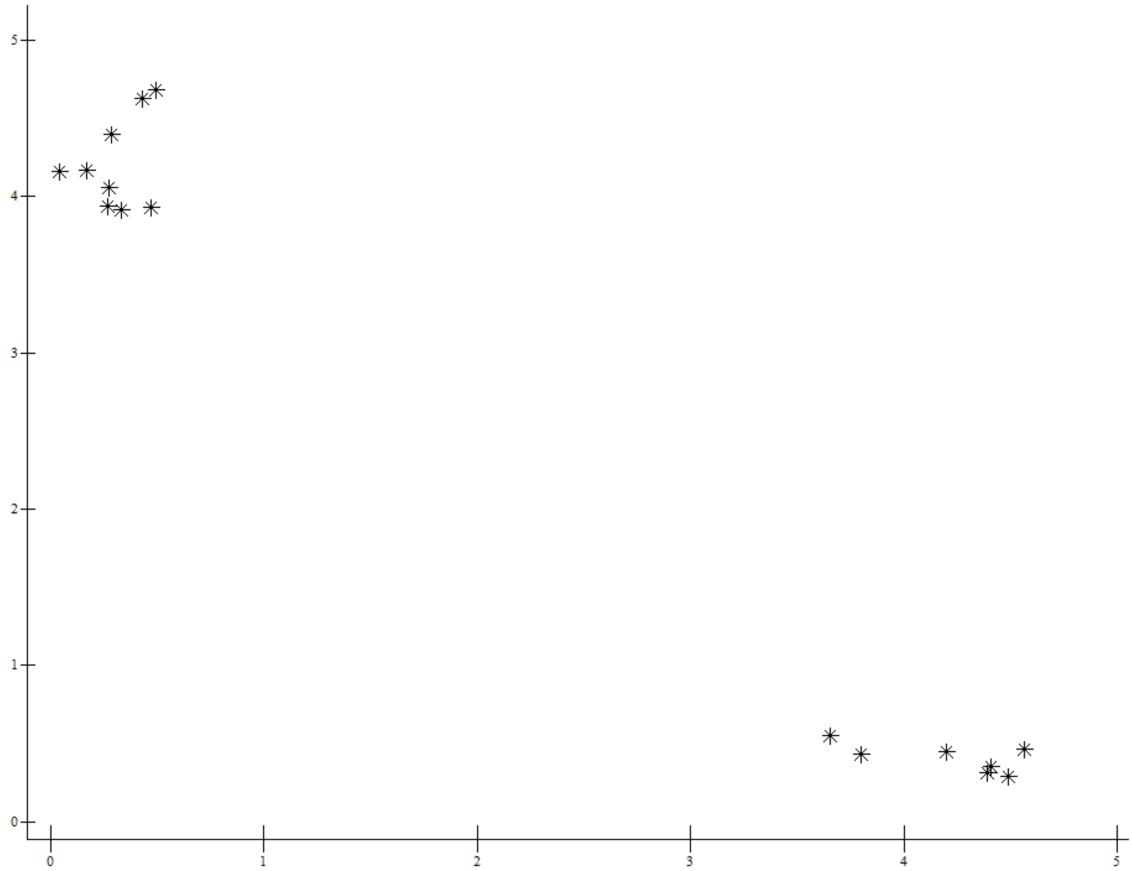


Теперь слепим a и b в одну матрицу, чтоб не мучиться и R1 - это растояние от M1 до всех точек. R2 аналогично.

```
ab←a⍪b
r1←(+/(ab-[2]m1)*2)*0.5
r2←(+/(ab-[2]m2)*2)*0.5
```

Построим

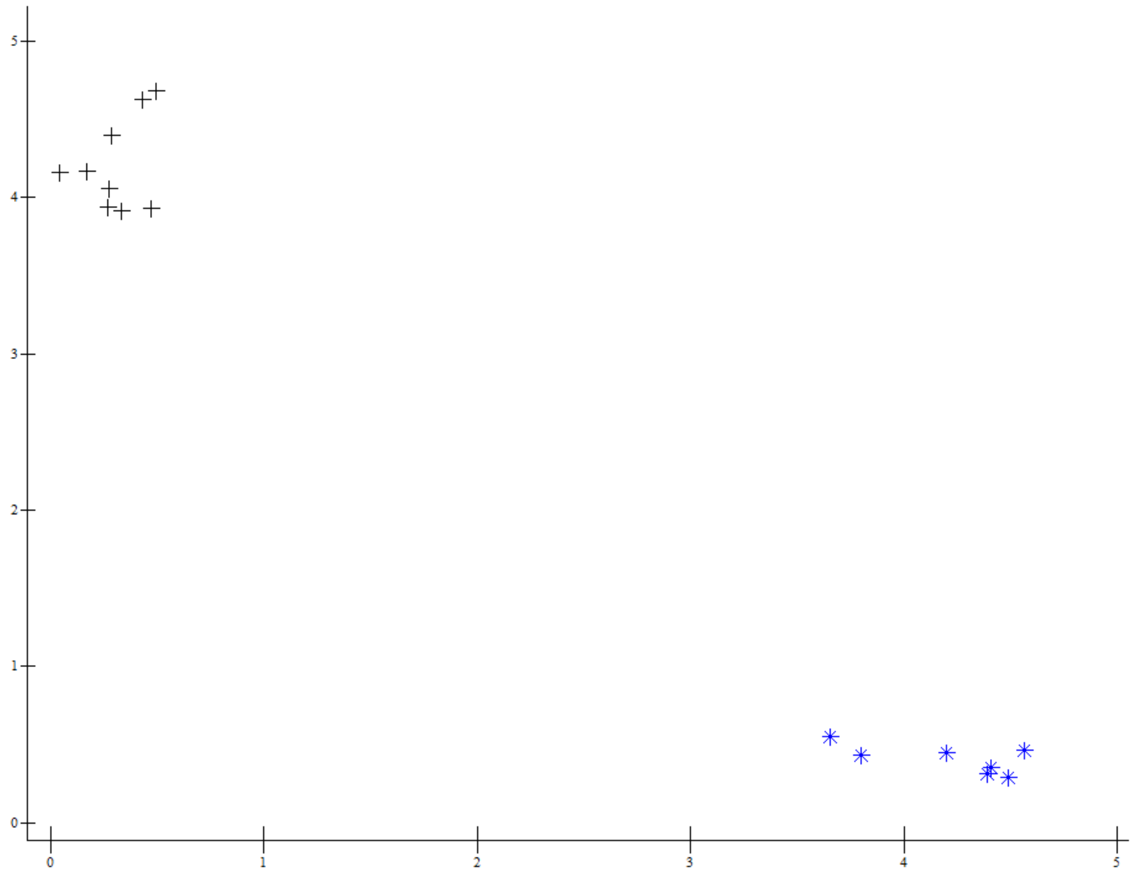
```
plot r1 r2
```



Распределение перевернулось, но ничего равным счетом не поменялось

Покрасим:

```
    (9 7/1 2)plotc r1 r2
```

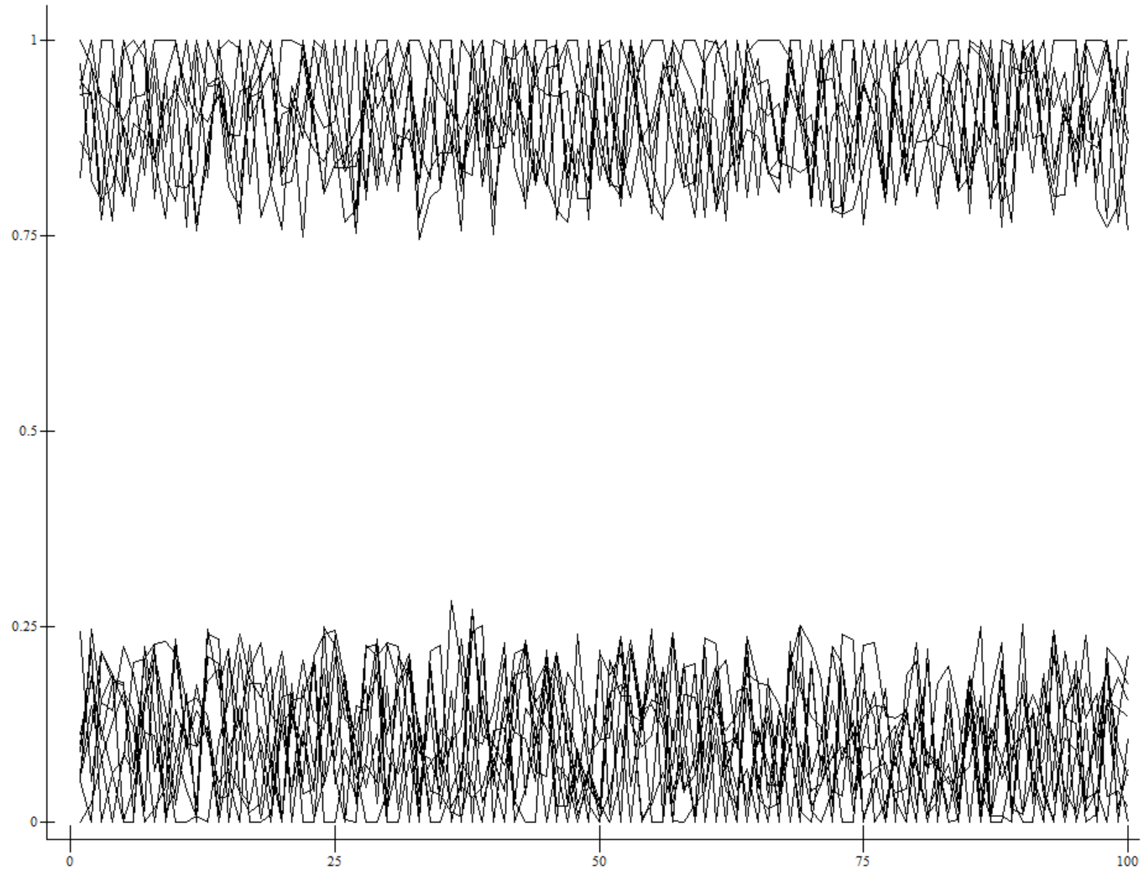


Когда у нас и так все было видно, то не имело смысла, что-то городить.
Тогда представим, что теперь у нас не двумерное, а ***стомерное пространство***!

```
      a←?9 100ρ0
      b←3+?7 100ρ0
      ρa
9 100
      ρb
7 100
```

Никакого графика мы здесь не нарисуем, можем сделать параллельные координаты. Всего у нас 16 точек.

```
      (16/1)parcoor a⍪b
```



Здесь трудно понять, что к чему относится.

Считаем опять M1 и M2 - центры:

```
      m1←ave a
      m2←ave b
      ρm1
100
      ρm2
100
```

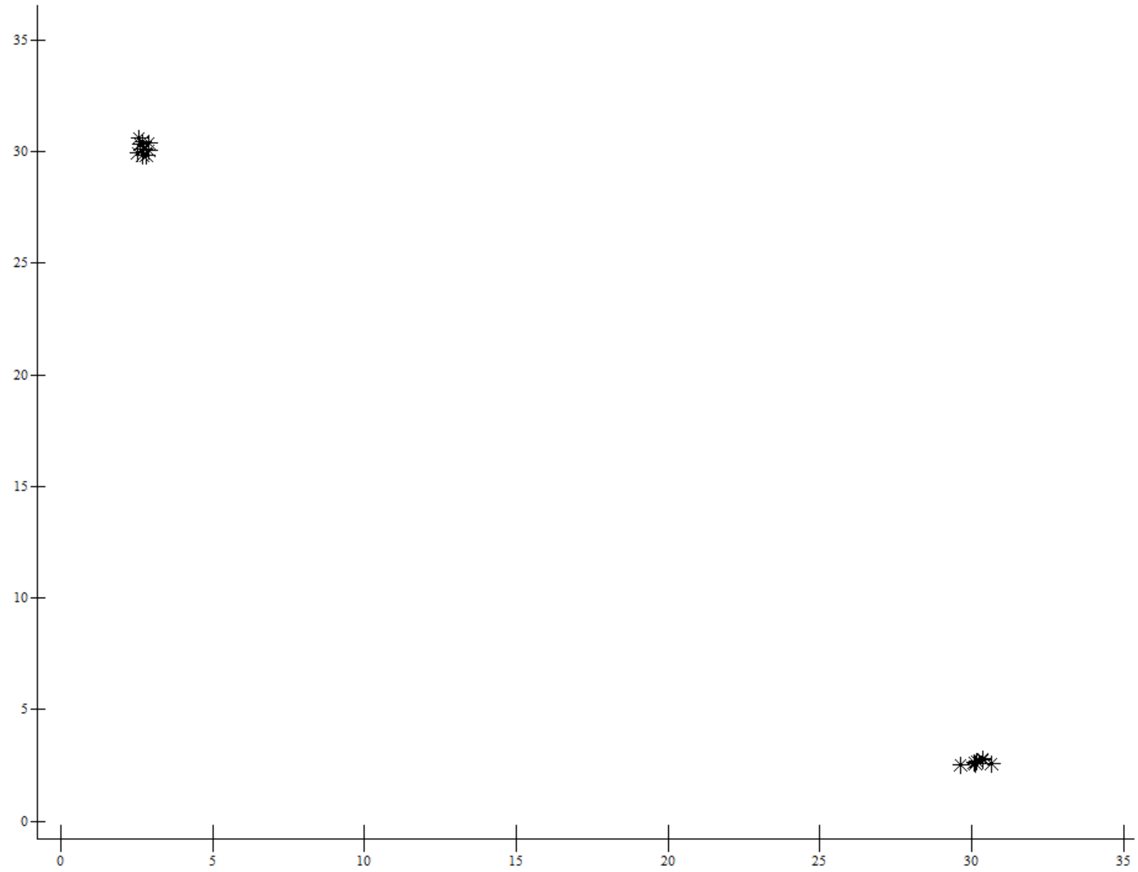
Они у нас теперь 100-мерные.

Делаем a и b. Считаем растояние первого центра до всех точек

```
      ab←a⍪b
      r1←(+/(ab-[2]m1)*2)*0.5
      r2←(+/(ab-[2]m2)*2)*0.5
```

И стоим графичек. Видим две группы на плоскости.

```
      plot r1 r2
```

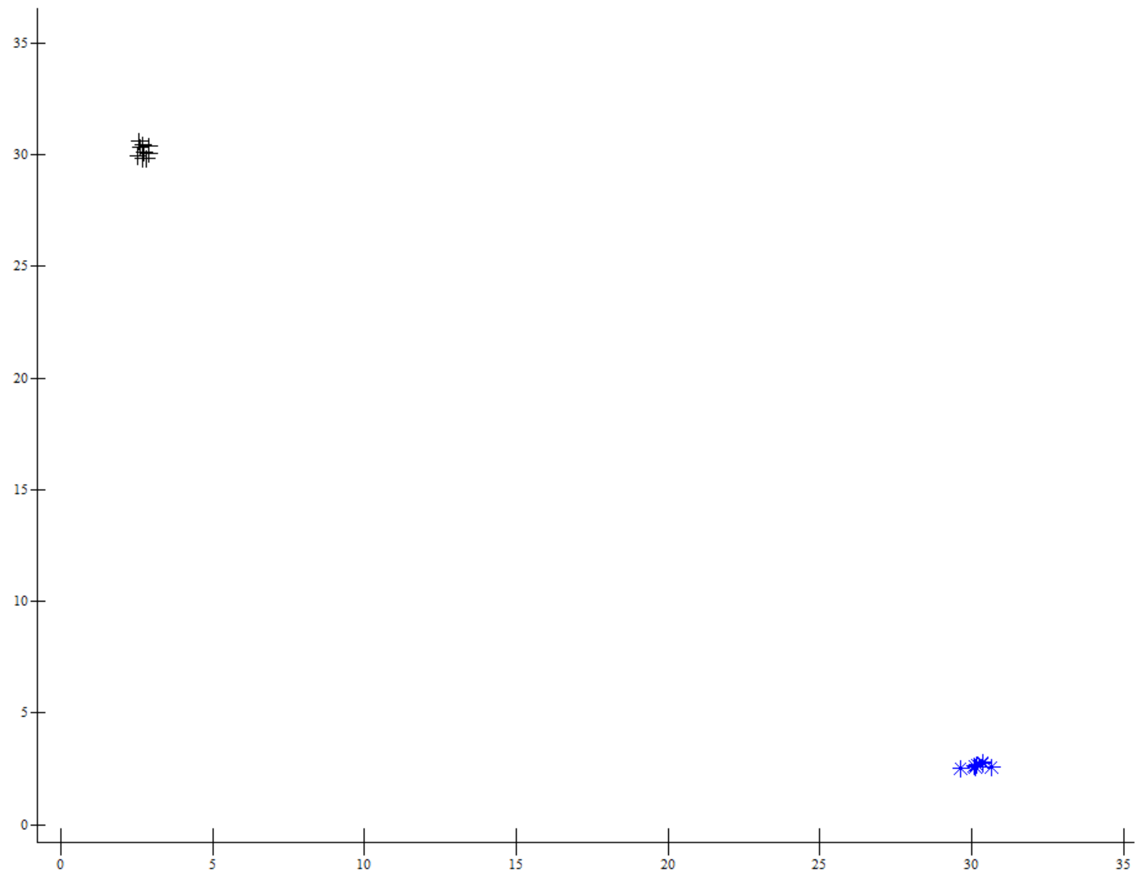


Можем взять 1 признак (кто меньше/больше 15)

```
      +/r1<15
9
      +/r1>15
7
```

Важно, что мы были в 100мерном пространстве и нифига не увидели, зная что первые точки относятся к одному классы, а другие ко второмы. Посчитали расстояния до центров от всех точек.

```
      (9 7/1 2)plotc r1 r2
```



Чтобы по 100 раз не набирать, сделаем ф-цию:

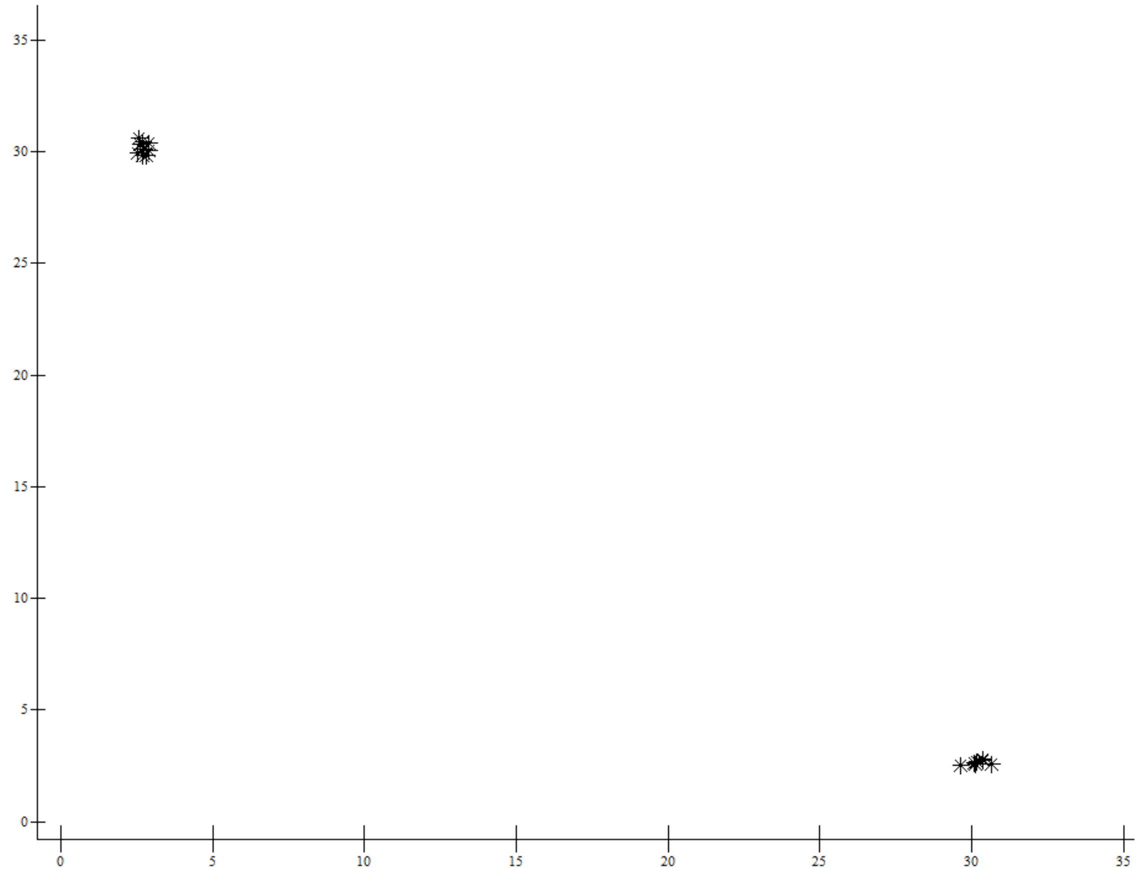
```
)ed m12proj

[0]     r←a m12proj b;ab;m1;m2;r1;r2  ⍝ аргументы матрицы a и b. Результат -
вектор                                           векторов расстояний.
[1]     m1←{(+⌿ω)÷≢ω}a                ⍝ считаем среднее по a
[2]     m2←{(+⌿ω)÷≢ω}b                ⍝ считаем среднее по b
[3]     ab←a⍪b                        ⍝  делаем матрицу a и b, где у нас слепятся
обе
[4]     r1←(+/(ab-[2]m1)*2)*0.5       ⍝  посчитаем расстояние
[5]     r2←(+/(ab-[2]m2)*2)*0.5
[6]     r←r1 r2                       ⍝ в результат передадим r1 и r2
     ∇
```

Зелененьки-переменные глобальные(они нужны только на момент исполнения ф-ции), когда записываем в [0] они становятся локальными и стали беленькими.

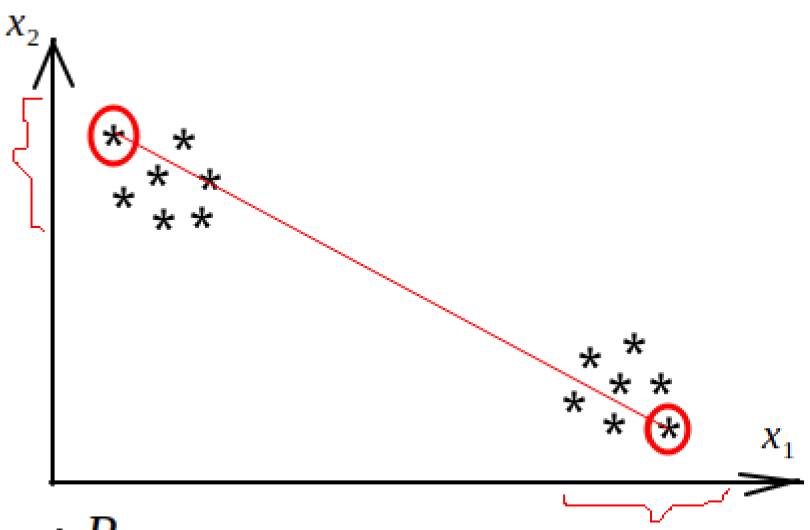
Постоим:

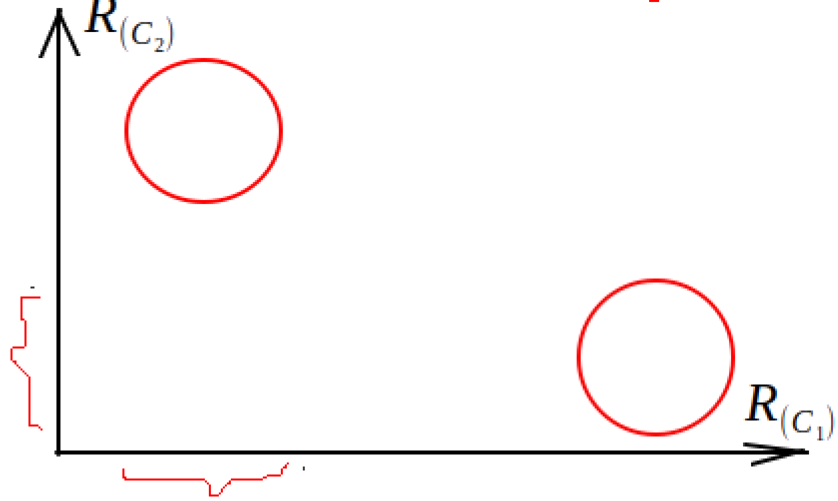
```
      plot a m12proj b
```



На выходе у нас будет вектор расстояний до первого класса и вектор расстояний до второго класса.

2. ***Если у нас тот случай, который очень часто встречается. Мы не знаем есть ли какие-то классы или их нет.***





Какие две точки кажутся особенными? (зелененьким обвели) Их особенность в том, что они наиболее удаленные друг от друга. Если посчитаем все возможные между всеми точками рассояния и найдем максимум, то он будет в этих двух точках.

Снова считаем расстояния. До С1 - маленькие, а до С2 - большие.

Нам наплевать в сколькимерном пространстве считать расстояния между точками

$$R_{xy} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Если n=2 - теорема Пифагора

Если n=3, берем ту же формулу, x и y - лежащий на столе и на полу шарик. Начало координат в углу комнаты. Один конец рулетки от шарика на столе, другой конец - шарик на полу и получим то же расстояние, что и по формуле.

Если n=4, то с рулеткой не залезем. И после 4-х нам уже наплевать.

Расстояние мы это посчитаем:

```
      ⍴ab
16 100
      dist←{(+/(⍺-⍵)*2)*.5}
```

есть точки

```
      1 2
1 2
      3 4
3 4
```

какое между ними расстояние? будет корень из 8(рисунок)

```
      8*.5
2.828427125
```

Имея dist ф-ции считаем расстояние между всеми a и b

Есть два вектора векторов, внешнее произведение - выполняет попарно(слева берется первый элемент и складывается со всеми эл-тами правого аргуметра и т.д.)

```
      3 2 4 ∘.+ 5 3
8 6
7 5
9 7
```

Прелесть в том, что мы можем любой операнд давать (умножение, сложение, вычитание), если дадим два вектора векторов в нашу ф-цию dist-вычислятся растояния между итым и житым и получаем результат в виде матрицы.

```
      3 2 4 ∘.× 5 3
15  9
10  6
20 12
      3 2 4 ∘.- 5 3
¯2  0
¯3 ¯1
¯1  1
      (1 2)(3 4) ∘.dist (2 1)(4 2)
1.414213562 3
3.16227766  2.236067977
```

Превратим a и b в 16 векторов (каждый размерности 100)

```
      ⍴ab
16 100
      ab←⊂[2]ab
      ⍴ab
16
      ⍴¨ab
 100  100  100  100  100  100  100  100  100  100  100  100  100  100  100  100
```

Получили матрицу 16 на 16 парных расстояний, по диагонали матрицы 0

```
      r←ab ∘.dist ab
```

Сделаем вот такую m

```
      m←?3 3⍴9
```

Надо найти максимальный элемент в этой m (будет равно 9)

```
      m=⌈/,m         ⍝ логическая матрица,где 1=max
0 0 1
0 0 0
0 0 0
      m
4 6 9
2 6 6
7 5 7
```

Координаты точки:

```
      ⍸m=⌈/,m
1 3
```

В нашем случае r:

```
      ⍸r=⌈/,r
3 15  15 3
```

Т.к. матрицы симметричные возьмем из них первые i=3, j=15

```
      i j←↑⍸r=⌈/,r
      i
3
      j
15
```

r1 - расстояния всех ab до первой точки:

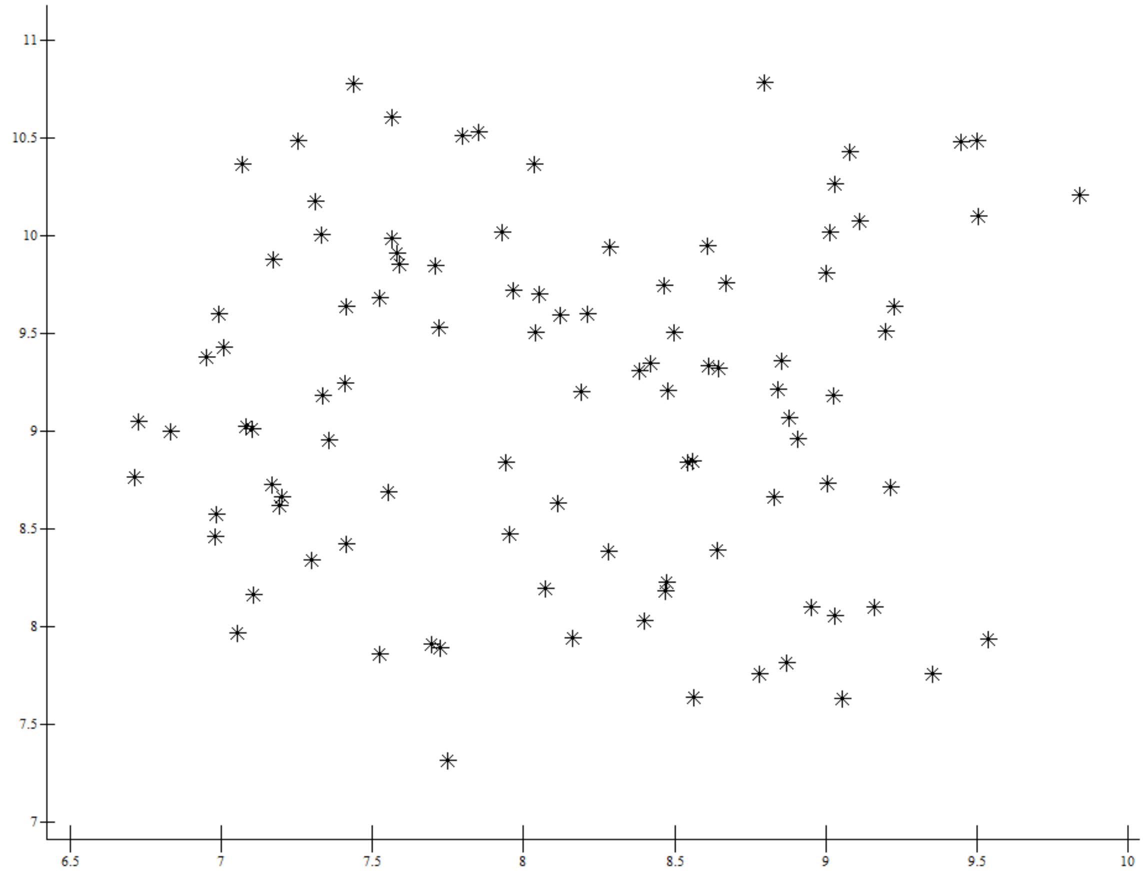
```
      r1←(+/(ab-ab[3])*2)*0.5
```

Так же считаем r2:

```
      r2←(+/(ab-ab[15])*2)*0.5
```

Строим:

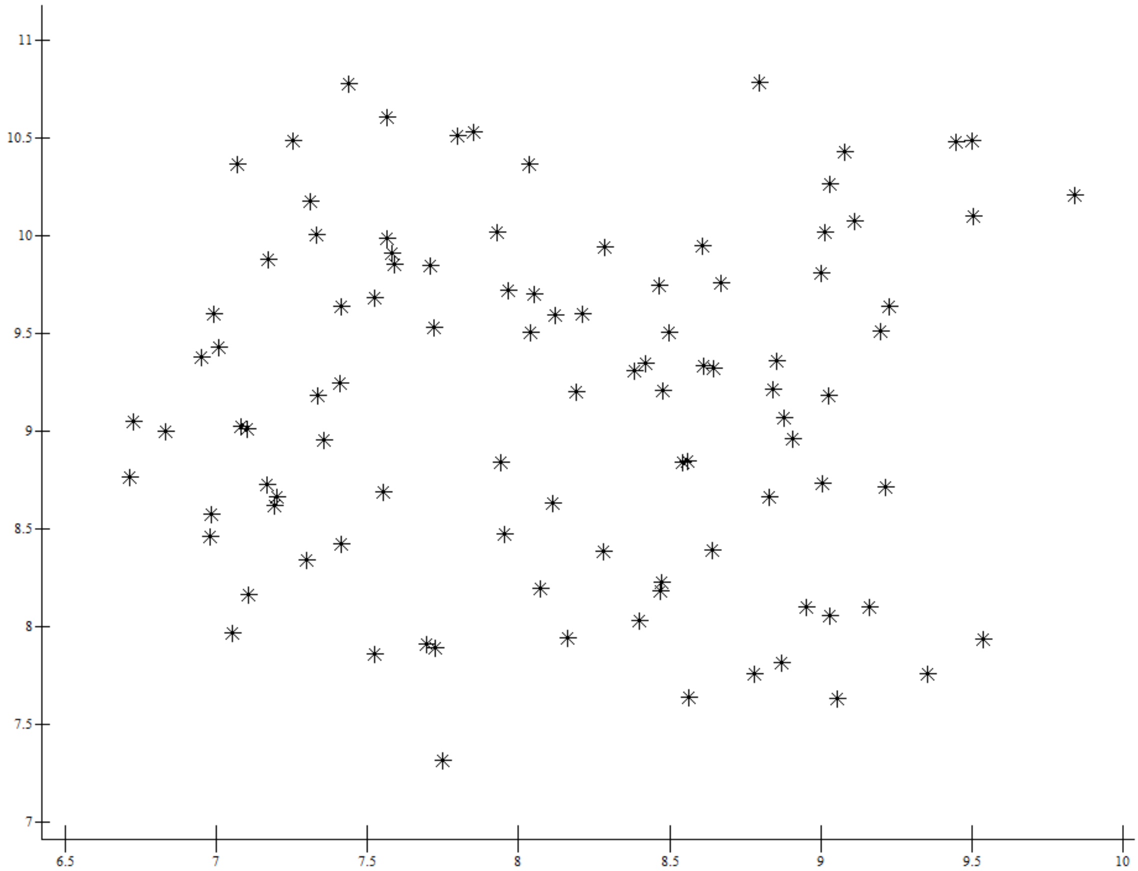
```
      plot ∊¨r1 r2
```



Вот все расстояния от мервой до второй точки, т.к. опять мы понизили все наши расстояния.

Соберем все это хозяйство в ф-цию maxproj (расстояние через бве максимальные равноудаленные точки) и проверим:

```
      ⎕vr'maxproj'
     ∇ r←maxproj ab;i;j;r1;r2
[1]    r←ab∘.{+/(⍺-⍵)*2}ab
[2]    i j←↑⍸r=⌈/,r
[3]    r1←(+/¨(ab-ab[i])*2)*0.5
[4]    r2←(+/¨(ab-ab[j])*2)*0.5
[5]    r←r1 r2
     ∇
      r1 r2←maxproj ab
      ⍴¨r1 r2
 16  16
      plot r1 r2
      plot maxproj ab
```


11
10.5
10
9.5
9
8.5
8
7.5
7
6.5
7
7.5
8
8.5
9
9.5
10