


Это сравнительный анализ, проведённый выдающимся нейрофизиологом Уайлдером Пенфилдом

Курпатов А. (с)



Выш. Мат. в машинном обучении

К.Т.Н. ПАК А.А.

ПО МАТЕРИАЛАМ ШАД ЯНДЕКС

Зачем на функции в ML?

1. $Y = \{0,1\}$ — бинарная классификация. Например, мы можем предсказывать, кликнет ли пользователь по рекламному объявлению, вернет ли клиент кредит в установленный срок, сдаст ли студент сессию, случится ли определенное заболевание с пациентом (на основе, скажем, его генома).
2. $Y = \{1, \dots, M\}$ — многоклассовая (multi-class) классификация. Примером может служить определение предметной области для научной статьи (математика, биология, психология и т.д.).
3. $Y = \{0,1\}^M$ — многоклассовая классификация с пересекающимися классами (multi-label classification). Примером может служить задача медицинской диагностики, где для пациента нужно определить набор заболеваний, которыми он страдает.
4. Ранжирование — задача, в которой требуется восстановить порядок на некотором множестве объектов. Основным примером является задача ранжирования поисковой выдачи, где для любого запроса нужно отсортировать все возможные документы по релевантности этому запросу.
5. Частичное обучение (semi-supervised learning) — задача, в которой для одной части объектов обучающей выборки известны и признаки, и ответы, а для другой только признаки. Такие ситуации возникают, например, в медицинских задачах, где получение ответа является крайне сложным (например, требует проведения дорогостоящего анализа).

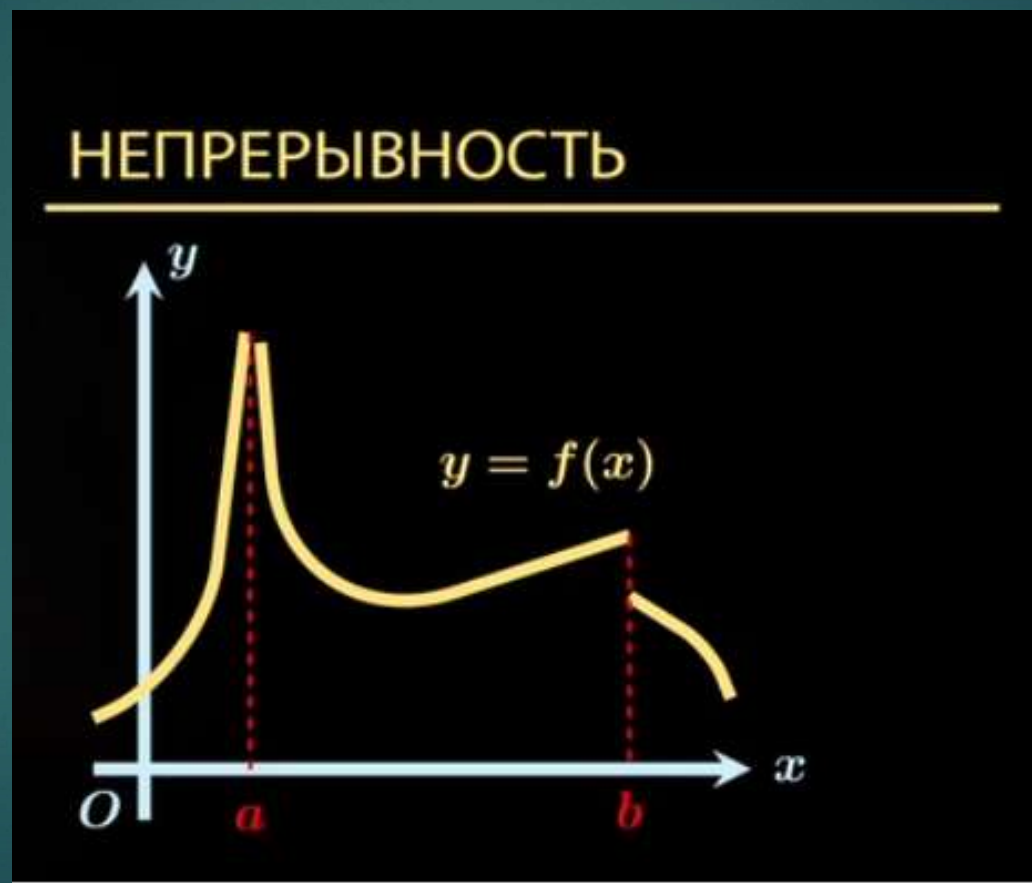
ФУНКЦИЯ

- ▶ $x \rightarrow f(x)$
- ▶ Функция – это некоторое соответствие между различными аргументами x и значениями функции $f(x)$, при этом каждому значению аргумента соответствует одно значение $f(x)$
- ▶ $x \in R$
- ▶ $D(f)$ -область определения функции
- ▶ $E(f)$ -область значения функции
- ▶ **Примеры:**
- ▶ $f(x) = \frac{1}{x-1}$; $D(f) = R \setminus \{1\}$; $E(f) = R \setminus \{0\}$;
- ▶ $f(x) = 2^x$; $D(f) = R$; $E(f) = (0; \infty)$

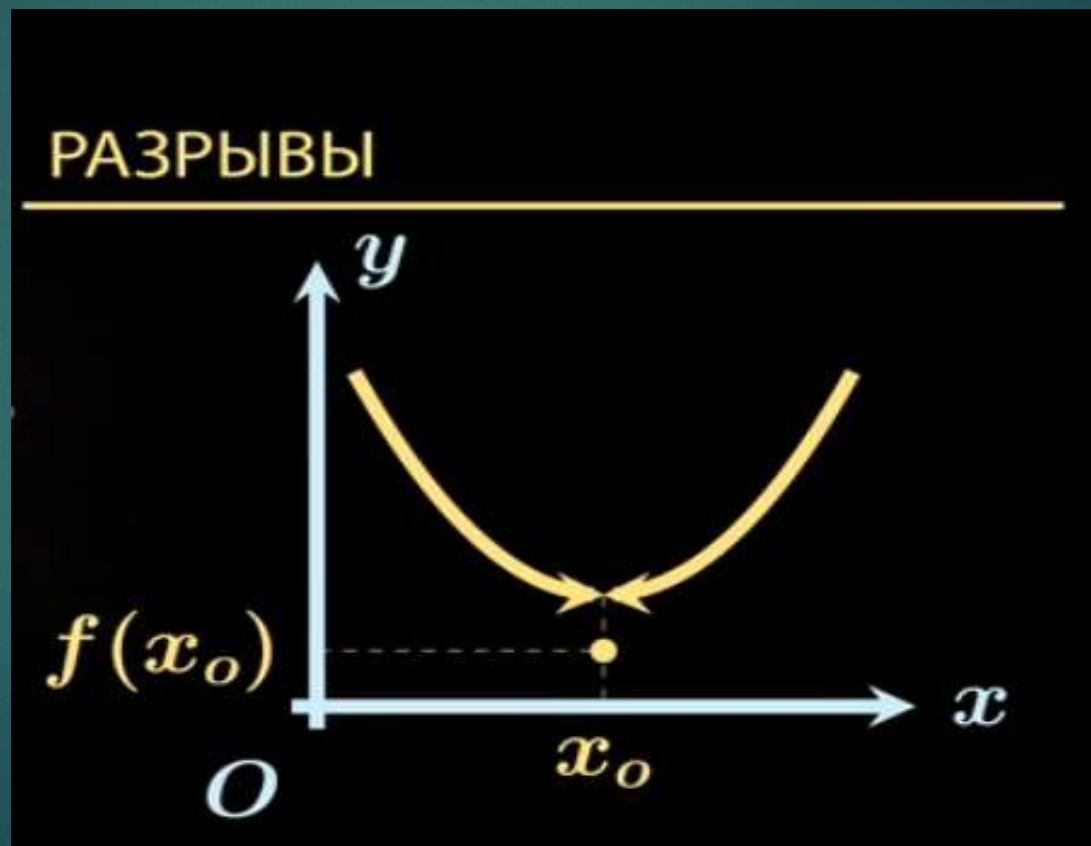
Примеры графиков функции:



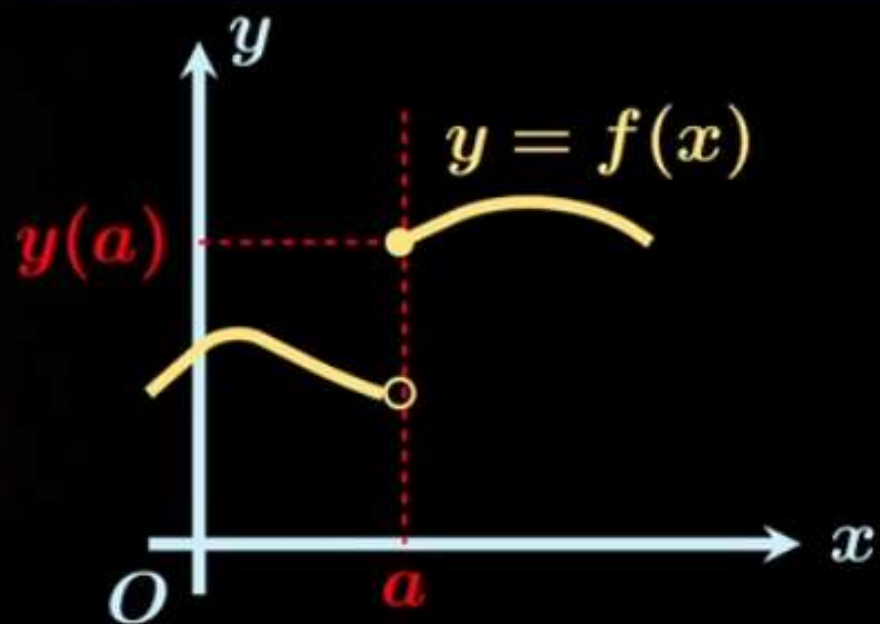
Пример непрерывности функции



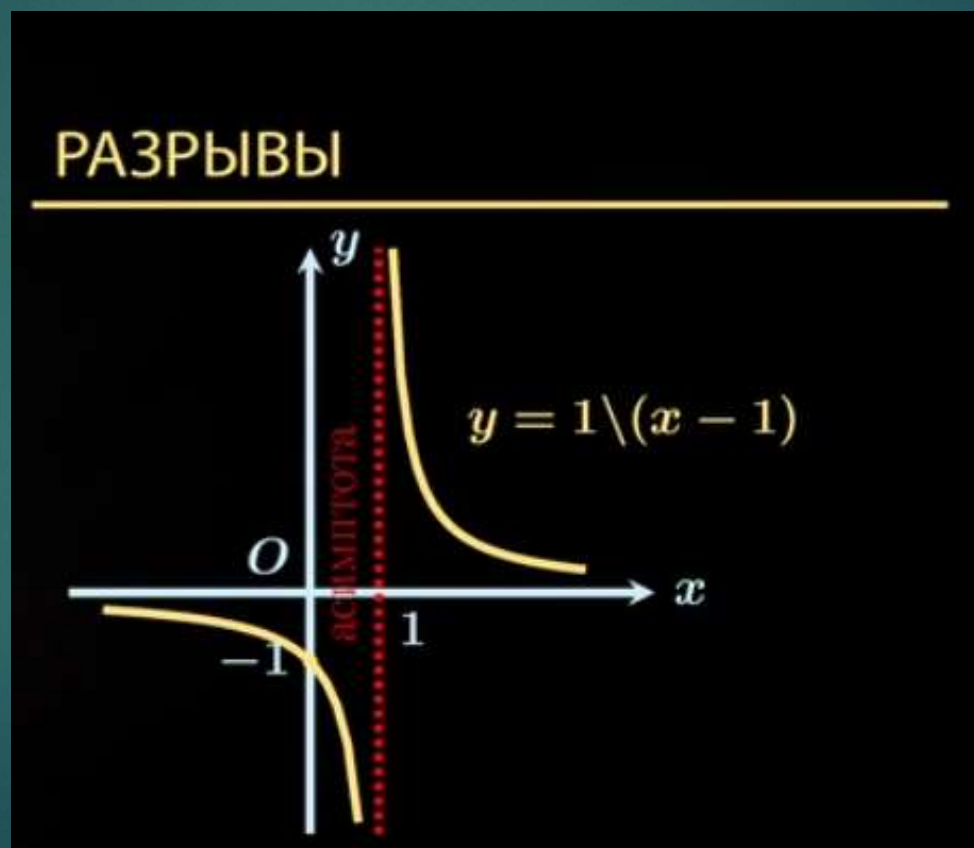
Разрывы в функции



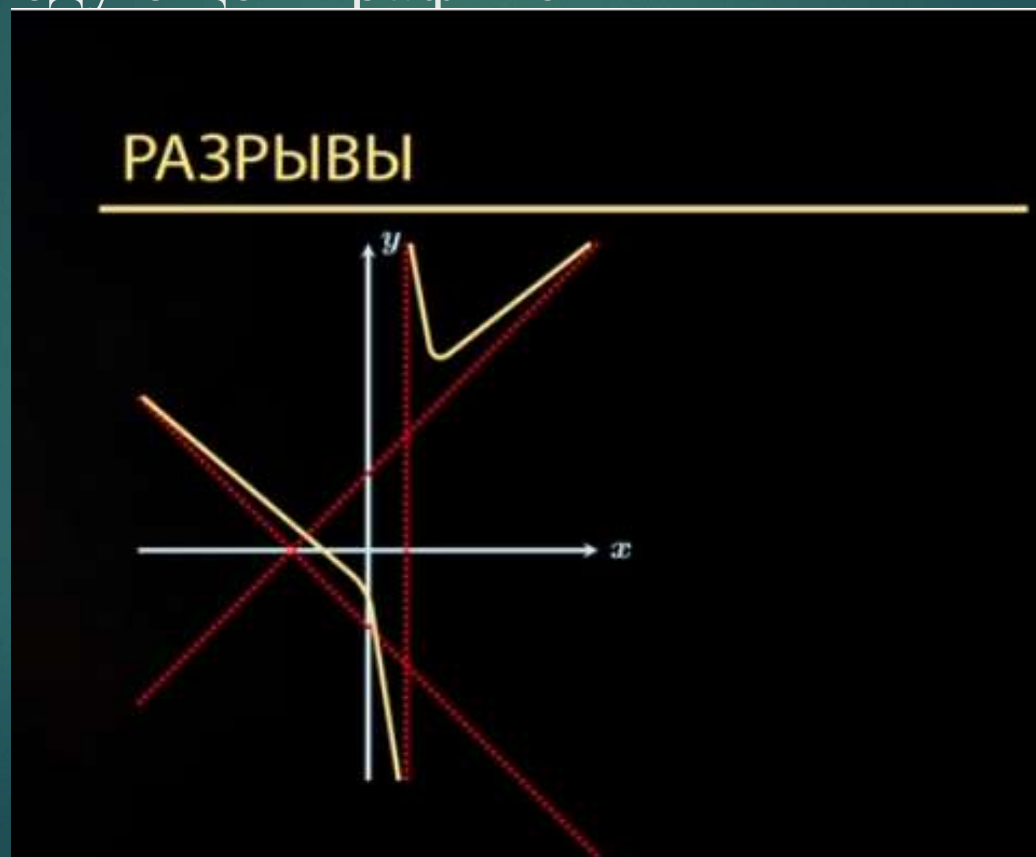
РАЗРЫВЫ



Асимптота-это прямая к которой функция может приближаться очень близко,но при этом не будет ее пересекать

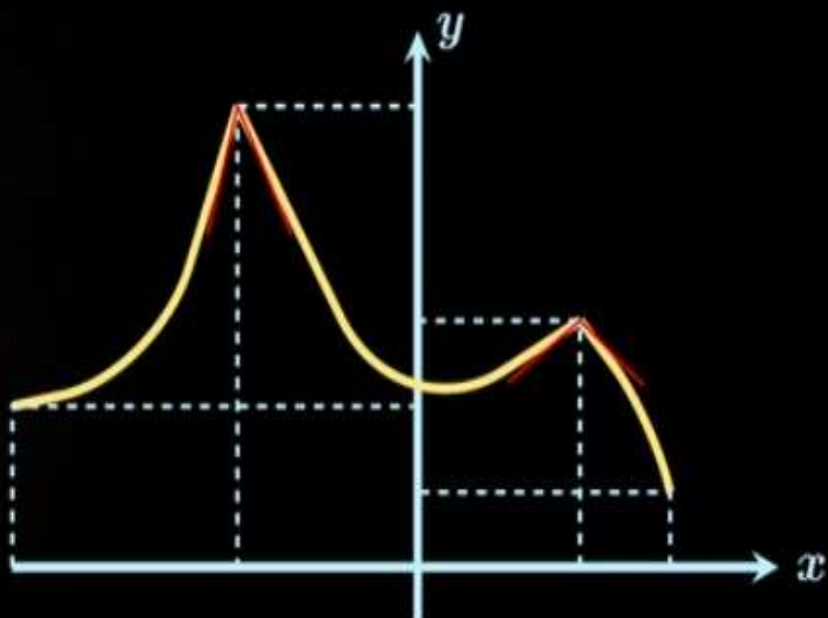


Асимптоты бывают разных видов, не только вертикальные, бывают и наклонные асимптоты, это показано в следующем графике

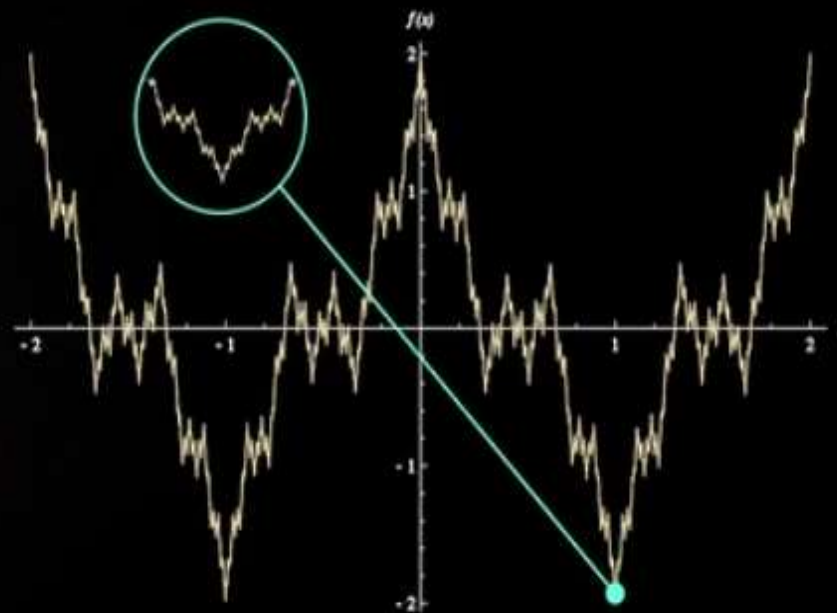


У графика функции могут быть не только разрывы, но и какие то углы, **гладкость** —это отсутствие углов

НЕФОРМАЛЬНЫЙ ВЗГЛЯД НА ГЛАДКОСТЬ



ПРИМЕР НИГДЕ НЕ ГЛАДКОЙ ФУНКЦИИ

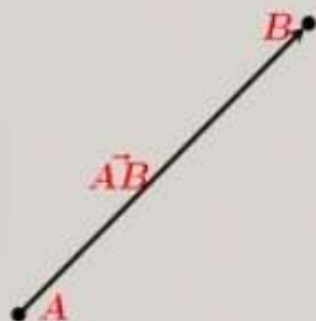


РЕЗЮМЕ

- ▶ Понятие функции
- ▶ Непрерывность функции, разрывы
- ▶ Гладкость функции

2.3. Линейная алгебра. Векторы

ВЕКТОРЫ И МАТРИЦЫ



$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

Что мы помним со
школьной программы
про векторы и матрицу?

Зачем нужны векторы и матрицы в анализе данных и откуда они берутся?!

- ▶ Рассмотрим сеть магазинов (для каждого из них нужно предсказать прибыль, это нужно для того чтобы, (если мы будем знать что в один из магазинов упадет прибыль, мы можем) принять меры, чтобы избежать этого)
- ▶ Признаковое описание:
- ▶ **Задача:** предсказать прибыль магазина в следующем месяце
- ▶ Магазин- это объект
- ▶ Признаки-числовые характеристики объектов



Признаковое описание

$x = (50, 47, 52, 55, 5, 1, 2, 55.73, 37.59, 1)$



Прибыль в предыдущие
4 месяца

$x = (50, 47, 52, 55, 5, 1, 2, 55.73, 37.59, 1)$



Планируемое число акций для
трех основных категорий:
кондитерские изделия,
овощи и фрукты, мясо



Акции под каждую категорию товаров

$x = (50, 47, 52, 55, 5, 1, 2, 55.73, 37.59, 1)$



Географические координаты
магазина



Местоположение магазинов

$x = (50, 47, 52, 55, 5, 1, 2, 55.73, 37.59, 1)$



Количество праздничных дней
в следующем месяце



Праздничные дни

Если 10 дней выходных, значит и прибыли
меньше

$x = (50, 47, 52, 55, 5, 1, 2, 55.73, 37.59, 1)$

Вектор — набор чисел
(очень неформально)

Пример 2. Из какого сорта винограда сделано вино?!



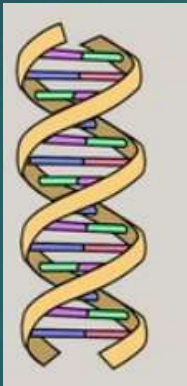
$x = (14.23, 1.71, 3.43, 5.64, 3.92)$

Химические анализы

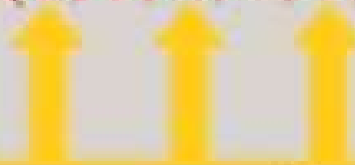
- ▶ Содержание алкоголя
- ▶ Щелочность
- ▶ Насыщенность цвета



Пример 3. Заболеет ли пациент раком в течение 5 лет?
(Каждый человек характеризуется своим геномом, геном определяет каким он вырастит, цвет волос, глаз и т.д. Именно геном определяет некоторые наследственные заболевания, напр. *рак.



$x = (0, 1, 0, 0, 1, \dots)$



Наличие мутаций в геноме

А если объектов несколько?

В этом случае данные приобретают двумерную структуру, напр. несколько магазинов, несколько бутылок вина

36,18	2	2	2	3	59090
46,47671233	1	0	4	3	14773
45,13434658	2	0	3	3	19376
25,88766123	4	1	4	3	16098
25,70410959	4	1	3	3	20338
33,03	1	0	3	1	501667
46,44931597	4	3	2	1	26100
51,24383562	2	0	4	2	20727
46,8739726	2	0	1	3	27861

36,18	2	2	2	3	59090
46,47671233	1	0	4	3	14773
45,13434658	2	0	3	3	19376
25,88766123	4	1	4	3	16098
25,70410959	4	1	3	3	20338
33,03	1	0	3	1	501667
46,44931597	4	3	2	1	26100
51,24383562	2	0	4	2	20727
46,8739726	2	0	1	3	27861

Объект

36,18	2	2	2	3	59090
46,47671233	1	0	4	3	14773
45,13434658	2	0	3	3	19376
25,88766123	4	1	4	3	16098
25,70410959	4	1	3	3	20338
33,03	1	0	3	1	501667
46,44931597	4	3	2	1	26100
51,24383562	2	0	4	2	20727
46,8739726	2	0	1	3	27861

Признак

36,18	2	2	1	2	3	59090	1
46,47671233	1	0	1	4	3	14773	1
45,13434658	2	0	1	3	3	19376	2
25,88766123	4	1	1	4	3	16098	0
25,70410959	4	1	1	3	3	20338	0
33,03	1	0	1	3	1	501667	2
46,44931597	4	3	1	2	1	26100	0
51,24383562	2	0	0	4	2	20727	0
46,8739726	2	0	1	1	3	27861	0
39,8630137	2	3	1	4	2	27861	0
37,09	2	0	1	4	3	55825	1
38,14	2	3	1	3	2	60000	1
45,46840315	1	2	1	1	1	40000	1

Матрица

Резюме

- ▶ Основные объекты-матрицы и векторы
- ▶ Какие операции на них можно ввести ?
- ▶ Какими свойствами они обладают?

Операции в линейных пространствах

- ▶ Векторное пространство-множества
- ▶ Вектор-элемент векторного пространства
- ▶ Множество- V
- ▶ Две операции: сумма векторов и умножение векторов на число
- ▶ Произведение любого вектора на любое число-вектор Сумма и умножение на число должны удовлетворять 8 аксиомам

- ▶ $x + y = y + x$, для любых $x, y \in V$
- ▶ $x + (y + z) = (x + y) + z$, для любых $x, y, z \in V$
- ▶ $\exists 0 \in V : x + 0 = x \forall x \in V$
- ▶ $\forall x \in V \exists -x \in V : x + (-x) = 0$
- ▶ $\alpha(\beta x) = (\alpha\beta)x$
- ▶ $1 \cdot x = x$
- ▶ $(\alpha + \beta)x = \alpha x + \beta x$
- ▶ $\alpha(x + y) = \alpha x + \alpha y$

Евклидово векторное пространство- \mathbb{R}^n (простые векторные пространства)-состоит из векторов, каждый элемент которых вещественные числа

» Евклидово векторное пространство — \mathbb{R}^n

▸ Точки на плоскости — \mathbb{R}^2

▸ Точки в пространстве — \mathbb{R}^3

Поэлементное сложение

- ▶ $a = (a_1, \dots, a_n)$

- ▶ $b = (b_1, \dots, b_n)$

- ▶ $a + b = (a_1 + b_1, \dots, a_n + b_n)$

Поэтапное умножение

- ▶ $\mathbf{a} = (a_1, \dots, a_n)$

- ▶ $\beta \in \mathbb{R}$

- ▶ $\beta \mathbf{a} = (\beta a_1, \dots, \beta a_n)$

Пример. Сорт вина (для каждой бутылки)

- ▶ Векторное пространство — это все возможные признаковые описания вина, при этом конкретный вектор — это конкретная бутылка вина.
- ▶ Векторы можно усреднять (можем искать сорт с наибольшим содержанием алкоголя)

$$\bar{x} = \frac{1}{m} \sum_{i=1}^m (x_1^i, \dots, x_n^i)$$

Резюме

- ▶ С помощью векторов в машинном обучении описывают объекты реального мира
- ▶ Неформально, вектор-набор чисел
- ▶ Формально, вектор- элемент векторного пространства
- ▶ Векторное пространство- множество с двумя операциями

2.5. Матрицы

$$A = \begin{pmatrix} 12 & 7 & 21 & 31 & 11 \\ 45 & -2 & 14 & 27 & 19 \\ -3 & 15 & 36 & 71 & 26 \\ 4 & -13 & 55 & 34 & 15 \end{pmatrix}$$

› Два индекса: строка и столбец

▶ $a_{12} = 7$

▶ $a_{31} = -3$

› Пространство всех матриц 4 на 5: $\mathbb{R}^{4 \times 5}$

4 бутылки разных вин, 5 описывающие признаки

36.18	2	2	1	2	3	59090	1
46.47671233	1	0	1	4	3	14773	1
45.13434658	2	0	1	3	3	19376	2
25.88766123	4	1	1	4	3	16098	0
25.70410959	4	1	1	3	3	20338	0
33.03	1	0	1	3	1	501667	2
46.44931597	4	3	1	2	1	26100	0
51.24383562	2	0	0	4	2	20727	0
46.8739726	2	0	1	1	3	27861	0
39.8630137	2	3	1	4	2	27861	0
37.09	2	0	1	4	3	55825	1
38.14	2	3	1	3	2	60000	1
45.46849315	4	3	1	1	1	40000	1
42.99726027	4	3	1	4	2	40343	0
29.98082192	4	3	0	4	2	27.583	2
46.20547945	4	3	1	2	2	45385	1

Двумерная структура: по строкам объекты, по столбцам признаки (дело в том, что бутылок много и каждый из них описывается числами или числовыми характеристиками этих бутылок)

Зачем нужны матрицы?

- ▶ Их используют в системах уравнений
- ▶ $y = (1, 0, 0, 1)$ возьмем 4 бутылки вина, если 1 и 4 бутылка подленная, а 2 и 3 подделка, за кадируем и вектором y
- ▶ вектор размера n —тоже

$$\begin{cases} 12w_1 + 7w_2 + 21w_3 + 31w_4 + 11w_5 = 1 \\ 45w_1 - 2w_2 + 14w_3 + 27w_4 + 19w_5 = 0 \\ -3w_1 + 15w_2 + 36w_3 + 71w_4 + 26w_5 = 0 \\ 4w_1 - 13w_2 + 55w_3 + 34w_4 + 15w_5 = 1 \end{cases}$$

▶ Вектор-столбец: $w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{pmatrix}$

▶ Вектор-строка: $w = (w_1, w_2, w_3, w_4, w_5)$

Умножение

Умножение матрицы $m \times n$ на вектор-столбец $n \times 1$:

$$Aw = \begin{pmatrix} \sum_{i=1}^n a_{1i}w_i \\ \sum_{i=1}^n a_{2i}w_i \\ \dots \\ \sum_{i=1}^n a_{mi}w_i \end{pmatrix}$$

Первая строка матрицы

Матричная запись системы уравнений

$$Aw=y$$

$$\begin{pmatrix} 12 & 7 & 21 & 31 & 11 \\ 45 & -2 & 14 & 27 & 19 \\ -3 & 15 & 36 & 71 & 26 \\ 4 & -13 & 55 & 34 & 15 \end{pmatrix} \times \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 47 \\ 55 \\ 63 \\ 33 \end{pmatrix}$$

Линейное преобразование

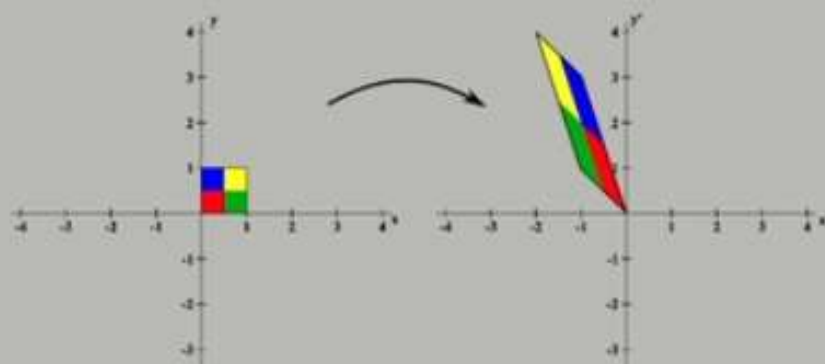
- › Матрица $m \times n$
- › Умножаем на вектор длины n , получаем вектор длины m
- › Матрица задает линейное преобразование

Резюме

- › Матрица — таблица с числами
- › Вектор — тоже матрица
- › Через матрицы можно записывать системы линейных уравнений
- › Матрицы задают линейные функции из одних векторных пространств в другие

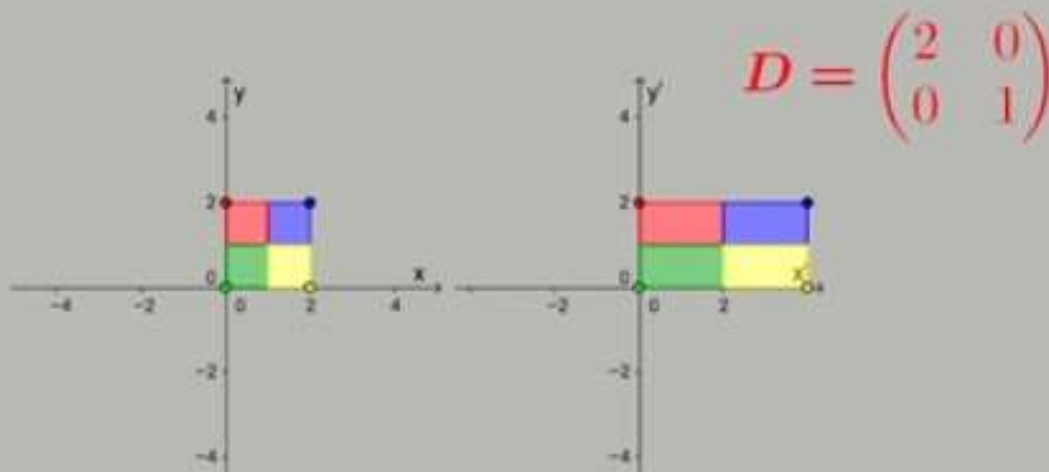
Типы матрицы

$$\triangleright T(x, y) = \begin{bmatrix} -1 & -1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$



Диагональные матрицы

- › Вне главной диагонали — только нули
- › Частный случай — единичная матрица I



Ортогональные матрицы

› $A^T A = A A^T = I$

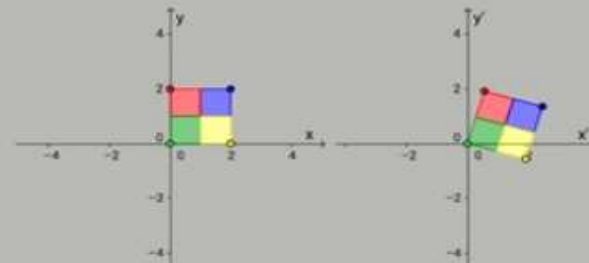
› Сохраняет длины векторов: $\|Ax\| = \|x\|$

› Сохраняет скалярные произведения:

$$\langle Ax, Az \rangle = \langle x, z \rangle$$

› Последовательность поворотов и вращений

› $A = \begin{pmatrix} 0.96 & 0.28 \\ -0.28 & 0.96 \end{pmatrix}$



› Определитель: во сколько увеличится площадь единичного квадрата после применения линейного преобразования

› У ортогональных матриц: $\det A = \pm 1$

Симметричные матрицы

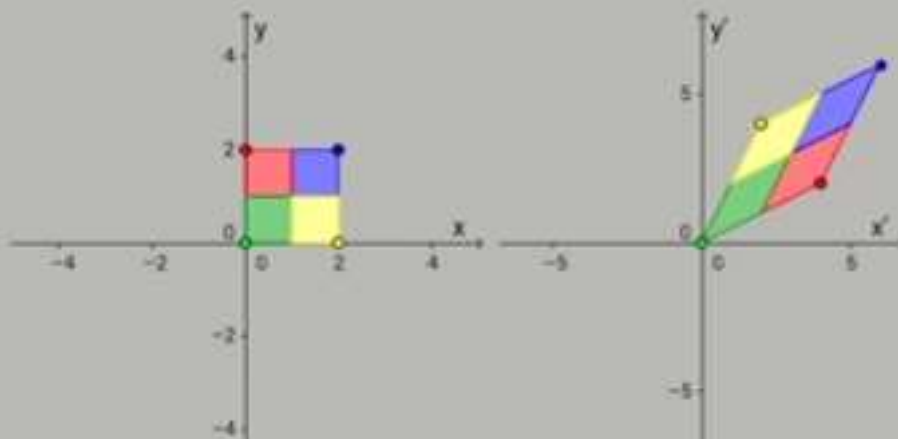
› $A = A^T$

› Можно представить в виде $A = QDQ^T$

› D — диагональная

› Q — ортогональная

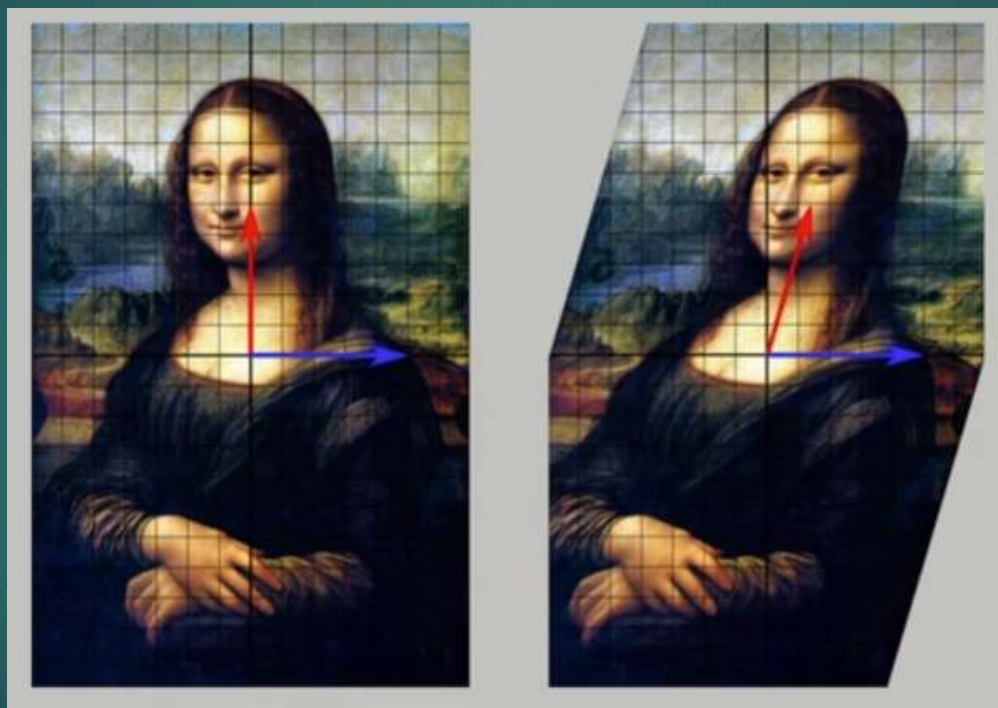
$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$



Резюме

- › Диагональные матрицы растягивают вдоль осей
- › Ортогональные матрицы вращают и отражают
- › Симметричные матрицы — композиция ортогональных и диагональных преобразований

Линейные преобразования



Собственные векторы

› $Ax = \lambda x, x \neq 0$

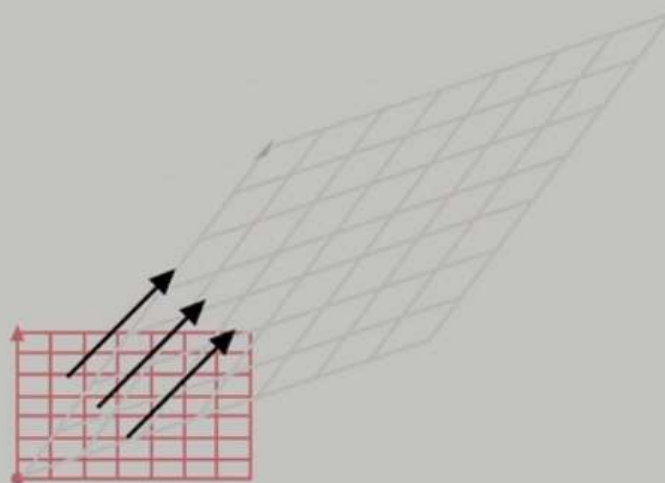
› x — собственный вектор

› λ — собственное значение

› У матрицы $n \times n$ — не более n собственных значений

Пример

$$A = \begin{pmatrix} 2 & 1 \\ 1.5 & 2 \end{pmatrix}$$



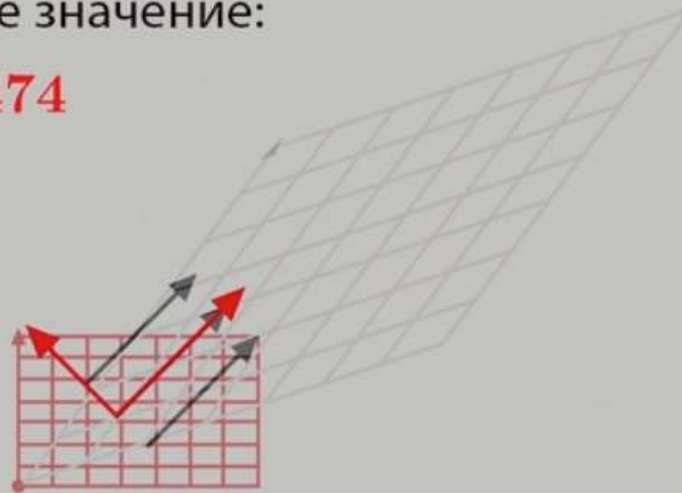
Пример

› Собственный вектор:

$$\nu_1 \approx (0.632456, 0.774597)$$

› Собственное значение:

$$\lambda_1 \approx 3.22474$$



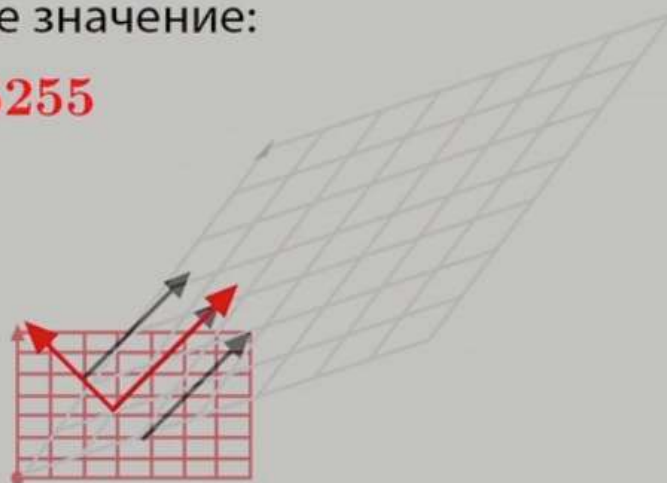
Пример

› Собственный вектор:

$$\nu_2 \approx (-0.632456, 0.774597)$$

› Собственное значение:

$$\lambda_2 \approx 0.775255$$



Зачем это нужно?

- ▶ Понижение размерности
- ▶ Собственные векторы дают наиболее характерные преобразования матрицы

Резюме

- ▶ Собственные векторы – это направления в которых матрица сжимает или растягивает, но не поворачивает
- ▶ Необходимо для сжатие матриц с наименьшей потерей данных

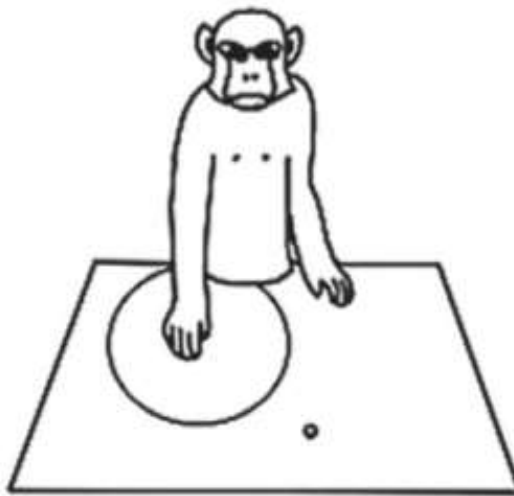
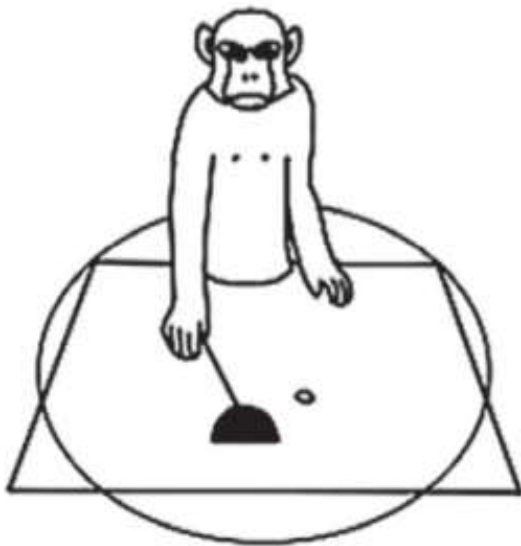


Рис. № 1. Реакция мозга обезьяны на объекты внешней среды при наличии лопатки и в её отсутствие.

Нейрофизиолог Ациси Ирики исследовал нейроны теменной доли обезьяны, отвечающие за пространственную ориентацию. Выяснилось, что, когда обезьяна получает лопатку, эти нейроны начинают реагировать на предметы, которые находятся в пределах досягаемости лопатки. Но стоит вам забрать у обезьяны лопатку, активность этих нейронов тут же спадает (рис. № 1).

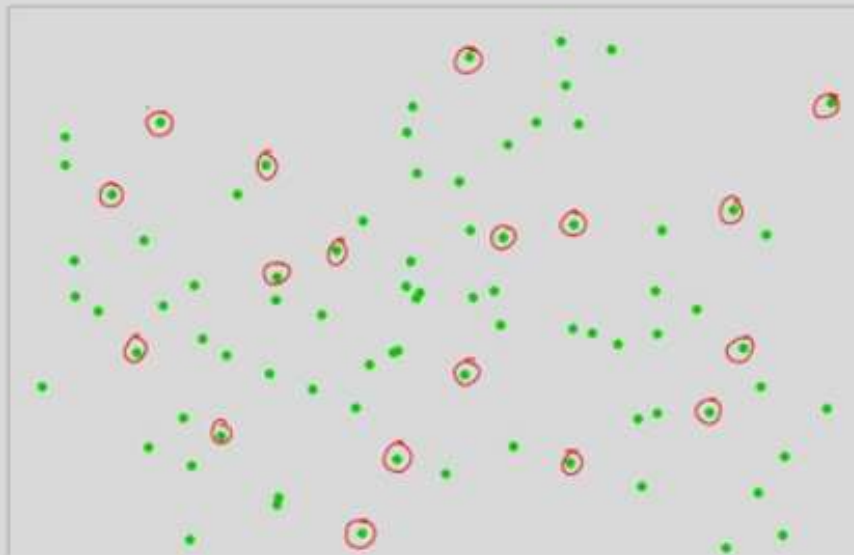
Курпатов А. (с)

Статистика

- ▶ Генеральная совокупность-множество всех тех объектов относительно которых будут сделаны выводы в рамках исследования научной проблемы
- ▶ Например: Если в рамках социального исследования рассмотрим, как очередное политическое событие повлияло на совершеннолетних жителей г.Алматы. Все те, кто является совершеннолетними будут представлять генеральную совокупность, если выбрать часть генеральной совокупности, это и будет выборкой, она отражает свойство генеральной совокупности

Выборка

- Простая случайная выборка (simple random sample)



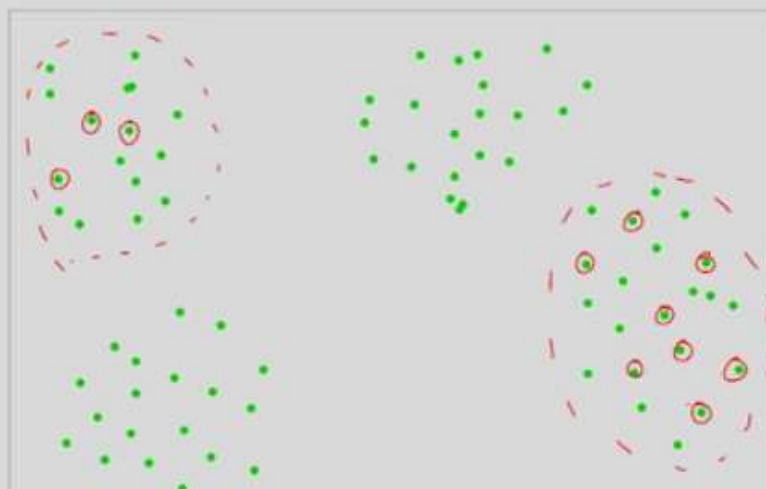
Выборка

- Стратифицированная выборка (stratified sample)



Выборка

- Групповая выборка (cluster sample)



Типы переменных. Количественные и номинативные переменные

Количественные

- непрерывные
- дискретные

Номинативные



Измеренное значение некоторого признака
Напр. рост (непрерывное [160; 190], дискретное —
количество детей в семье, мы не можем сказать 3,5 ребенка)



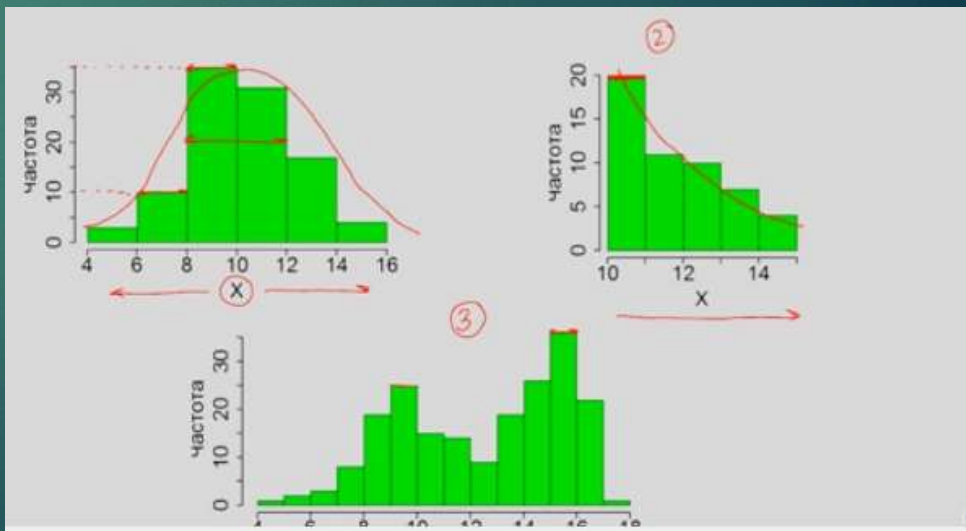
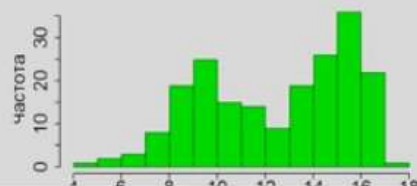
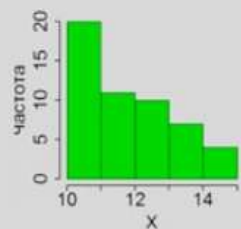
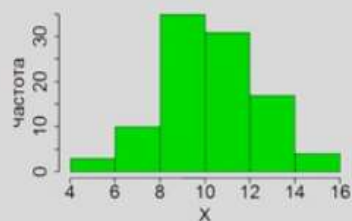
1-ж

за цифрами не стоит никакого
мат.смысла

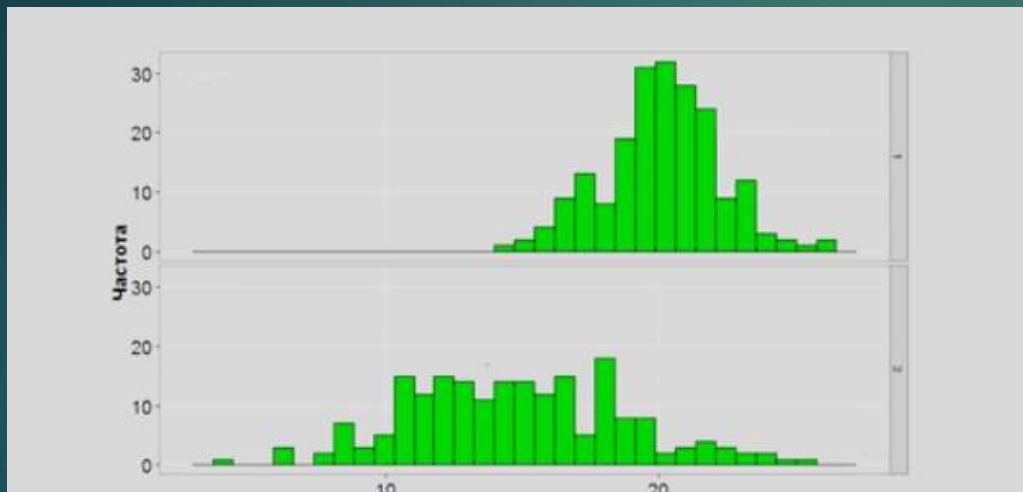
2-м

Меры центральной тенденции

Гистограмма частот



Описательные статистики



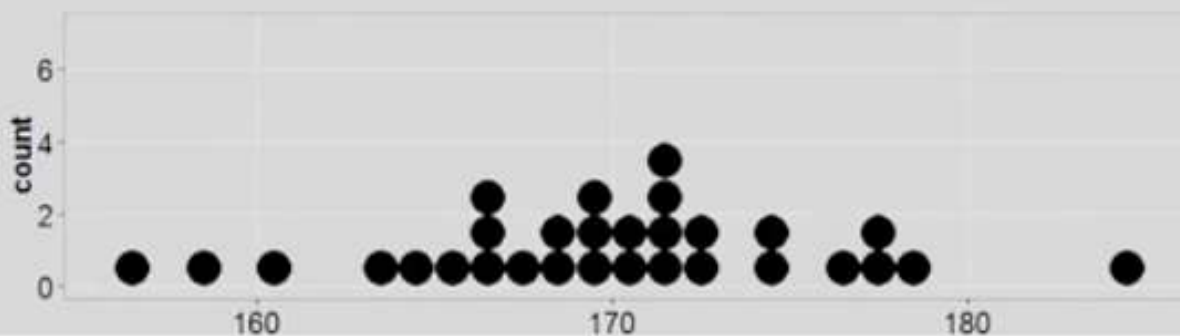
Меры центральной тенденции-это насколько высокие значения принимает переменная

Мера изменчивости- это изменчивость признака

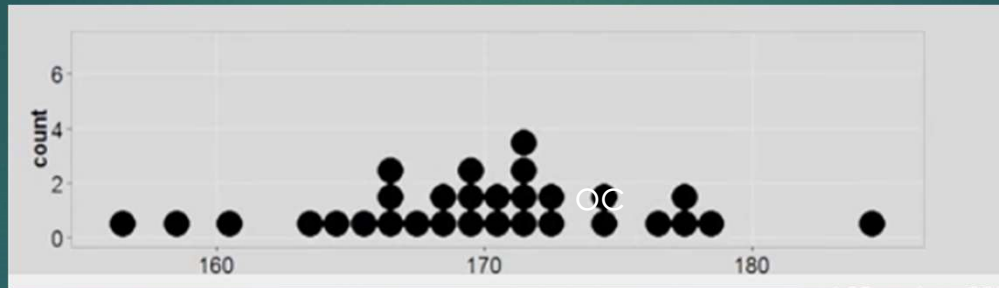
Меры центральной тенденции

Мода (Mode) – значение измеряемого признака, которое встречается максимально часто.

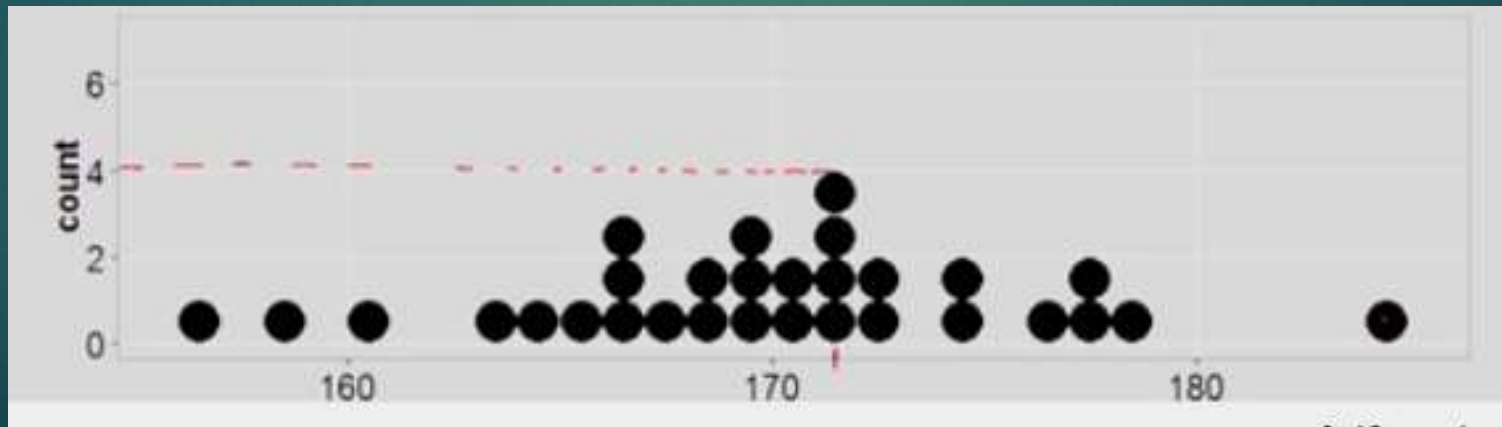
185 175 170 169 171 172 175 157 170 172 167 173 168 167 166
167 169 172 177 178 165 161 179 159 164 178 172 170 173 171



Предположем, что у нас есть выборка из $N=30$ наблюдений, допустим это рост наших испытуемых



Значение выборки будет равно 172 (мода), это означает что, чаще всего в выборке присутствует значения равные 172



Медиана

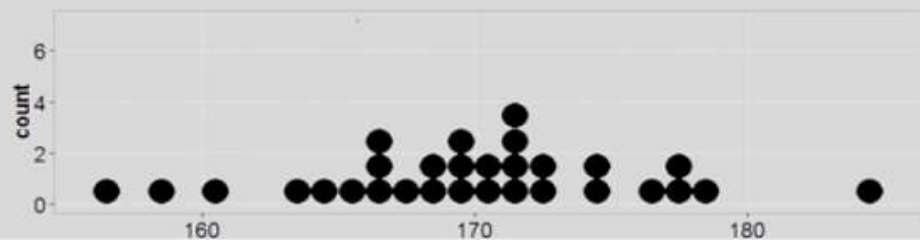
Медиана (median) – значение признака, которое делит упорядоченное множество данных пополам.

157 159 161 164 165 166 167 167 167

$N=9$
 $Me=165$

157 159 161 164 165 166 167 167 167 168 169 169 170 170 170
171 171 172 172 172 172 173 173 175 175 177 178 178 179 185

$N=30$
 $Me=\frac{170+171}{2} = 170,5$



Среднее значение

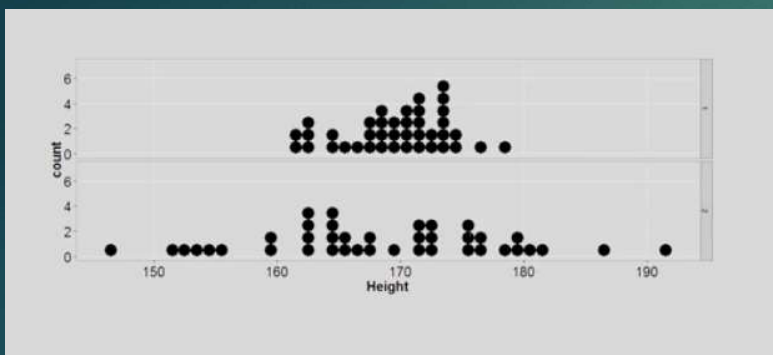
Среднее значение (mean, среднее арифметическое)
сумма всех значений измеренного признака, деленная
на количество измеренных значений.

157 159 161 164 165 166 167 167 167 168 169 169 170 170 170
171 171 172 172 172 172 173 173 175 175 177 178 178 179 185

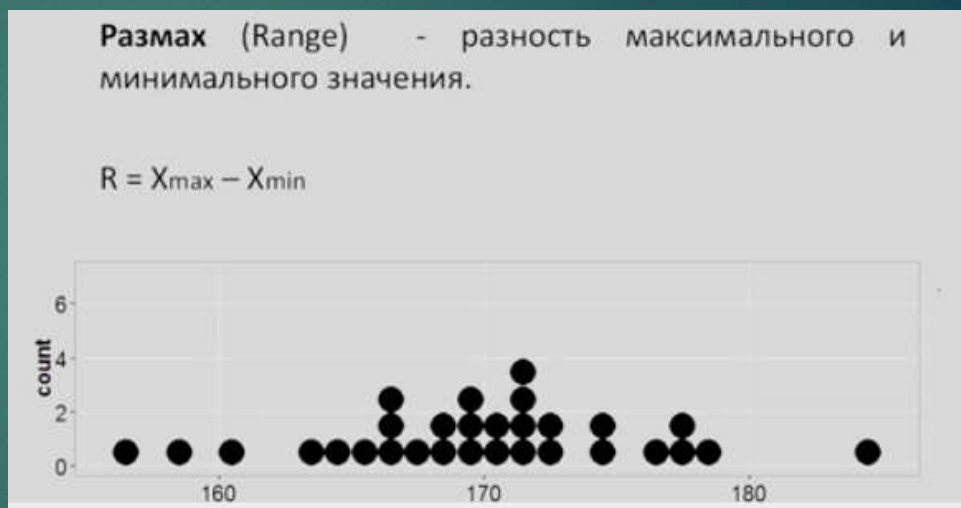


$$\bar{X} = \frac{x_1 + \dots + x_{30}}{30} = 170,4$$

Меры изменчивости



$$R = 185 - 157 = 28$$



Дисперсия

Дисперсия (variance) – средний квадрат отклонений индивидуальных значений признака от их средней величины.



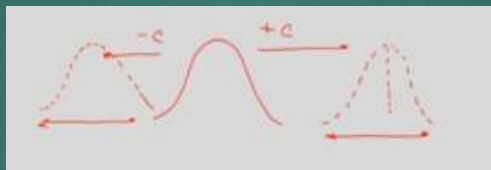
$$\bar{X}=170,4$$

$$D=\frac{\sum(x_i-\bar{x})^2}{n}$$

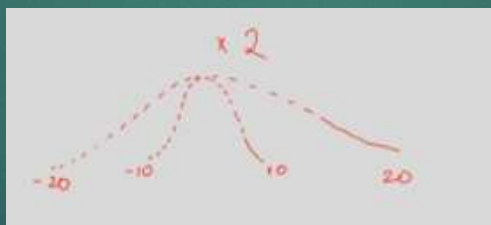
$$\sqrt{D} = \delta$$

Свойство дисперсии

$$D_{x+c} = D_x$$
$$sd_{x+c} = sd_x$$



$$D_{x*c} = D_x * c^2$$
$$sd_{x*c} = sd_x * c$$

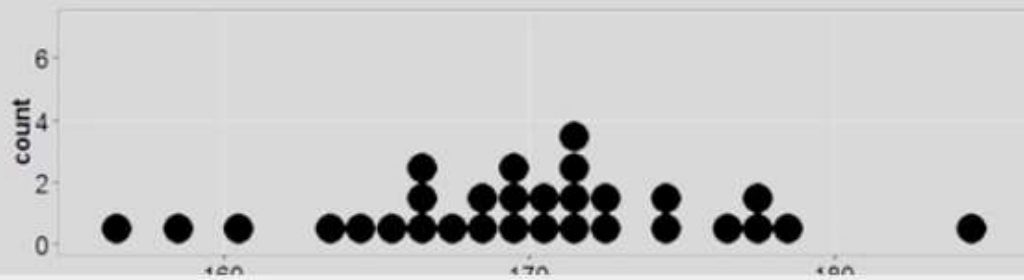


Квартили распределения и график box-plot

- ▶ Квартили распределения-это такие значения признака, которые делят упорядоченные данные на некоторое число равных частей.

Квартили – три точки (значения признака), которые делит упорядоченное множество данных на четыре равные части.

157 159 161 164 165 166 167 167 167 168 169 169 170 170 170
171 171 172 172 172 172 173 173 175 175 177 178 178 179 185

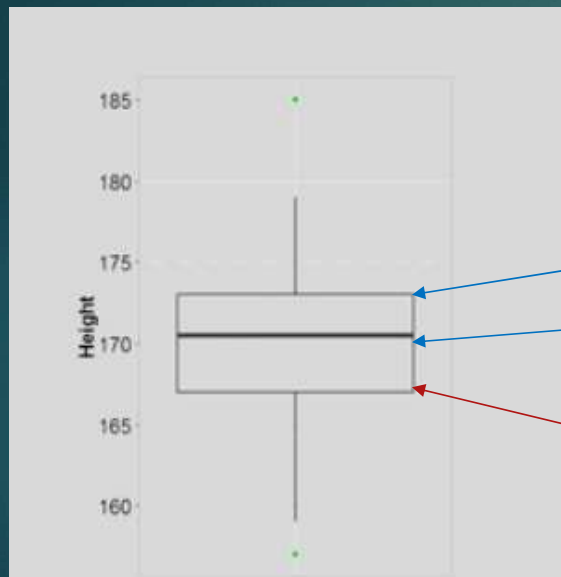


1 квартал=167

2 квартал=170,5

3 квартал=173

Box plot



3 кв

Me=2 кв

1 кв



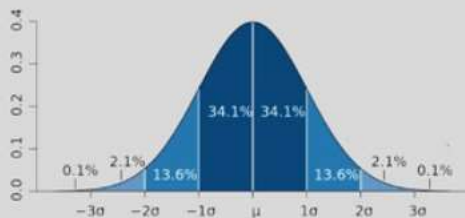
Разность между 3 кв и 1 кв называется **межквартильный размах**, если нарисовать 1,5 межквартильного размаха от 3 кв и 1,5 межквартильного размаха от 1 кв., то это и будет границами усов нашего графика

Нормальное распределение

- Унимодально

- Симметрично

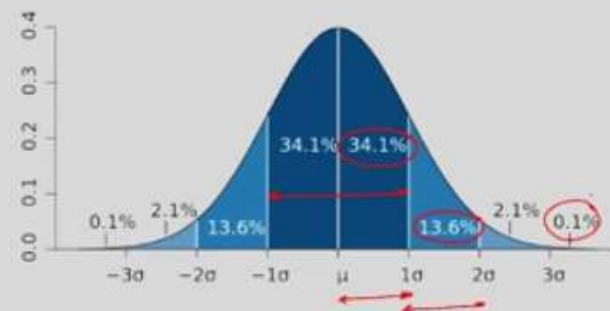
- Отклонения наблюдений от среднего подчиняются определенному вероятностному закону.



- Унимодально

- Симметрично

- Отклонения наблюдений от среднего подчиняются определенному вероятностному закону.



Стандартизация

Стандартизация или *z-преобразование* – преобразование полученных данных в стандартную Z-шкалу (Z-scores) со средним $M_z = 0$ и $D_z = 1$

$$Z_i = \frac{x_i - \bar{X}}{\delta_x} \quad (1)$$

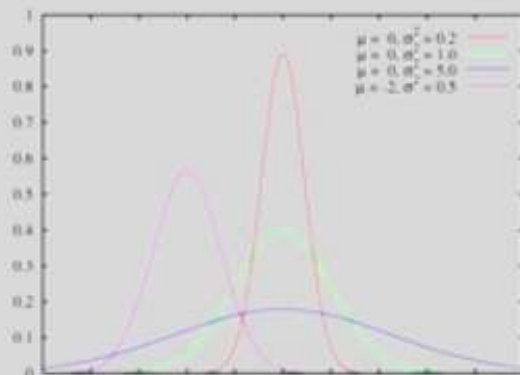
$$\bar{X} - c = \bar{X} - \bar{X} = 0 \quad (2)$$

$$D_x = \left(\frac{1}{\delta_x}\right)^2 = D_x * \frac{1}{D_x} = 1 \quad (3)$$

Правило «двух» и «трех» сигм

И Насим Талиб с его Черным Леб(е-я)дем

- $M_x \pm \sigma \approx 68\%$ наблюдений
- $M_x \pm 2\sigma \approx 95\%$ наблюдений
- $M_x \pm 3\sigma \approx 100\%$ наблюдений

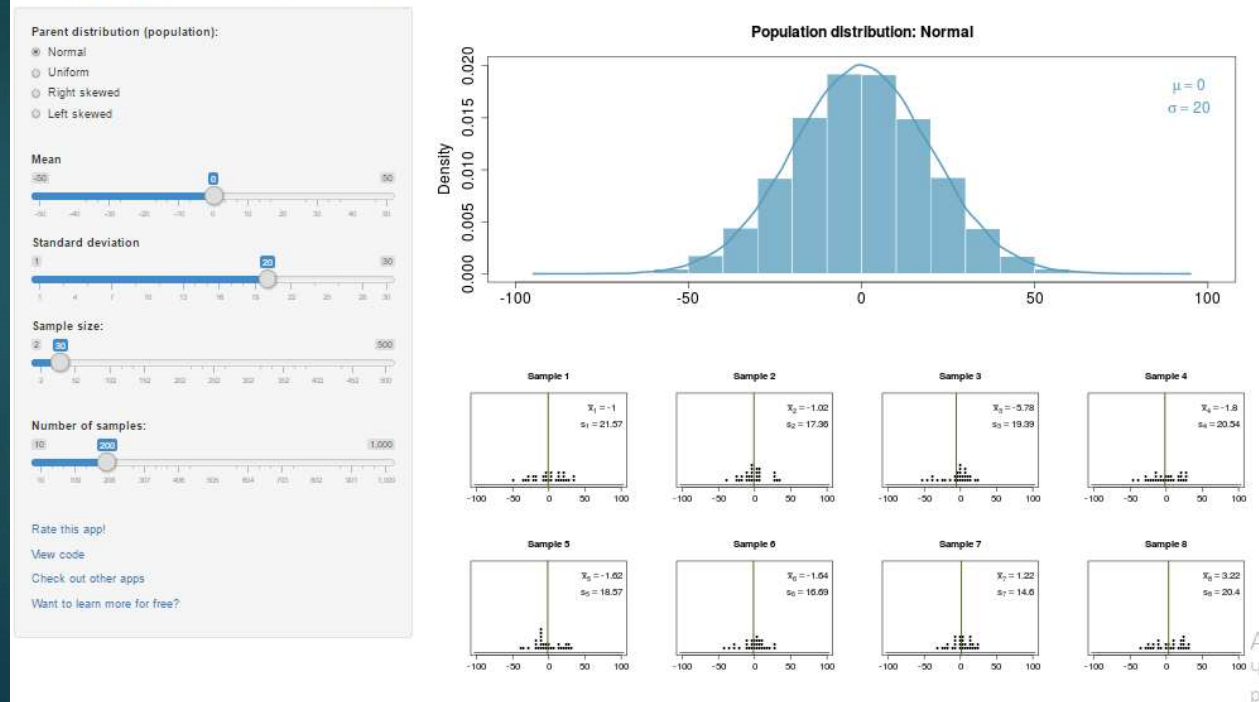


z	0,00	0,01	0,02	0,03	0,04
0,0	0,5000	0,4960	0,4920	0,4880	0,4840
0,1	0,4602	0,4562	0,4522	0,4482	0,4441
0,2	0,4207	0,4168	0,4129	0,4090	0,4052
0,3	0,3821	0,3783	0,3745	0,3707	0,3669
0,4	0,3446	0,3409	0,3372	0,3336	0,3300
0,5	0,3085	0,3050	0,3015	0,2981	0,2946
0,6	0,2743	0,2709	0,2676	0,2643	0,2611
0,7	0,2420	0,2386	0,2356	0,2327	0,2296
0,8	0,2119	0,2086	0,2054	0,2023	0,1993
0,9	0,1841	0,1814	0,1788	0,1762	0,1736
1,0	0,1587	0,1562	0,1539	0,1515	0,1492

Центральная предельная теорема

https://gallery.shinyapps.io/CLT_mean/

Central Limit Theorem for Means



Список литературы

- [1] Weber, R., Schek, H. J., Blott, S. (1998). A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. // Proceedings of the 24th VLDB Conference, New York C, 194–205.
- [2] Boriah, S., Chandola, V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. // In Proceedings of the 2008 SIAM International Conference on Data Mining (pp. 243–254).