

2. Математический аппарат



Случайные величины, выборка, случайные процессы

Случайная величина

Что такое случайная величина?

В рамках нашего курса – это некоторая функция, принимающая в зависимости от случая те или иные значения с определёнными вероятностями.

Определения

Функция распределения

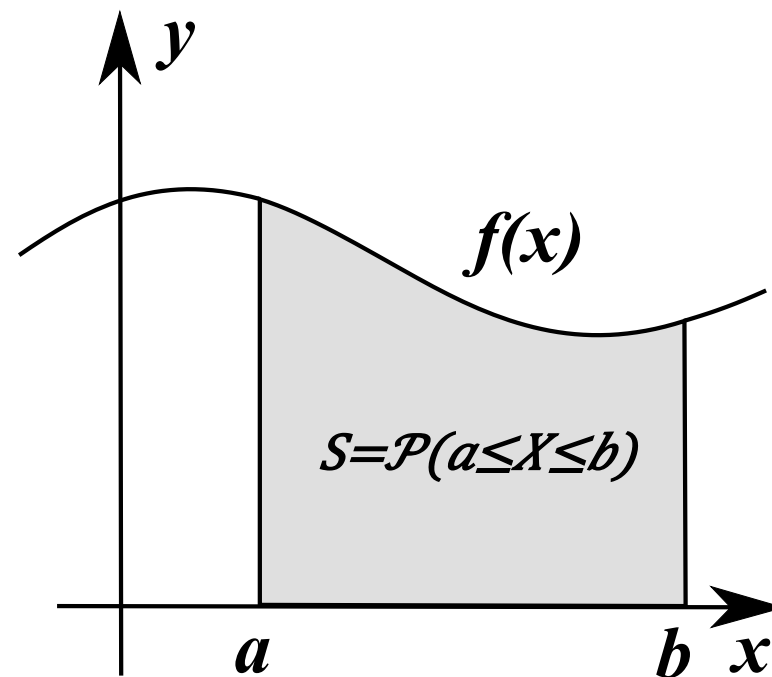
$$F(x) = \mathcal{P}(X \leq x)$$

Плотность вероятности

$$\int_{-\infty}^x f(t) dt = F(x)$$

Свойства плотности вероятности:

- 1) $\int_{-\infty}^{+\infty} f(t) dt = 1$
- 2) $\int_a^b f(t) dt = \mathcal{P}(a \leq X \leq b)$



Числовые характеристики СВ

Мат. ожидание

$$M[X] = \int_{-\infty}^{+\infty} xf(x)dx = \sum_{i=1}^n x_i p_i$$

Дисперсия

$$D[X] = M[(X - M[X])^2]$$

Квантиль

$$x_{\alpha} = \min\{x: F(x) \leq \alpha\}$$

Медиана – квантиль уровня $\alpha=0.5$.

Мода – значение СВ, которое встречается наиболее часто.

Нормальное распределение

Нормальное распределение, также называемое распределением Гаусса или Гаусса — Лапласа — распределение вероятностей, которое в одномерном случае задаётся функцией плотности вероятности, совпадающей с функцией Гаусса:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

где параметр μ — математическое ожидание (среднее значение), медиана и мода распределения, а параметр σ — среднеквадратическое отклонение (σ^2 — дисперсия) распределения.

Нормальное распределение

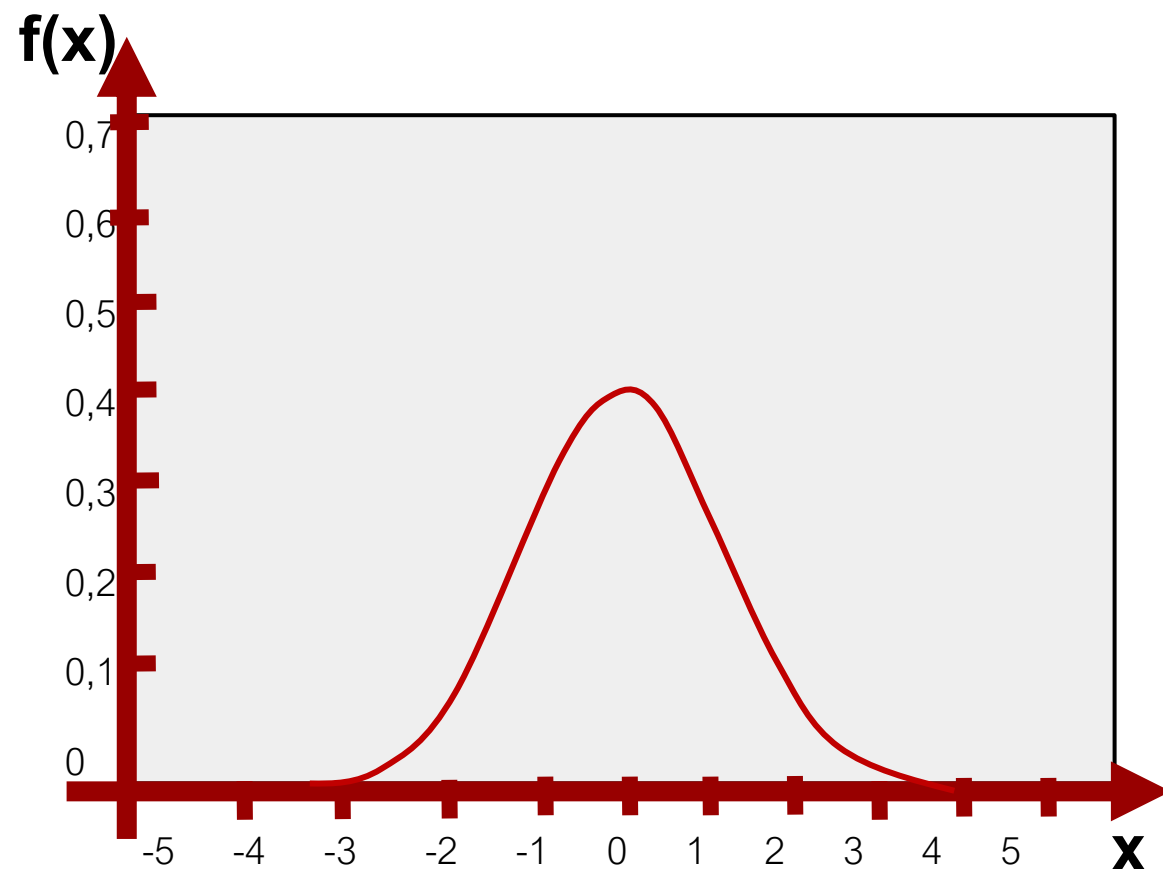
Нормальное распределение

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

$$M[X] = \mu$$

$$D[X] = \sigma^2$$

где параметр μ — математическое ожидание (среднее значение), медиана и мода распределения, а параметр σ — среднеквадратическое отклонение (σ^2 — дисперсия) распределения.



Непрерывное равномерное распределение



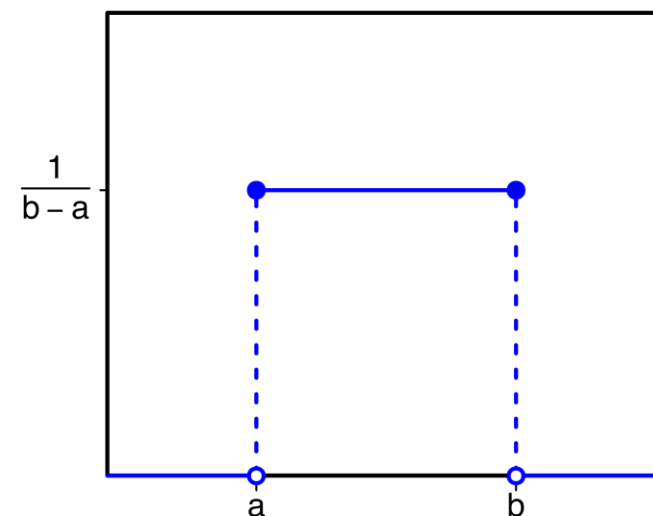
Непрерывное равномерное распределение в теории вероятностей — распределение случайной вещественной величины, принимающей значения, принадлежащие некоторому промежутку конечной длины, характеризующееся тем, что плотность вероятности на этом промежутке почти всюду постоянна.

Непрерывное равномерное распределение

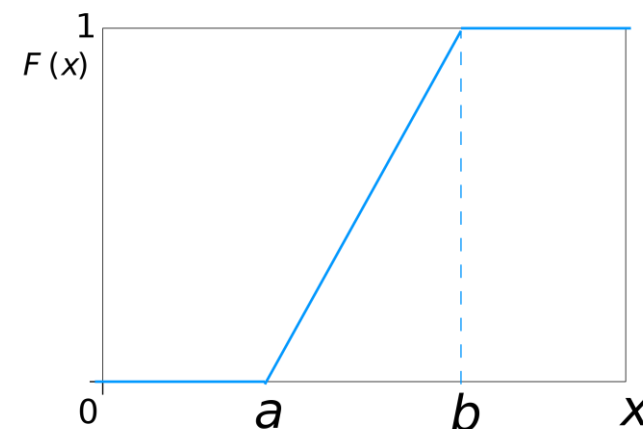
Непрерывное равномерное распределение

Говорят, что случайная величина имеет непрерывное равномерное распределение на отрезке $[a, b]$, где $a, b \in \mathbb{R}$, если её плотность $f_X(x)$ имеет вид:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}$$



Плотность вероятности

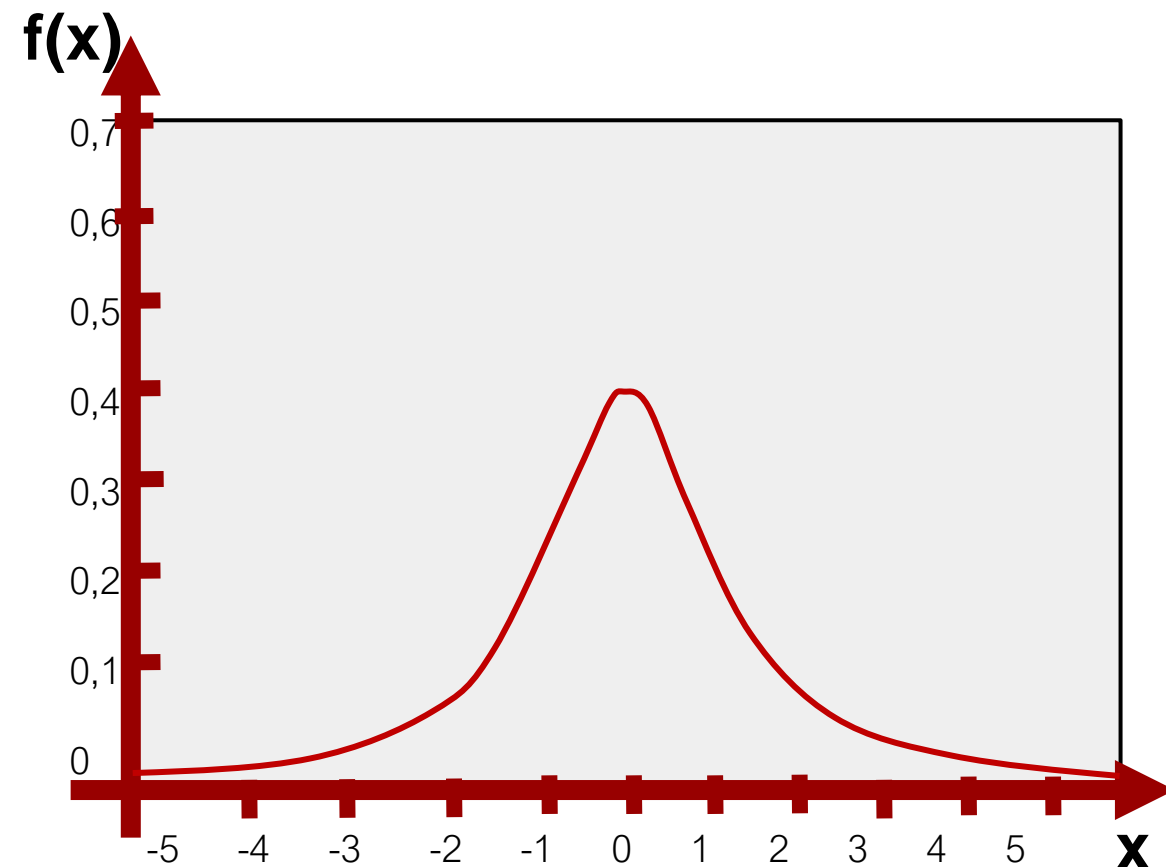


Функция распределения

Распределение Коши

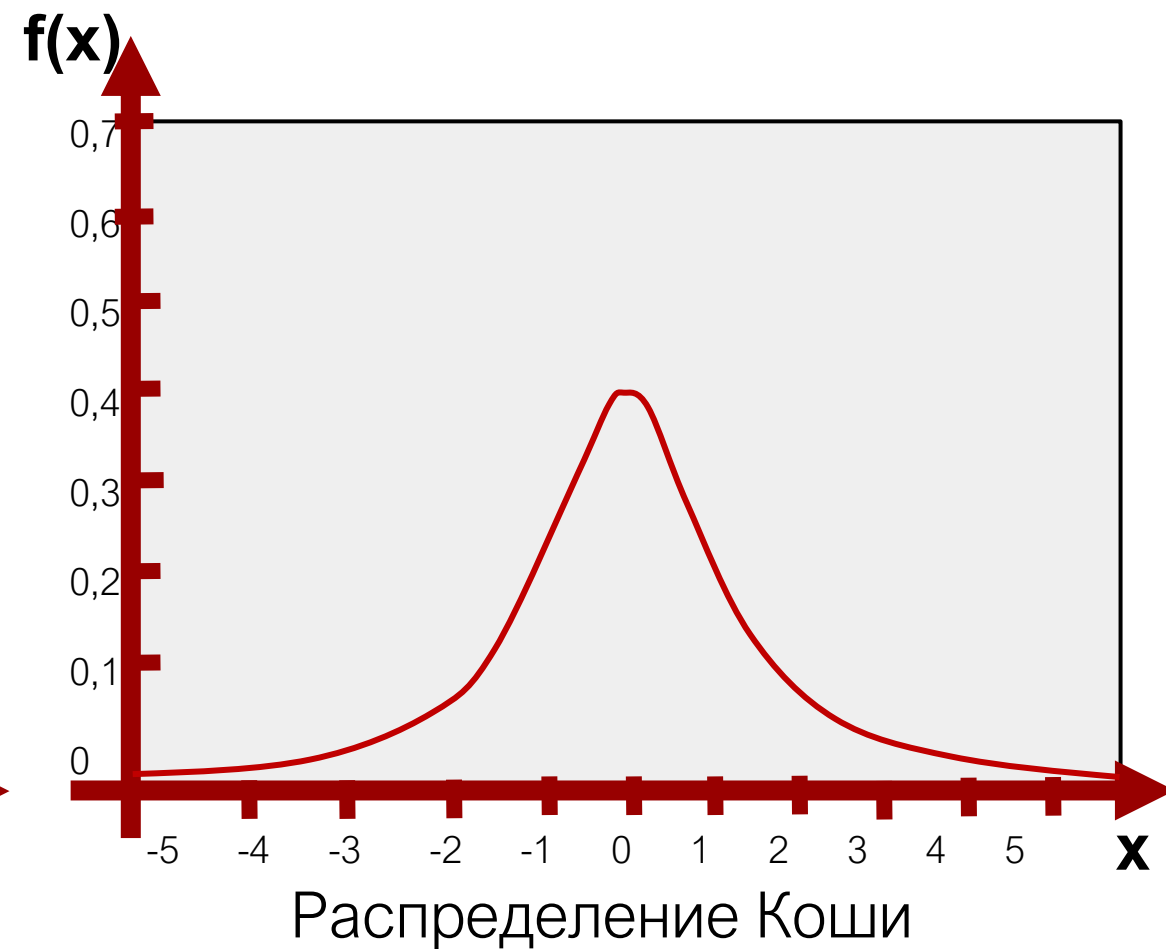
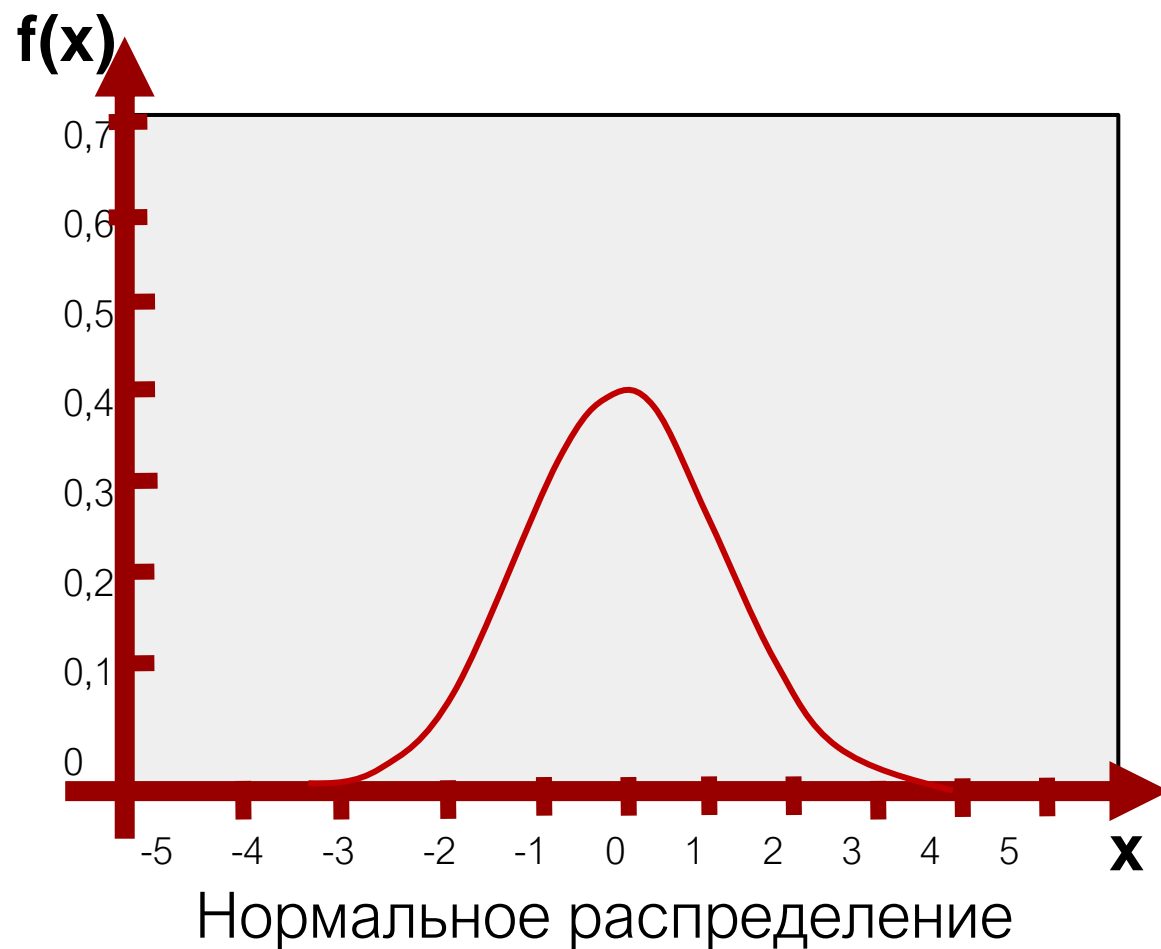
$$\begin{aligned} X_1 &\sim \mathcal{N}(0,1) \\ X_2 &\sim \mathcal{N}(0,1) \end{aligned} \quad X_1/X_2 \sim \mathcal{C}$$

$$f(x) = \frac{1}{\pi} \left[\frac{1}{x^2 + 1} \right]$$



У распределения Коши нет мат. ожидания!

Нормальное распр. и распр. Коши

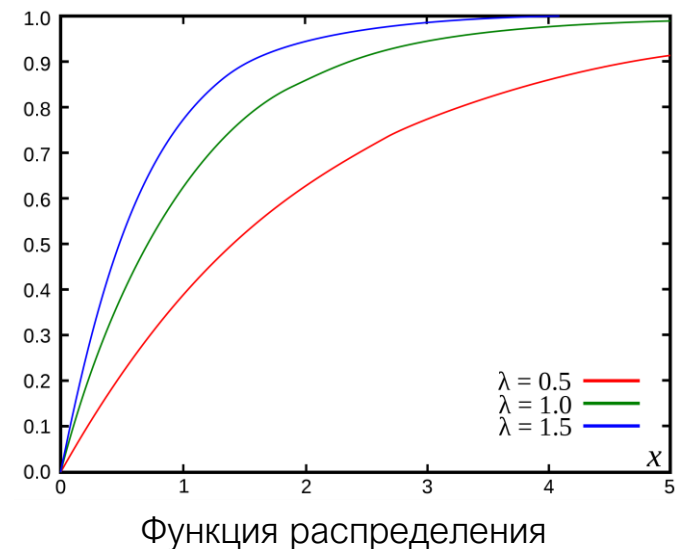
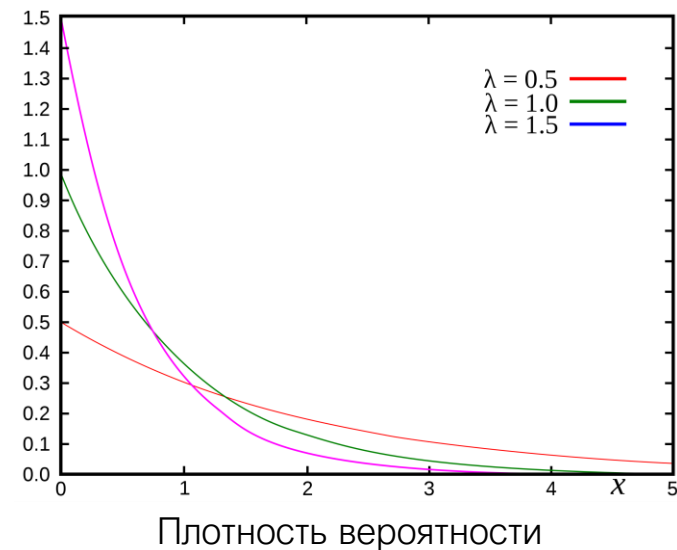


Экспоненциальное распределение

Экспоненциальное (или показательное) распределение — абсолютно непрерывное распределение, моделирующее время между двумя последовательными свершениями одного и того же события.

Случайная величина X имеет экспоненциальное распределение с параметром $\lambda > 0$, если её плотность вероятности имеет вид:


$$f_x(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0 \end{cases}$$



Выборка и реализация

Однородной выборкой (выборкой) объекта n при $n \geq 1$ называется случайный вектор $Z_n \triangleq (X_1, \dots, X_n)$, компоненты которого $X_i, i = \overline{1, n}$, называемые элементами выборки, являются независимыми СВ с одной и той же функцией распределения $F(x)$. Будем говорить, что выборка Z_n соответствует функции распределения $F(x)$.

Реализацией выборки называется неслучайный вектор $Z_n \triangleq (x_1, \dots, x_n)$, компонентами которого являются реализации соответствующих элементов выборки $X_i, i = \overline{1, n}$.



К сожалению,
все регулярно путают
выборку с реализацией

Вариационный ряд

Вариационный ряд

Упорядочим элементы реализации выборки x_1, \dots, x_n по возрастанию $x^{(1)} \leq x^{(2)} \leq \dots \leq x^{(n)}$, где верхний индекс соответствует номеру элемента в упорядоченной последовательности. Обозначим $X^{(k)}, k = \overline{1, n}$, случайные величины, которые при каждой реализации z_n выборки Z_n принимают k -е (по верхнему номеру) значения $x^{(k)}$. Упорядоченную последовательность СВ $X^{(1)} \leq \dots \leq X^{(n)}$ называют **вариационным рядом выборки**.

Элементы $X^{(k)}$ вариационного ряда называют **порядковыми статистиками**, а крайние члены вариационного ряда $X^{(1)}, X^{(n)}$ - **экстремальными порядковыми статистиками**.

ЦПТ и ЗБЧ

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \mathbb{E}(X_1) = \mathbb{E}(X_2) = \dots = \mu$$

ЦПТ: $\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \rightarrow N(0, 1)$

ЗБЧ: $\bar{X}_n \rightarrow \mu$



Задача регрессии

Регрессия

Регрессия - частный случай задачи обучения с учителем, при котором целевая переменная принадлежит бесконечному подмножеству вещественной оси.

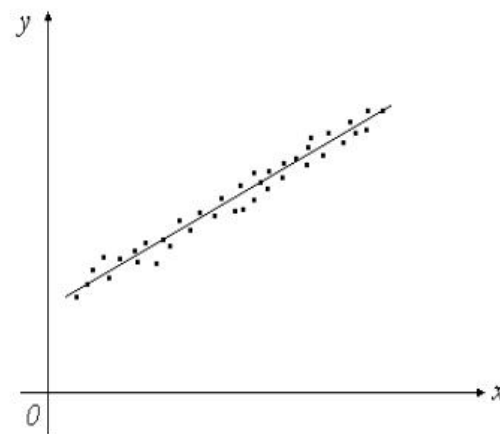


Рисунок 1 – Линейная регрессия

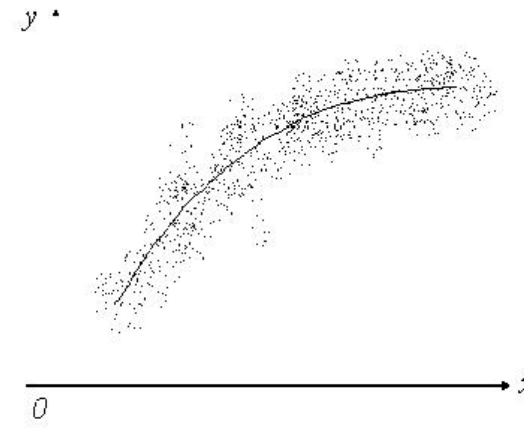


Рисунок 2 – Нелинейная регрессия

8



Регрессия

$y = f(x, \varepsilon) = f(x) + \varepsilon$ – некоторая неизвестная функция, которую мы ХОТИМ ВОССТАНОВИТЬ.

Есть некоторая обучающая реализация выборки (X, Y) размера N .

Нужно выбрать $\hat{f}(x) = \hat{y}$ - (из некоторого класса), так что она будет «лучшей» в смысле некоторой метрики.

Задача регрессии (regression)

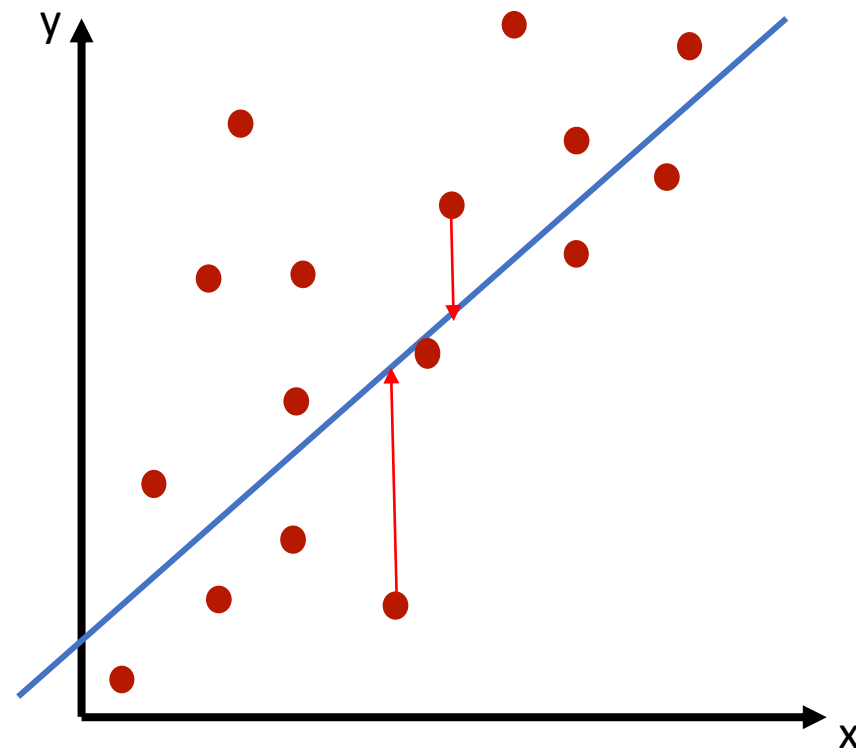
1. MAE
2. RMSE
3. MedAE

—

MAE

Средний модуль отклонения (MAE – Mean Absolute Error или MAD – Mean Absolute Deviation):

$$MAE = \frac{1}{m} \sum_{i=1}^m |\hat{y}_i - y_i|$$



MSE

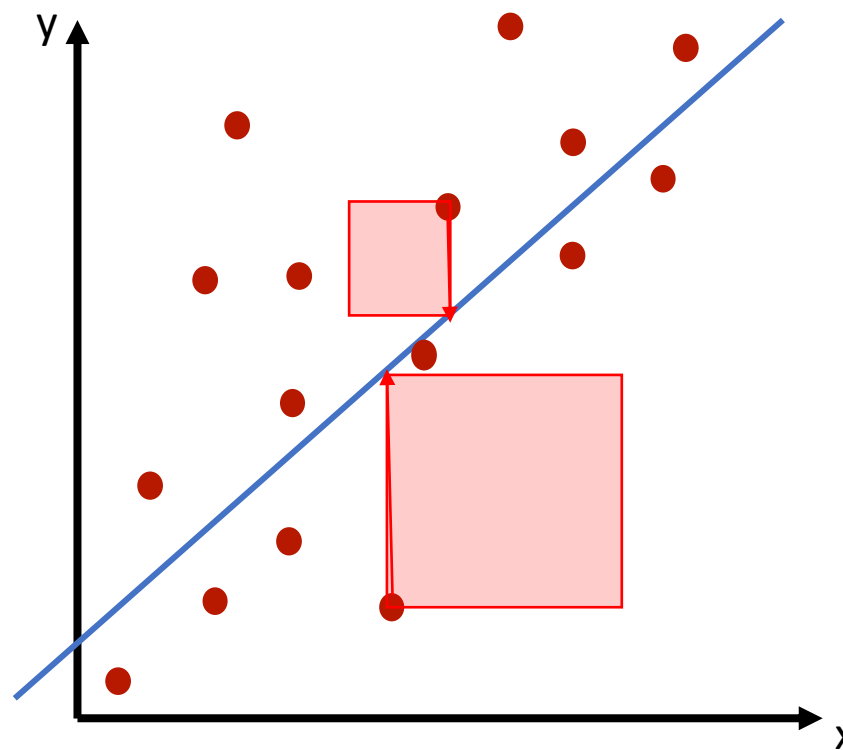
Средний квадрат отклонения (MSE – Mean Squared Error):

$$MSE = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2$$

или корень из этой ошибки:

RMSE – Root Mean Squared Error или
RMSD – Root Mean Square Deviation

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2}$$



MedAE

MedAE – Median Absolute Error

$$\text{MedAE} = \text{median} (|\hat{y}_1 - y_1|, \dots, |\hat{y}_m - y_m|)$$

Метод наименьших квадратов

1. Хотим восстановить зависимость $y = ax + b + \varepsilon$

Метод наименьших квадратов

1. Хотим восстановить зависимость $y = ax + b + \varepsilon$
2. Выпишем функцию ошибки

$$MSE = \sum_{i=1}^n e_i^2 = 1/n \sum_{i=1}^n (y_i - (ax_i + b))^2$$

Метод наименьших квадратов

1. Хотим восстановить зависимость $y = ax + b + \varepsilon$

2. Выпишем функцию ошибки

$$MSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

3. Нужно найти минимум $F(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$ по a и b .

Метод наименьших квадратов

1. Хотим восстановить зависимость $y = ax + b + \varepsilon$
2. Выпишем функцию ошибки

$$MSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

3. Нужно найти минимум $F(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$ по a и b .

$$\begin{cases} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} 2 \sum_{i=1}^n (ax_i^2 + bx_i - x_i y_i) = 0 \\ 2 \sum_{i=1}^n (ax_i + b - y_i) = 0 \end{cases} \Rightarrow \begin{cases} a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a \sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i \end{cases}$$

Метод наименьших квадратов

1. Хотим восстановить зависимость $y = ax + b + \varepsilon$

2. Выпишем функцию ошибки

$$MSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

3. Нужно найти минимум $F(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$ по a и b .

$$\begin{cases} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} 2 \sum_{i=1}^n (ax_i^2 + bx_i - x_i y_i) = 0 \\ 2 \sum_{i=1}^n (ax_i + b - y_i) = 0 \end{cases} \Rightarrow \begin{cases} a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a \sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i \end{cases}$$

4. Дальше нужно решить систему уравнений

Мы про это много чего знаем



- Если шумы нормальные – оптимальная оценка
- Теорема Гаусса — Маркова
- Обобщение на многомерный случай
- Обобщенный МНК

Многомерный случай

$$y_i = x_i^T b + \varepsilon_i$$
$$y = Xb + \varepsilon$$

X - матрица наблюдений факторов (строки матрицы — векторы значений факторов в данном наблюдении, по столбцам — вектор значений данного фактора во всех наблюдениях)

b — вектор коэффициентов, которые нужны найти.

Оптимальная оценка:

$$\hat{b} = (X^T X)^{-1} X^T y$$

Вопросы



- Как оценить важность элементов выборки?
- Что если шум не нормальный?
- Что если не MSE оценка, а MAE?



Поиск экстремума

Задача минимизации функции

$$x^* = \operatorname{argmin}_{x \in X} f(x)$$

Т.е. мы ищем любой $x \in X$, который доставит минимум функции $f(x)$.

Градиентный алгоритм

$$x^* = \operatorname{argmin}_{x \in X} f(x)$$

Методы решения:

- Градиентный метод (+разные модификации)
- Метод внутренней точки (Interior-point method)

$$x_i := x_{i-1} - \rho \nabla f(x_{i-1})$$

Градиентный метод

$$x^* = \operatorname{argmin}_{x \in X} f(x)$$

«Классическая» теорема о сходимости градиентного метода.

Если:

- $f(x)$ дифференцируема и ограничена снизу.
- Выполняется условие Липшица:

$$|f(x) - f(y)| \leq L \cdot |x - y|^\alpha, \quad \alpha \leq 1$$

Тогда: градиентный алгоритм сходится.

Методы случайного поиска

$$x^* = \operatorname{argmin}_{x \in X} f(x)$$

Если $f(x)$ имеет несколько локальных оптимумов.

Методы случайного поиска:

- Рандомизация «классических» алгоритмов
- Алгоритм имитации отжига
- Генетический алгоритм

Вопросы, на которые можно ответить



- Какие вы знаете методы условной и безусловной оптимизации?
- Зачем нужно условие Липшица?
- Можно ли ослабить требования на дифференцируемость?
- Критерий остановки?



Динамическое программирование

Динамическое программирование



Динамическое программирование — это метод, который позволяет решать некоторые задачи комбинаторики, оптимизации, обладающие свойством декомпозиции на подзадачи. Задачи оптимизации, как правило, связаны с задачей максимизации или минимизации той или иной целевой функции.

Динамическое программирование



Метод динамического программирования **сверху** — это простое запоминание результатов решения тех подзадач, которые могут повторно встретиться в дальнейшем.

Динамическое программирование снизу включает в себя переформулирование сложной задачи в виде рекурсивной последовательности более простых подзадач.

Метод Беллмана



Оптимальное поведение обладает тем свойством, что каковы бы ни были начальное состояние и начальное решение, последующие решения должны быть оптимальными относительно состояния, полученного в результате первоначального решения.

Метод Беллмана

Наша задача найти $u_i^* = [u_1^*, \dots, u_N^*]$ – вектор оптимального управления.

x – состояние системы

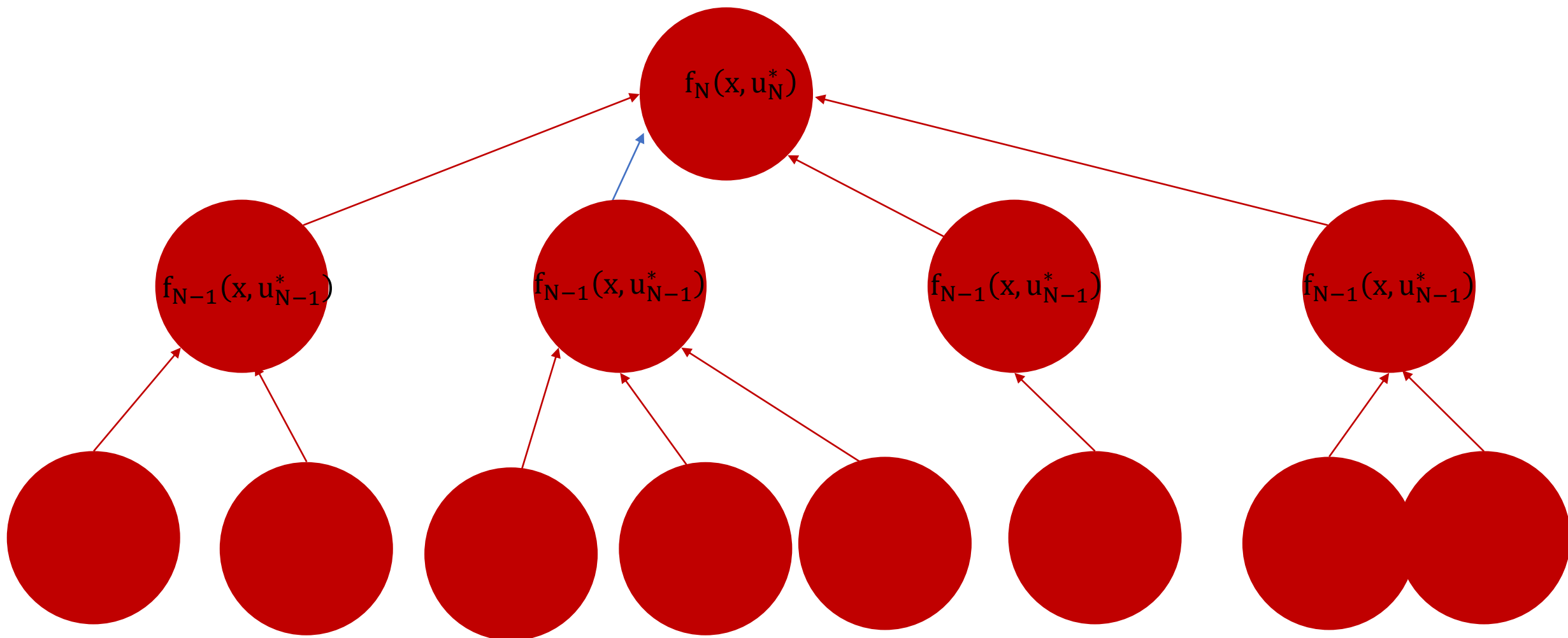
$f_i(x, u)$ – критерий качества на i -ом шаге.

$f_{i+1}(x')$ – оптимальный выигрыш с $i+1$ шага.

Уравнение Беллмана:

$$f_i(x, u_i^*) = f_i(x) = \max_u (f_i(x, u) + f_{i+1}(x'))$$

Метод Беллмана в виде дерева



Что происходит на самом деле?



U_{ij} – матрица (управление, итерация).

Вместо перебора по всем NP – перебор сокращается.

Но не сильно. 😊

Проклятие размерности


Проклятие размерности — проблема, связанная с экспоненциальным возрастанием количества данных из-за увеличения размерности пространства.

Термин «проклятие размерности» был введен Ричардом Беллманом в 1961 году.

Вопросы на которые можно ответить



- Как применить динамическое программирование к выбору оптимального пути?
- Можно ли как-то обойти проклятие размерности?



Планирование эксперимента и Active learning

Неформальная постановка

Когда мы решаем «классическую» задачу регрессии – нам дана реализация выборки (обучающая выборка). И мы считаем, что эта реализация взята из генеральной совокупности без смещения.

А что если бы могли сами выбирать элементы обучающей выборки?

Планирование эксперимента

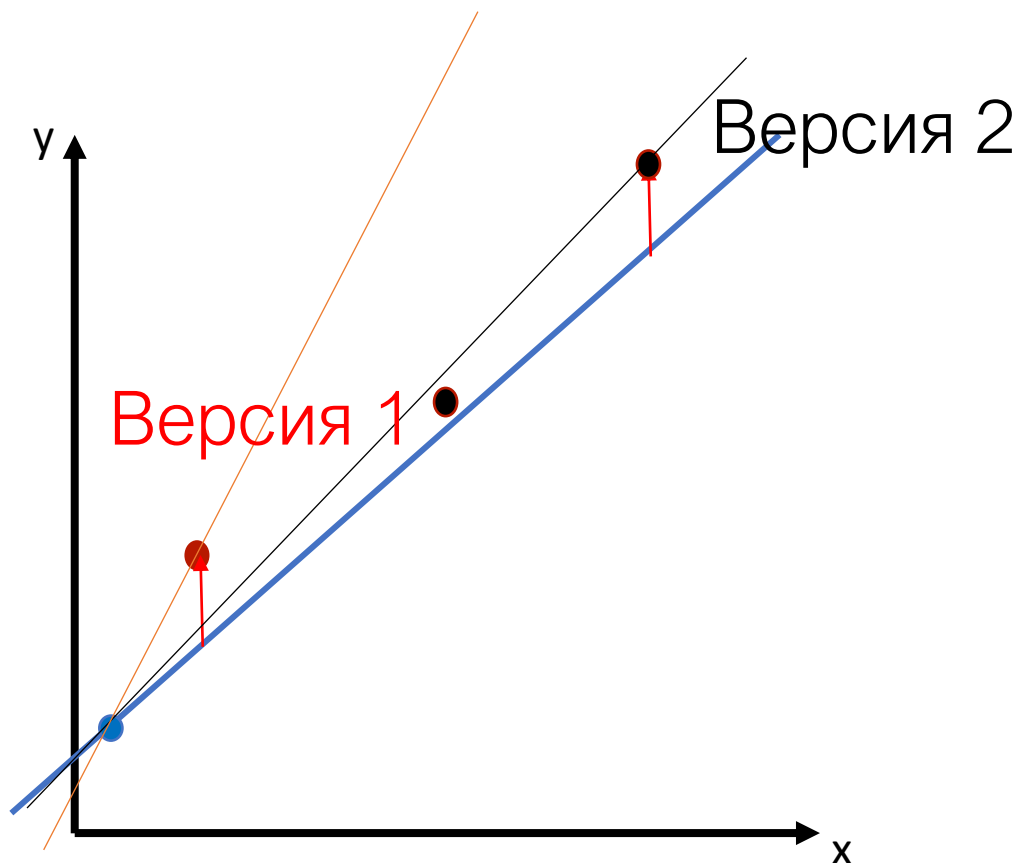


Планирование эксперимента - это процедура выбора числа и условий проведения опытов (физических или расчетных), необходимых и достаточных для решения поставленной задачи с требуемой точностью.

Пример 1.

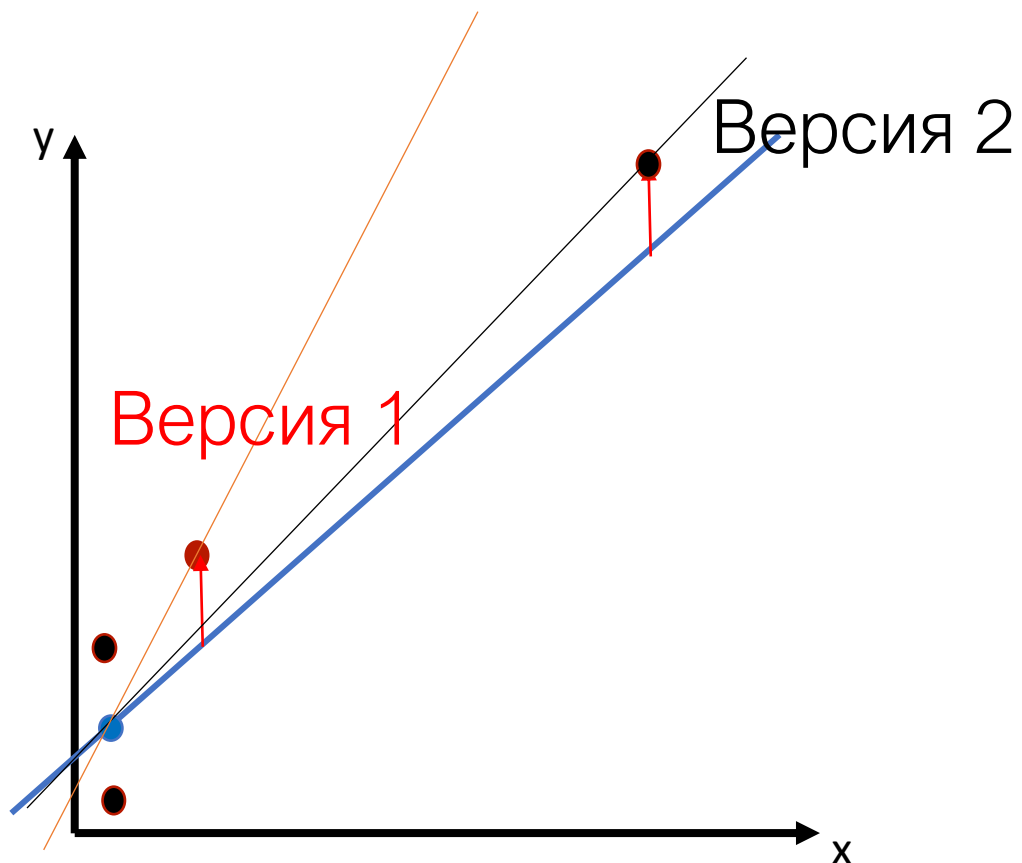
Хотим решить задачу линейной регрессии.
При этом выборка состоит из 2-ух элементов.
При этом шум - н.о.р.с.в.

Пример 1.



Синяя прямая – истинная зависимость. Очевидно, что чем дальше мы возьмем 2-ую точку, тем точнее будет оценка.

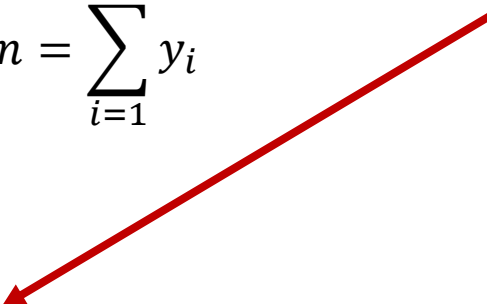
Вопросы



- Как оптимально выбрать 3 точки?
- Есть ли разница, если мы выбираем последовательно или все сразу?

Линейный случай

$$\begin{cases} a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a \sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i \end{cases} \quad \begin{cases} \hat{b} = \frac{n \sum_{t=1}^n x_t y_t - (\sum_{t=1}^n x_t)(\sum_{t=1}^n y_t)}{n \sum_{t=1}^n x_t^2 - (\sum_{t=1}^n x_t)^2}, \\ \hat{a} = \frac{\sum_{t=1}^n y_t - \hat{b} \sum_{t=1}^n x_t}{n}. \end{cases}$$


$$\sum_{i=1}^n (y_i - (\hat{a}x_i + \hat{b}))^2 \rightarrow \min_{x_i \in X}$$

Многомерный случай

$$y_i = x_i^T b + \varepsilon_i \sim N(0, \sigma)$$
$$y = Xb + \varepsilon$$

Оптимальная оценка:

$$\hat{b} = (X^T X)^{-1} X^T y$$

Таким образом:

$$\hat{y} = x^T \hat{b} = x^T (X^T X)^{-1} X^T y$$
$$D[\hat{y}] \approx x^T D[\hat{b}] x = \sigma x^T (X^T X)^{-1} x$$

Далее нужно решить задачу:

$$\min_x \sigma x^T (X^T X)^{-1} x$$

Пример 2

Одномерный дизайн эксперимента

Design	Number of trials N	Values of x								
6.1	3	-1	0	1						
6.2	6	-1	-1	0	0	1	1			
6.3	8	-1	-1	-1	-1	-1	-1	1	1	
6.4	5	-1	-0.5	0	0.5	1				
6.5	7	-1	-1	-0.9	-0.85	-0.8	-0.75	1		
6.6	2	-1	1							
6.7	4	-1	-1	0	1					
6.8	4	-1	0	0	1					

- Модель первого порядка $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Для иллюстрации предположим, что для любого дизайна мы всегда получаем данные, оценивающие $\hat{\beta} = (16, 7.5)$

$$\hat{y}(x) = 16 + 7.5x$$

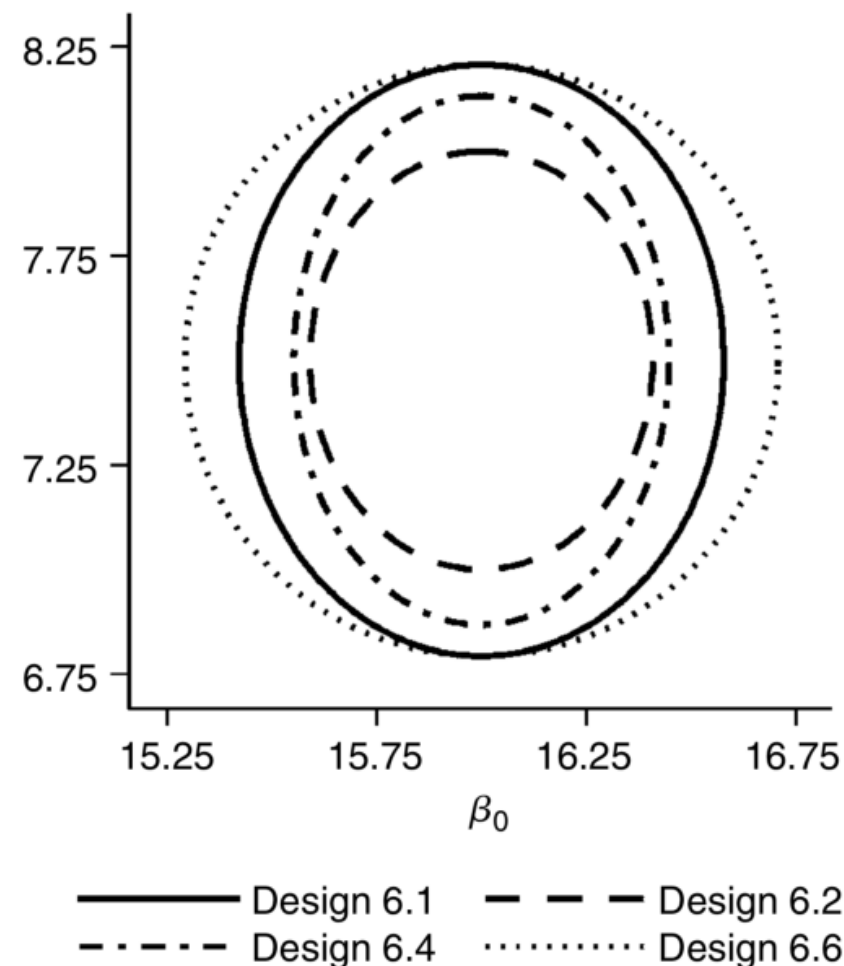
Пример 2

- Обозначьте доверительный эллипс $(\beta - \hat{\beta})X^T X(\beta - \hat{\beta})=1$
- $\hat{\beta} = (16, 7.5)$ обозначение Гауссовского

Design	Number of trials N	Values of x							
6.1	3	-1	0	1					
6.2	6	-1	-1	0	0	1	1		
6.3	8	-1	-1	-1	-1	-1	-1	1	1
6.4	5	-1	-0.5	0	0.5	1			
6.5	7	-1	-1	-0.9	-0.85	-0.8	-0.75	1	
6.6	2	-1	1						
6.7	4	-1	-1	0	1				
6.8	4	-1	0	0	1				

пример для дизайна 6.1

$$X = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \quad (X^T X)^{-1} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$



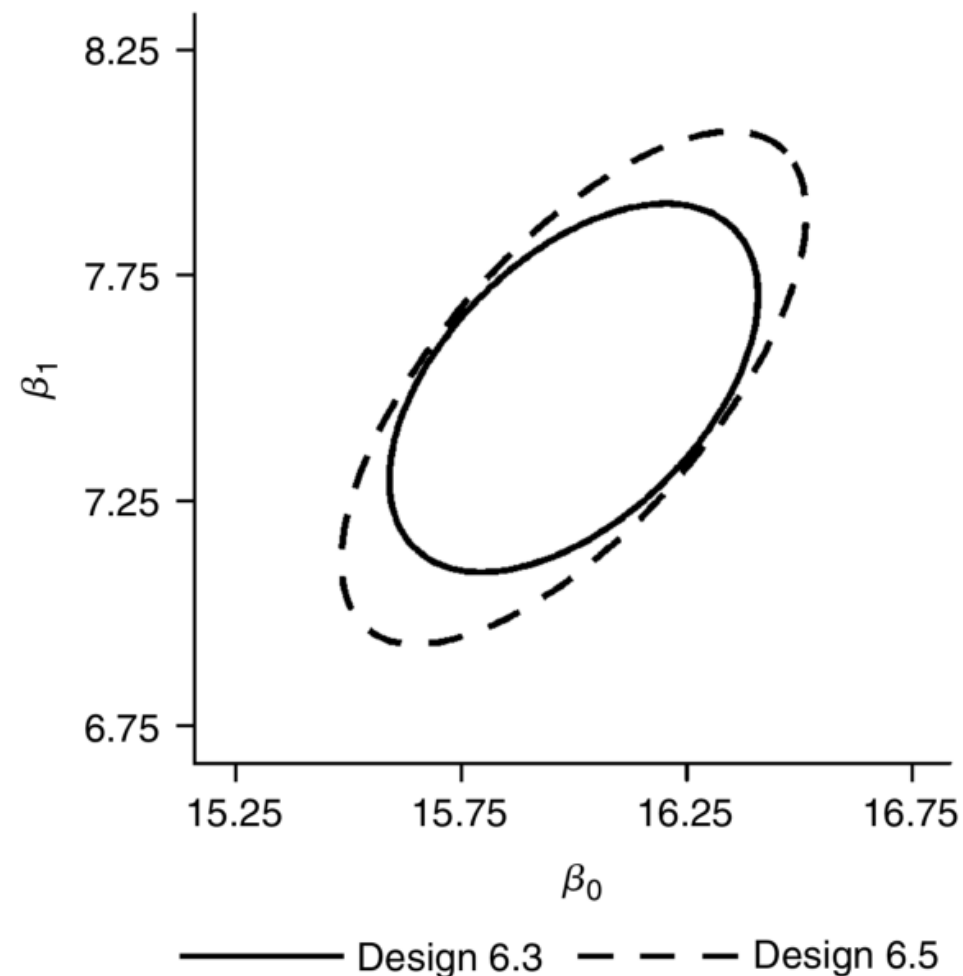
Пример 2

- Обозначьте доверительный эллипс $(\beta - \hat{\beta})X^T X(\beta - \hat{\beta}) = 1$
- $\hat{\beta} = (16, 7.5)$ обозначение Гауссовского

Design	Number of trials N	Values of x							
6.1	3	-1	0	1					
6.2	6	-1	-1	0	0	1	1		
6.3	8	-1	-1	-1	-1	-1	-1	1	1
6.4	5	-1	-0.5	0	0.5	1			
6.5	7	-1	-1	-0.9	-0.85	-0.8	-0.75	1	
6.6	2	-1	1						
6.7	4	-1	-1	0	1				
6.8	4	-1	0	0	1				

пример для дизайна 6.3

$$X = \begin{bmatrix} 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad (X^T X)^{-1} = \begin{bmatrix} \frac{1}{6} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{6} \end{bmatrix}$$



Active learning



Active Learning – класс задач машинного обучения, где выборка не является случайной, а может быть выбрана по некоторым правилам.

Постановка

$$\sum_{i=1}^n C_i \cdot L(b; x_i, y_i) \rightarrow \min_b$$

L – это некоторый функционал качества.

Например MSE.

$C_i = C_i(x_i, y_i)$ - стоимость получения пары (x_i, y_i) .

В общем случае стоимость может быть случайной.

Стратегии (теория)

Отбор объектов из выборки (pool-based active learning).

Имеется некоторая выборка, и алгоритм использует объекты (X) из нее в качестве запросов к оракулу, чтобы получить разметку (Y). В данной стратегии каждому объекту присваивается степень информативности — сколько выгоды принесет информация об истинной метке объекта, и оракулу отправляются самые информативные объекты.

Стратегии (теория)

Отбор объектов из потока (selective sampling)

Алгоритм пользуется не статической выборкой, а потоком данных, и для каждого объекта из потока принимается решение, запрашивать оракула на этом объекте или нет. В случае, если принято решение запросить оракула, объект и его метка используются в дальнейшем обучении модели, в противном случае объект просто отбрасывается. В отличие от отбора объектов из выборки отбор из потока не строит никаких предположений насчет плотности распределения объектов, не хранит сами объекты и работает значительно быстрее.

Стратегии (теория)

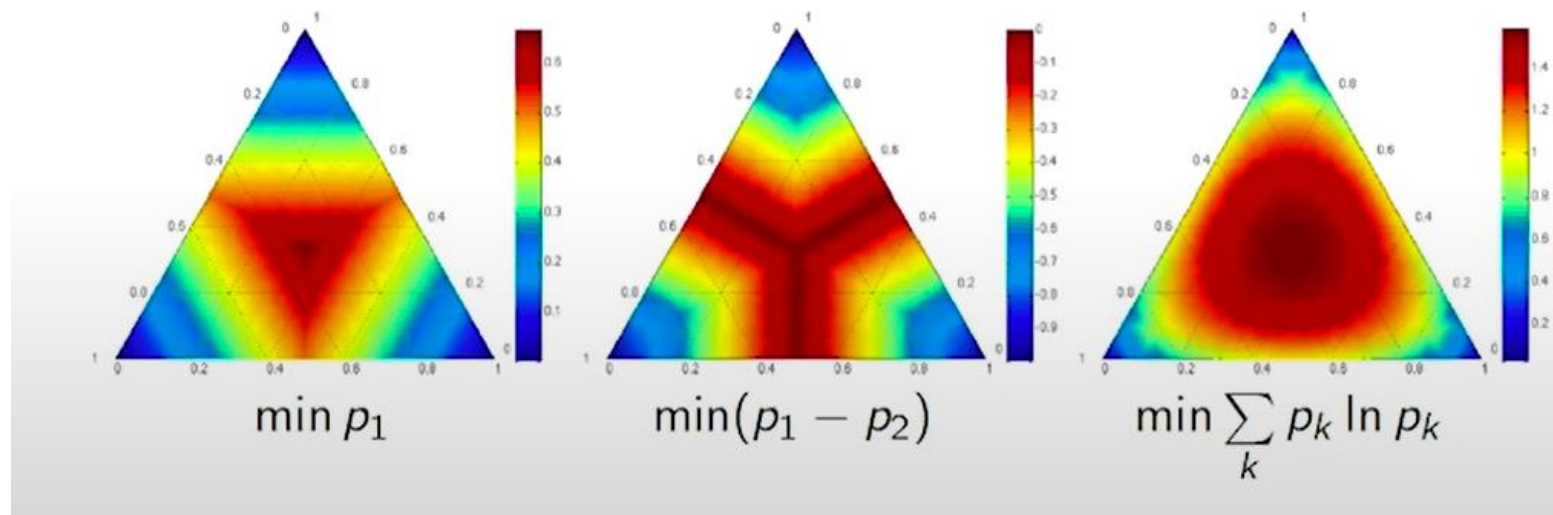
Синтез объектов (query synthesis)

Вместо использования заранее заданных объектов, алгоритм сам выбирает объекты (X) и подает их на вход оракулу (Y). Например, если объекты — это вектора в n -мерном пространстве, разделенные гиперплоскостью и решается задача бинарной классификации, имеет смысл давать оракулу на вход синтезированные вектора, близкие к границе.

Сэмплирование по неувренности

Основная идея: брать объект, с наибольшей неопределенностью.
В качестве критерия:

- Наименьшей достоверности
- Наименьшей разности отступов
- Максимуму энтропии



Сэмплирование по неувверенности

Аналог для регрессии:

1. Переформулировать в виде многоклассовой классификации. И см. выше.
2. Брать объект с самым широким доверительным интервалом. Легкая проблема: нужно строить доверительные интервалы для каждой точки (не для параметров модели)

Несогласие в комитете



Идея: у нас есть R алгоритмов, выбираем x_i – где модели между собой максимально расходятся.

Что значит максимально расходятся?

- 1) Максимальная выборочная дисперсия
- 2) Максимальный размах (+интерквантильный)
- 3) ...

Ожидаемое изменение модели

Идея: выбираем x_i – который бы привел к наибольшему изменению параметров модели.

Если мы оптимизируем (стох.) градиентом методом:

$$b := b - \alpha \frac{\partial L(b; x_i, y_i)}{\partial b}$$

То $\frac{\partial L(b; x_i, y_i)}{\partial b}$ и есть величина изменения. Нужно найти ее максимум по x_i .

Ожидание сокращения ошибки



Идея: выбираем объект, который после добавления даст наиболее качественную классификацию (регрессию) по неразмеченным данным.

Сокращение дисперсии

Идея: выбирать объекты, которые максимально сократят дисперсию.

$$MSE(b) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i, b))^2$$

$$MSE(\hat{b}) = \min_b (MSE(b))$$

$$D[MSE(\hat{b})] \approx \left(\frac{\partial f(x)}{\partial b} \right)^T \left(\frac{\partial MSE(b)}{\partial b^2} \right)^{-1} \left(\frac{\partial f(x)}{\partial b} \right)$$

$$x = \arg \min_x D[MSE(\hat{b})]$$

Сокращение дисперсии

A-optimal: $\min \text{tr}(D[MSE(\hat{b})])$

D-optimal: $\min \det(D[MSE(\hat{b})])$

E-optimal: $\min \max_{\lambda} (D[MSE(\hat{b})])$ (по собственным значениям)

L-optimal: $\min \text{tr}$ от линейной комбинации объектов матрицы

Полезные ссылки



- <http://active-learning.net/>
- https://github.com/SimiPixel/pool_based_active_learning
- Федоров В.В. (1971) Теория оптимального эксперимента (планирование регрессионных экспериментов)

Весь наш курс мы будем
пытаться скрестить:
Оптимальное планирование
эксперимента
с поиском оптимума