# ОСНОВНЫЕ МАТЕМАТИЧЕСКИЕ МЕТОДЫ

# МАТЕМАТИКА В ИИ

| Дисциплина | Важность |
|---|---|
| Линейная алгебра | 35% |
| Теория вероятности и мат статистика | 25% |
| Многомерный анализ | 15% |
| Алгоритмы и их сложность | 15% |
| Остальное | 15% |

https://habr.com/ru/company/otus/blog/525636/

# АЛГЕБРА

## Экспоненты

- Экспонента — **показательная функция exp(x)**

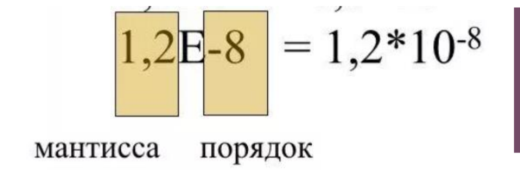
## Радикалы

- Результат извлечения корня

## Факториалы

## Суммирование

## Экспоненциальная запись




https://techrocks.ru/2021/12/23/math-concepts-you-need-for-ai/

# ЛИНЕЙНАЯ АЛГЕБРА

- это **систематизированное представление знаний**, которое может понять компьютер, и все операции в линейной алгебре являются систематизированными правилами.
- Матричные вычисления

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 18 \\ 27 \end{bmatrix}$$

# ЛИНЕЙНАЯ АЛГЕБРА

## Скаляр

- одиночное число (действительное или натуральное)

## Вектор

- список чисел, расставленных по порядку. Их можно представить как точки в пространстве

## Матрица

- двумерный массив чисел, где каждое число идентифицируется двумя индексами

## Тензор

- N-мерный массив (N>2) чисел, выстроенных в обычной сетке с N-осями.

## Собственный вектор и собственное значение

## Сингулярное выражение

## Метод главных компонент

- Способ уменьшения размерности

## Векторное и скалярное произведения

# СОБСТВЕННЫЙ ВЕКТОР

**Определение**: ненулевой вектор $\bar{u}$, который при умножении на некоторую квадратную матрицу $A$ превращается в самого же себя с числовым коэффициентом $\lambda$, называется **собственным вектором** матрицы $A$. Число $\lambda$ называют **собственным значением** или **собственным числом** данной матрицы.

$$A\bar{u} = \begin{pmatrix} -1 & -6 \\ 2 & 6 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \cdot 2 - 6 \cdot (-1) \\ 2 \cdot 2 + 6 \cdot (-1) \end{pmatrix} = \begin{pmatrix} 4 \\ -2 \end{pmatrix} = 2 \cdot \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

http://www.mathprofi.ru/sobstvennye_znachenija_i_sobstvennye_vektory.html

# СИНГУЛЯРНОЕ ВЫРАЖЕНИЕ

**Теорема 1** (Дж. Форсайт). *Для любой вещественной $(n \times n)$-матрицы $A$ существуют две вещественные ортогональные $(n \times n)$-матрицы $U$ и $V$ такие, что*

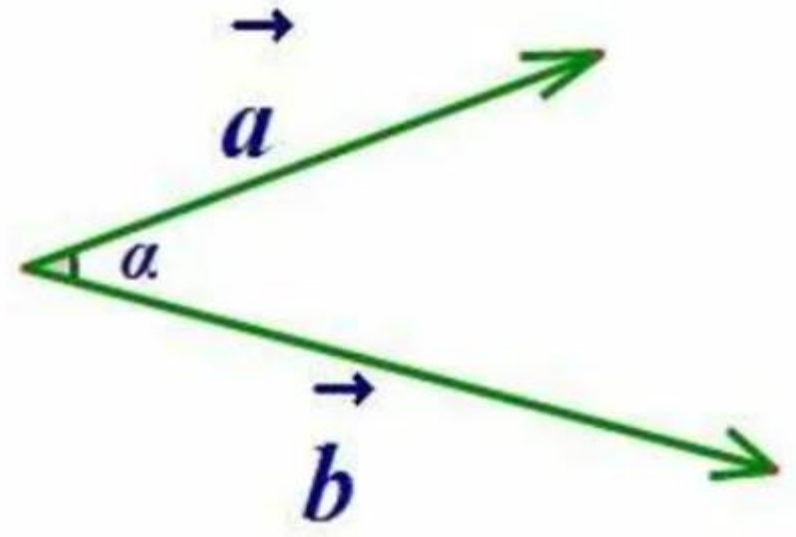
$$U^T AV = \Lambda.$$

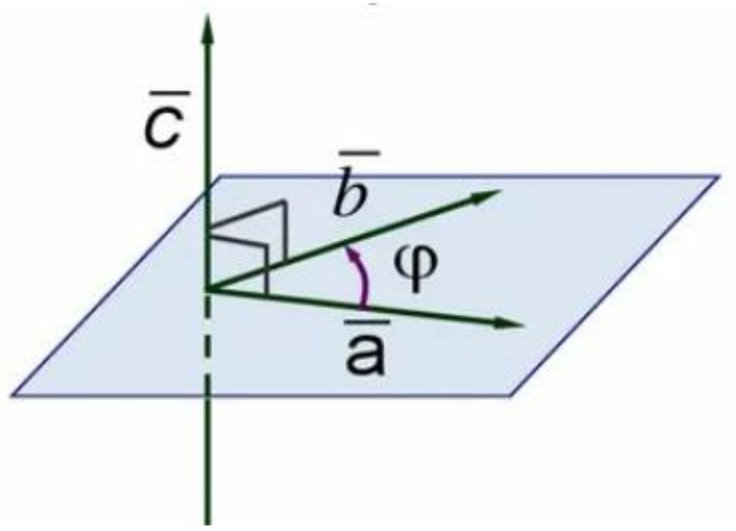
Столбцы матриц U и V называются соответственно левыми и правыми сингулярными векторами, а значения диагонали матрицы Λ называются сингулярными значениями.

Ортогональная матрица – матрица, у которой обратная матрица равна транспонированной матрице.

# СКАЛЯРНОЕ И ВЕКТОРНОЕ ПРОИЗВЕДЕНИЕ ВЕКТОРОВ

$$\vec{a} \cdot \vec{b} = |\vec{a}| \cdot |\vec{b}| cos\alpha$$




$$|\bar{c}| = |\bar{a}| \cdot |\bar{b}| \cdot \sin(\bar{a}; \bar{b}).$$

$$\bar{c} \perp \bar{a}; \quad \bar{c} \perp \bar{b}$$

# ВЕРОЯТНОСТНАЯ ИНТЕРПРЕТАЦИЯ МАШИННОГО ОБУЧЕНИЯ

- Машинное обучение — это наука о том, как на основании данных делать выводы, откуда эти данные взялись, и предсказания, какие данные встретятся нам в будущем.
- Невозможен точный результат: доля неопределенности выводов, оценка точности предсказаний

# ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТИ

- Дискретные случайные величины: конечный набор исходов, сумма вероятностей = 1
- Непрерывные случайные величины: набор исходов – вещественная прямая, вероятность = производная от функции распределения

$$F(a) = p(x < a) \qquad p(x) = \frac{\mathrm{d}F}{\mathrm{d}x}; \qquad \int_{-\infty}^{\infty} p(x)\mathrm{d}x = F(\infty) - F(-\infty) = 1.$$

- Совместная вероятность - это вероятность одновременного наступления двух событий, p(x, y).

$$p(x, y) = p(x)p(y)$$

# ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТИ

- Условная вероятность - вероятность наступления одного события, если известно, что произошло другое

$$p(x \mid y) = \frac{p(x, y)}{p(y)}. \qquad p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x)$$

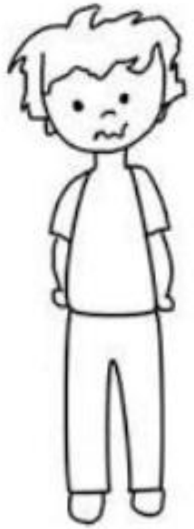
- **Теорема Байеса**

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

позволяет переоценивать наши априорные представления о мире (p(y)) на основе частичной информации (данных), которую мы получили в виде наблюдений (p(x|y)), в качестве вывода получая новое состояние наших представлений p(y|x).

# ПРИМЕР

Мы думаем, что он интроверт



Боб

P(A) – Боб редко любит заводить новых друзей (априорная вероятность)

Эрик - экстраверт

P(B) – Вероятность того, что Эрик дружелюбен

Likelihood

Prior

Posterior

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Normalizing constant

P(A) Предположение: Боб - типичный интроверт! →
P(B|A)P(A) Сведения: Боб завел нового друга! →
P(A|B) Обновленное предположение: В Бобе что-то есть от интроверта!

***Мы меняем наше предположение*** о Бобе - он ***не такой уж и интроверт***.

# ПРИМЕР

- Тест на болезнь имеет вероятность успеха = 95%
- Вероятность ошибки первого рода (ложного срабатывания) = 5%
- Вероятность ошибки  второго рода (пропуск больного человека) = 5%
- Распространение болезни – у 1% респондентов

Получил положительный результат на тест

**С какой вероятностью он действительно болен?**

# ПРИМЕР

Пусть: t – результат теста, d – наличие болезни

$$p(t=1) = p(t=1 \mid d=1)p(d=1) + p(t=1 \mid d=0)p(d=0).$$

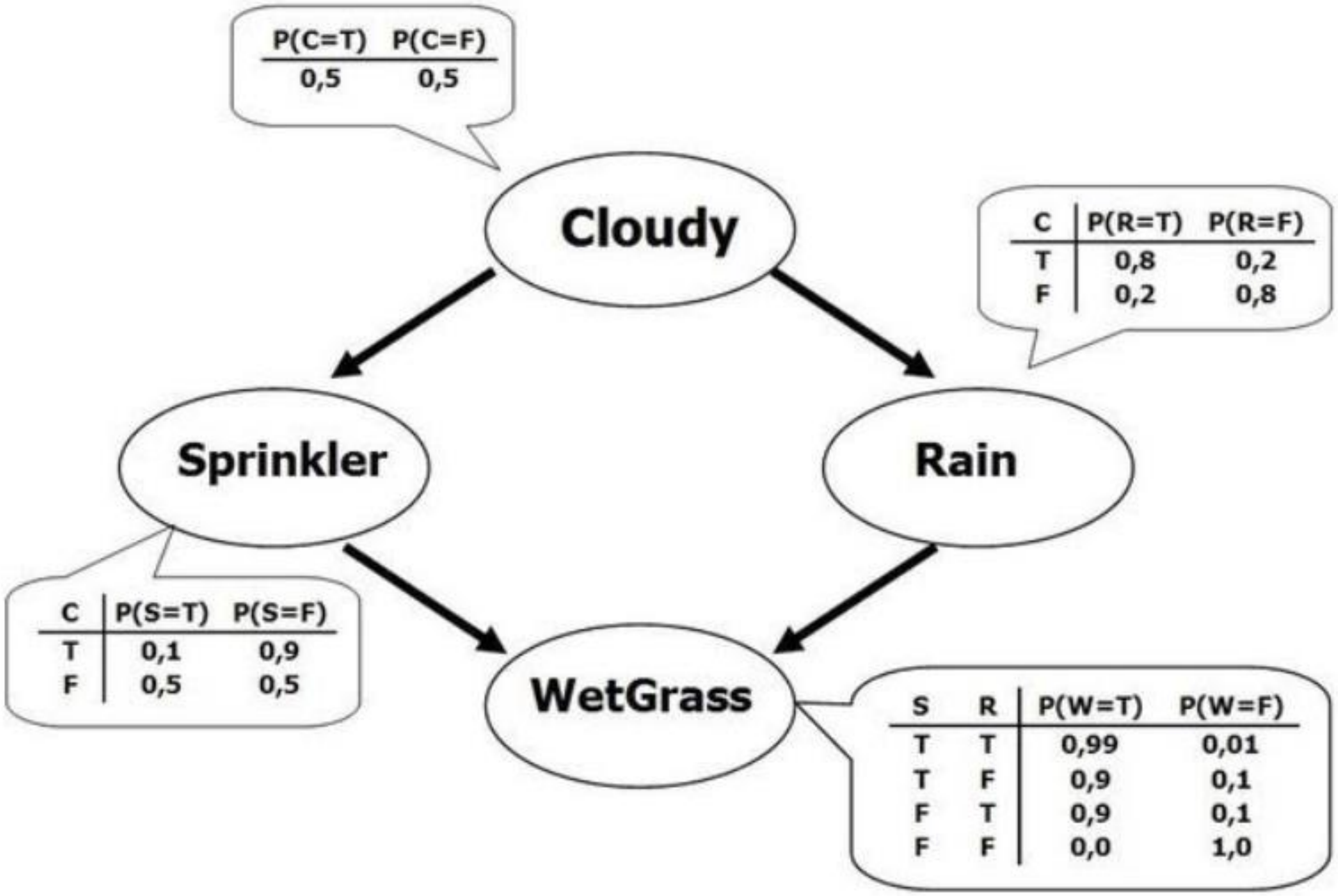
По теореме Байеса

$$p(d=1 \mid t=1) = \frac{p(t=1 \mid d=1)p(d=1)}{p(t=1 \mid d=1)p(d=1) + p(t=1 \mid d=0)p(d=0)} =$$

$$= \frac{0{,}95 \times 0{,}01}{0{,}95 \times 0{,}01 + 0{,}05 \times 0{,}99} = 0{,}16.$$

**Вероятность действительно оказаться больным — всего 16 %!**

# БАЙЕСОВСКИЕ СЕТИ (ВЕРОЯТНОСТНЫЕ ГРАФИЧЕСКИЕ МОДЕЛИ)



# СТАТИСТИКА

***Машинное обучение** - это не только построение прогностических моделей, но и **извлечение как можно большего объема информации** из данных с помощью доступных нам статистических инструментов.*

**Базовая статистика**

- среднее значение, медиана, мода, дисперсия, ковариация и т. д.

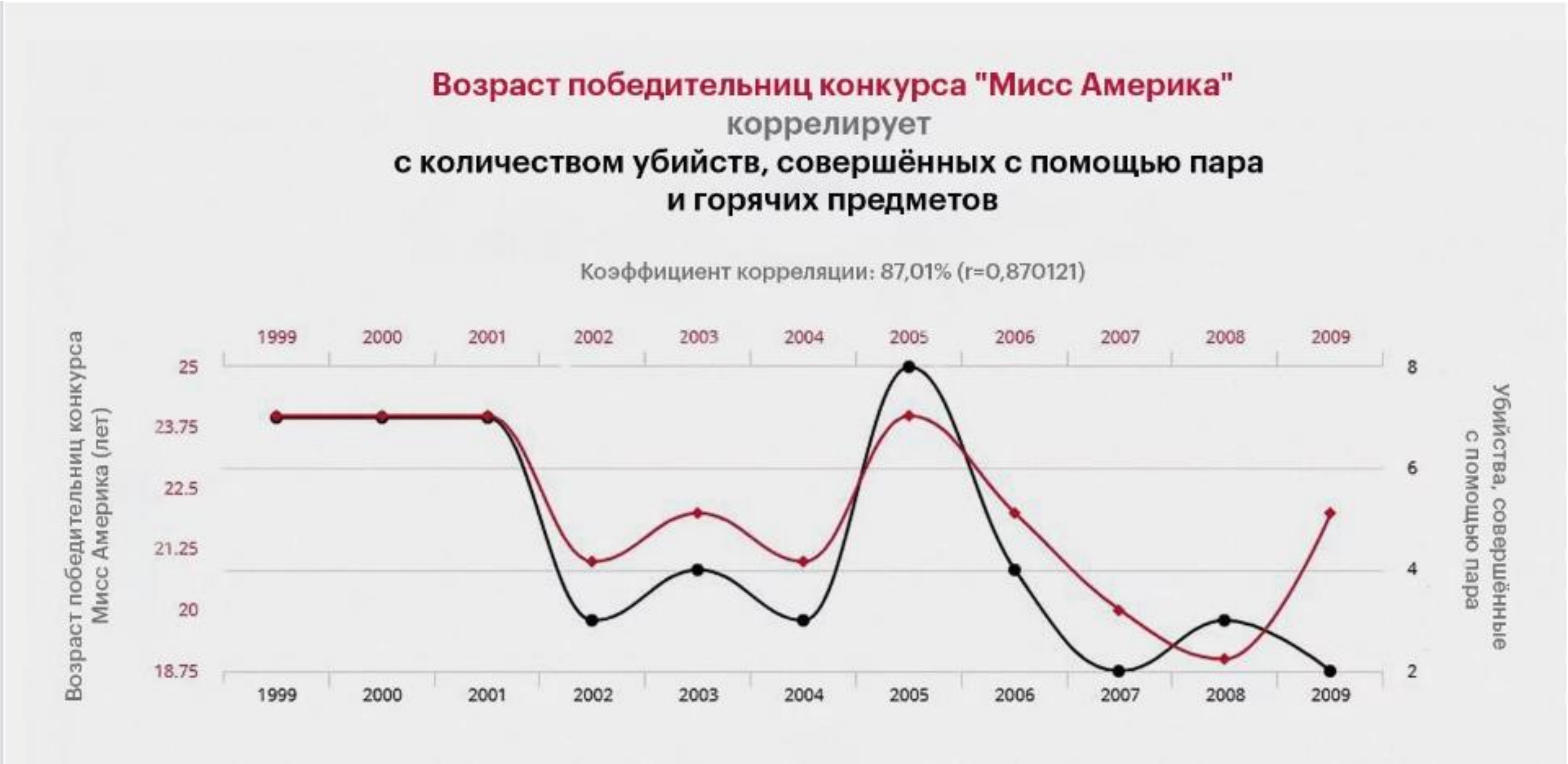
**Основные распределения**

- биномиальное, пуассоновское, бернуллиевское, гауссовское, экспоненциальное

# БАЗОВАЯ СТАТИСТИКА

- **Случайная величина** – это переменная, значения которой определяются случайным экспериментом (дискретная и  непрерывная).
- **Среднее**: вычисляется как среднее арифметическое.
- **Медиана**: если выстроить все данные по возрастанию и найти середину этого ряда, это будет медиана. Одна половина из значений данных будет больше медианы, а другая — меньше.
- **Мода**: значение в наборе данных, которое встречается чаще всего.
- **Математическое ожидание**:  среднее (взвешенное по вероятностям возможных значений) значение случайной величины. Среднеожидаемое значение при многократном повторении испытаний.
- **Дисперсия** — это величина, показывающая, как именно и насколько сильно разбросаны значения. Мера вариации между значениями.
- **Корреляция** - изменения одной величины сопутствуют изменениям другой, оценка тесноты связи между переменными (**не означает причинно-следственную связь**)

https://skillbox.ru/media/code/7-bazovykh-statisticheskikh-ponyatiy-neobkhodimykh-datasayentistu/



Возраст победительниц конкурса "Мисс Америка"
коррелирует
с количеством убийств, совершённых с помощью пара
и горячих предметов
Коэффициент корреляции: 87,01% (r=0,870121)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
25
23.75
22.5
21.25
20
18.75
8
6
4
2
Возраст победительниц конкурса
Мисс Америка (лет)
Убийства, совершённые
с помощью пара

# МОДЕЛЬ ПРАКТИЧЕСКИХ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Задача состоит в том, чтобы по данным D подобрать описывающие их параметры θ наилучшим образом. В классической статистике для этого обычно ищут гипотезу максимального правдоподобия (maximum likelihood, ML):

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta})$$

- апостериорное распределение (posterior): $p(\boldsymbol{\theta} \mid D) \propto p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$ (пропорционально)
- максимальная апостериорная гипотеза: $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid D) = \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$

# АНАЛИЗ

## Производные

- Частные производны, правила взятия производных

## Векторное/матричное исчисление

- Дифференциальные операторы производных

## Градиентные алгоритмы

- локальный/глобальный максимум и минимум, седловые точки, выпуклые функции

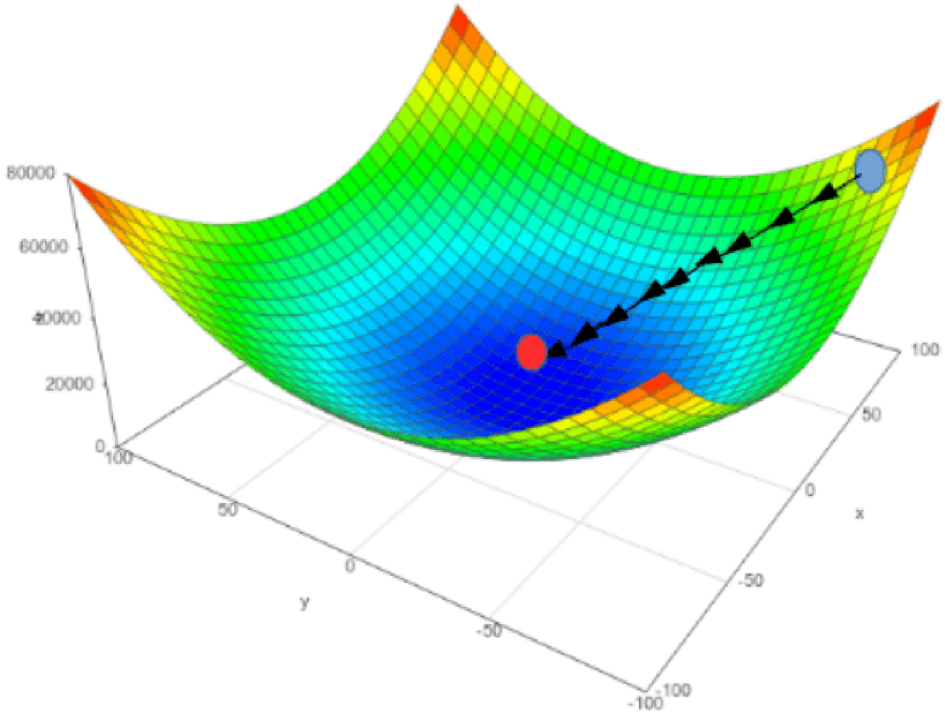
# ПРОИЗВОДНЫЕ

- Производная показывает отношение приращении функции к приращению переменной (скорость).

$$f'(x) = y'(x) = \frac{dy}{dx}$$

- Точки, в которых функция перестаёт изменять значение своего приращения называются точками локального экстремума (производная равна нулю)
- Частные производные 1-го порядка напоминают **«обычную» производную**
- $z'_x, z'_y$ - это функции, которые определяют скорость изменения функции $z=f(x,y)$ в направлении осей $OX$ и $OY$ соответственно

# МЕТОД ГРАДИЕНТНОГО СПУСКА
## (МЕТОД ОПТИМИЗАЦИИ)




- Поверхность – это значение функции ошибки E(θ) от весов модели, которая обучается.
- Шарик – текущее значение параметров
- Найти минимум функции ошибки (шарик катится вниз)
- Найти направление, в котором будет скатываться шарик
- *Многомерное исчисление, или частная производная, используются для **математической оптимизации заданной функции***

# МЕТОД ГРАДИЕНТНОГО СПУСКА

- Градиент поверхности - это направление, в котором функция быстрее всего возрастает.

$$\nabla_{\boldsymbol{\theta}} E = \begin{pmatrix} \frac{\partial E}{\partial \theta_1} \\ \vdots \\ \frac{\partial E}{\partial \theta_{n-1}} \\ \frac{\partial E}{\partial \theta_n} \end{pmatrix}$$

- Шарик должен катиться по направлению обратному градиенту $-\nabla_\theta E$.
- Дискретизация времени у непрерывного процесса катящегося шарика

# МОДЕЛЬ ГРАДИЕНТНОГО СПУСКА

- $\theta_t$ вектор параметров модели на шаге t,
- E — минимизируемая функция,
- Вектор обновления параметров на шаге t

$$\boldsymbol{u}_t = -\eta \nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta}_{t-1}), \qquad \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{u}_t.$$

# ЛИНЕЙНАЯ РЕГРЕССИЯ

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \boldsymbol{x}^\top \boldsymbol{w}$$

$$\hat{y}(\boldsymbol{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \boldsymbol{x}^\top \hat{\boldsymbol{w}}.$$

Метод наименьших квадратов для оценки ошибки

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2.$$

# ПРИМЕР ЛИНЕЙНОЙ РЕГРЕССИИ

x — рост песика, y — его вес.

1. Чем крупнее песик, тем больший вес он имеет;
2. Песики одинакового роста могут иметь разный вес.

Выводы:

1. Для фиксированного роста песика x его вес y = f (x) является случайной величиной;
2. В среднем вес f (x) возрастает при увеличении роста песика x.

$$y = \theta_0 + \theta_1 x + \varepsilon,$$

Неизвестные параметры и случайная составляющая

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 + \varepsilon,$$

$x_1$ — рост песика,

$x_2$ — обхват туловища песика,

$y$ — вес песика,

$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры,

$\varepsilon$ — случайная составляющая с нулевым средним.

Зависимость линейна по параметрам, квадратична по аргументам.

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

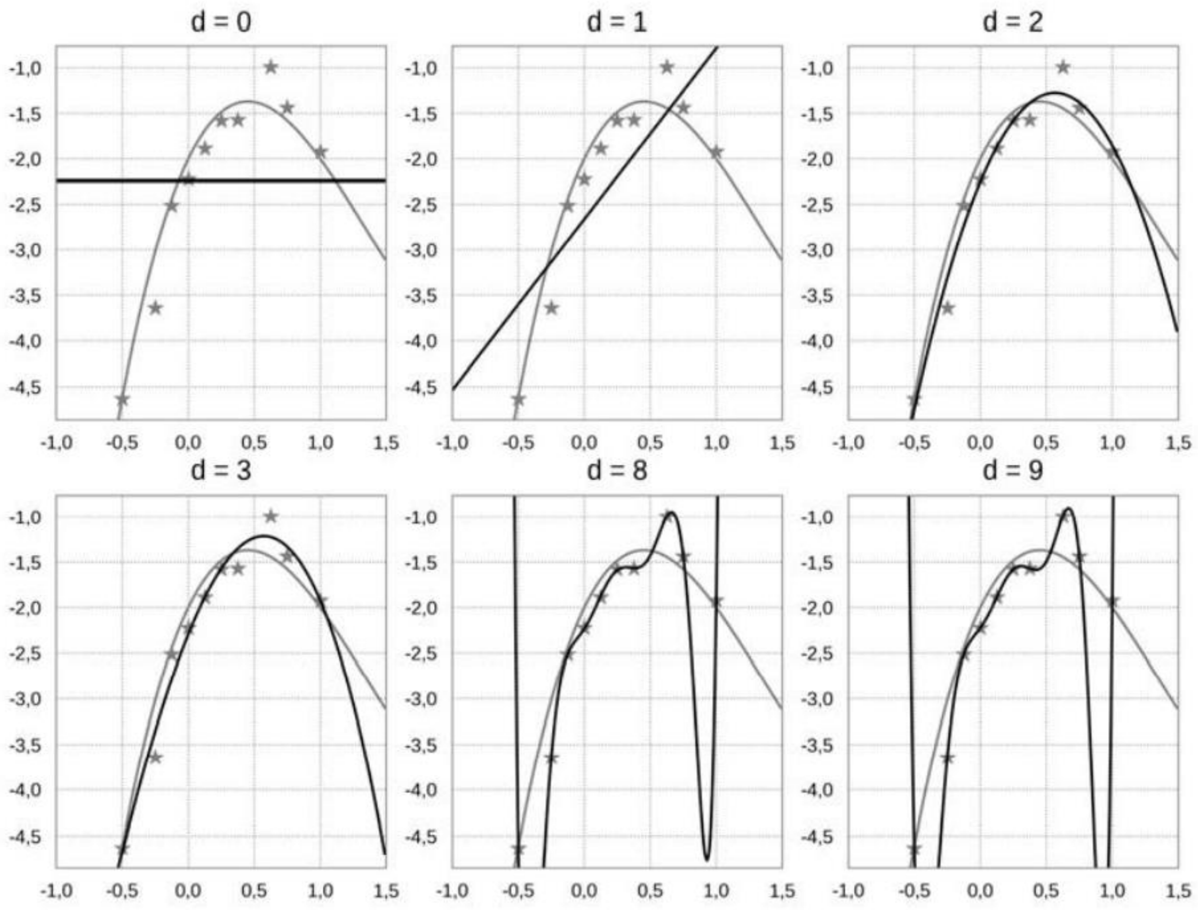
Матричная форма записи проведенных испытаний

$$Y = X\theta + \varepsilon.$$

# РЕГУЛЯРИЗАЦИЯ

- Высокая точность определения элементов из тренировочного набора данных
- Низкая предсказательная способность

# ПРИМЕР: АППРОКСИМАЦИЯ МНОГОЧЛЕНА



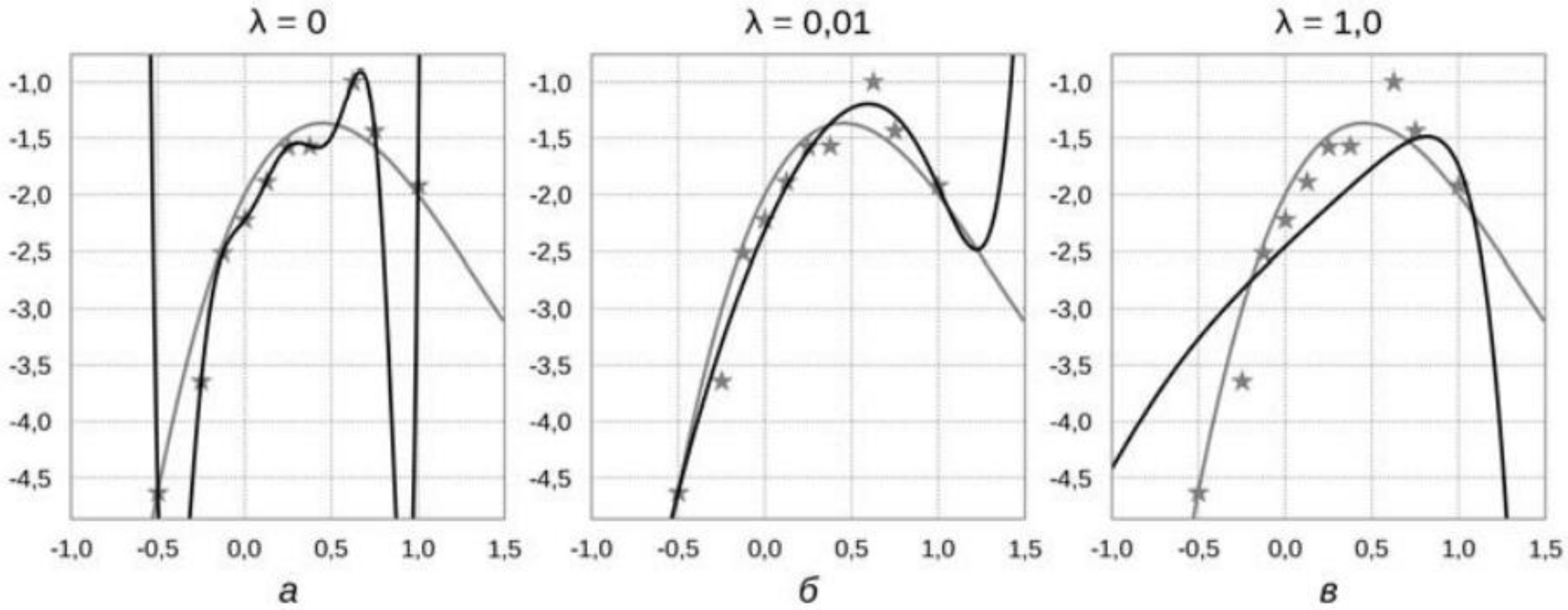

$$f(x) = x^3 - 4x^2 + 3x - 2$$

d – степень многочлена

**Проблема** — чем больше степень многочлена, тем, конечно, точнее им можно приблизить данные, но в какой-то момент результаты приближения перестанут иметь отношение к действительности.

Решение – добавление дополнительных слагаемых  регуляризаторов

# ПРИМЕР ДОБАВЛЕНИЯ РЕГУЛЯРИЗАТОРОВ




d = 9

Добавление регуляризатора - формализация факта, что небольшие, короткие векторы коэффициентов более вероятны, чем длинные.

# ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Задача классификации с вероятностной точки зрения: каждому классу $C_k$ сопоставляется плотность $p(x \mid C_k)$, где $p(C_k)$ – это априорные распределения (размеры классов)
- По теореме Байеса

$$p(\mathcal{C}_1 \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x} \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1+e^{-a}} = \sigma(a),$$

где

$$a = \ln \frac{p(\boldsymbol{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)},$$

$$\boxed{\sigma(a) = \frac{1}{1+e^{-a}}.}$$

**Логистический сигмоид**

a - линейная функция от входных признаков:

$$a = \boldsymbol{w}^{\top}\boldsymbol{x}$$

# ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Сигмоид переводит любое вещественное число на отрезок [0, 1]; чем меньше аргумент, тем меньше результат (на минус бесконечности получается 0), и наоборот, на плюс бесконечности получается 1.

$$p(\mathcal{C}_1 \mid \boldsymbol{x}) = y(\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x}), \qquad p(\mathcal{C}_2 \mid \boldsymbol{x}) = 1 - p(\mathcal{C}_1 \mid \boldsymbol{x}),$$

и для обучения можно просто напрямую оптимизировать правдоподобие по $\boldsymbol{w}$.

- SoftMax – функция (вместо сигмоида) - сглаженный максимум

$$p(\mathcal{C}_k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{j=1}^{K} p(\boldsymbol{x} \mid \mathcal{C}_j)p(\mathcal{C}_j)} = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}}$$

# ТЕОРИЯ ИНФОРМАЦИИ

## Энтропия (Шеннона)

- измерения неопределенности в эксперименте

## Перекрестная энтропия

- сравнивает два распределения вероятности и говорит, насколько они похожи

## Расстояние Кульбака-Лейблера

- мера того, насколько схожи два распределения вероятности

## Алгоритм Витерби

- используется в NLP. **Алгоритм** поиска наиболее вероятной последовательности скрытых состояний

## Encoder-Decoder

https://techrocks.ru/2021/12/23/math-concepts-you-need-for-ai/

# ЭНТРОПИЯ

- Энтропия – мера недостающей информации в системе
- Энтропия – мера беспорядка, хаотичности. Информация направлена на устранение беспорядка
- Система имеет минимальную энтропию при наличии достоверного состояния
- Клод Шеннон - 1948

**Задача.** На факультете учатся 100 студентов, из них 20 на ИБ, 30 на ИВТ, а остальные – на ПИЭ. *Сколько информации* несет сообщение о том, что студент учится на ИБ (ИВТ, ПИЭ)?

**Формула:** $\boxed{I_i = -\log_2 p_i = \log_2 \frac{1}{p_i}}$

**Решение:**

ИБ $p_1 = \frac{20}{100} = 0{,}2$ $I_1 = -\log_2 0{,}2 = \log_2 5 \approx 2{,}32$ бита

ИВТ $p_2 = \frac{30}{100} = 0{,}3$ $I_2 = -\log_2 0{,}3 \approx \log_2 3{,}33 \approx 1{,}74$ бита

ПИЭ $p_3 = \frac{50}{\ } = 0{,}5$ $I_3 = -\log_2 0{,}5 = \log_2 2 = 1$ бит

# РАССТОЯНИЕ КУЛЬБАКА - ЛЕЙБЛЕРА

- Кросс-**энтропия** (**Перекрестная энтропия**) – это Функция потерь (Loss Function), которую можно использовать для количественной оценки разницы между двумя Распределениями вероятностей
- Разница между двумя вероятностными распределениями: истинным и приближенным – оценка качества приближения.

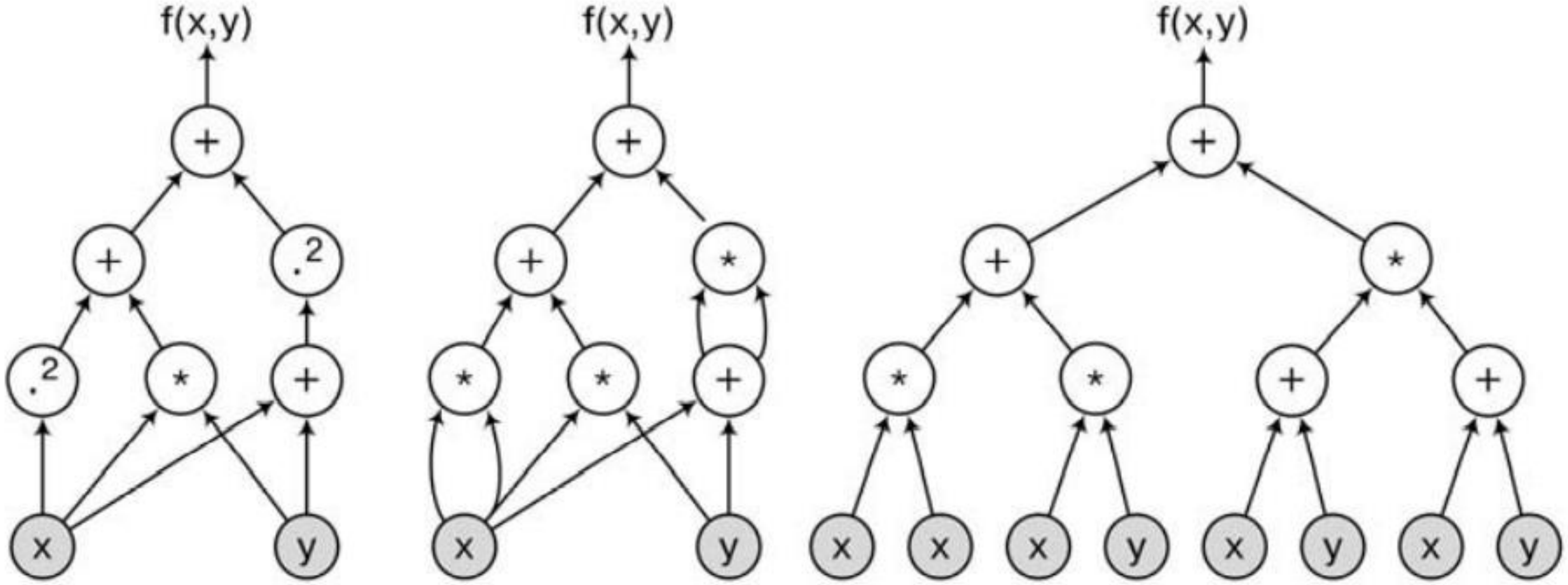
$$\mathrm{KL}\left(P\|Q\right) = \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

- Насколько распределение на тестовых примерах, порожденное классификатором Q, похоже или непохоже на «истинное» распределение, задаваемое данными P.

# ГРАФ ВЫЧИСЛЕНИЙ

$$f(x,y) = x^2 + xy + (x+y)^2.$$

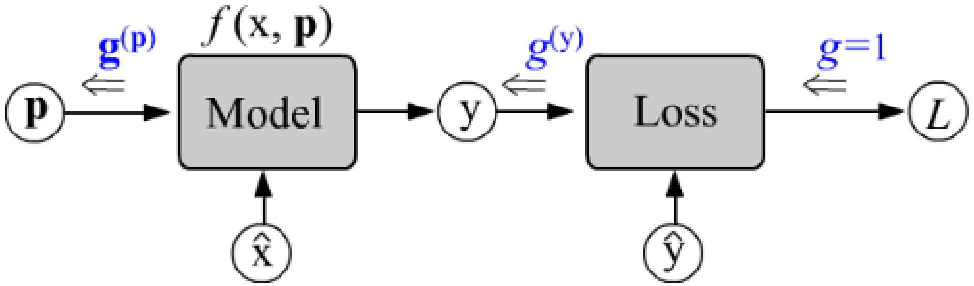
- Это способ представления композиции сложной функции из простых
- Граф вычислений — это граф, узлами которого являются функции (обычно достаточно простые, взятые из заранее фиксированного набора), а ребра связывают функции со своими аргументами.
- Значения, которые вводятся в узлы и выходят из узлов, называются тензорами



Пусть модель задана функцией $\mathbf{y} = f(\mathbf{x}, \mathbf{p})$, где $\mathbf{x}$ - вектор признаков объекта, а $\mathbf{p}$ - вектор параметров.
Кроме этого, есть $N$ обучающих примеров $\{\hat{\mathbf{x}}, \hat{\mathbf{y}}\} = (\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{y}}^{(1)}), \dots, (\hat{\mathbf{x}}^{(N)}, \hat{\mathbf{y}}^{(N)})$, которые помечаются шляпкой.
Параметры $\mathbf{p}$ модели подбираются так, чтобы минимизировать среднюю ошибку $L$ на один пример:

$$L = L(\mathbf{p}) = \frac{1}{N}\sum_{i=1}^{N} L(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}), \qquad \mathbf{y}^{(i)} = f(\hat{\mathbf{x}}^{(i)}, \mathbf{p}).$$

Чтобы найти минимум $L$, необходимо сдвигать вектор параметров в направлении обратном к градиенту (частным производным) $L$ по $\mathbf{p}$. Величина сдвига определяется скалярным гиперпараметром $\lambda$ (скорость обучения):

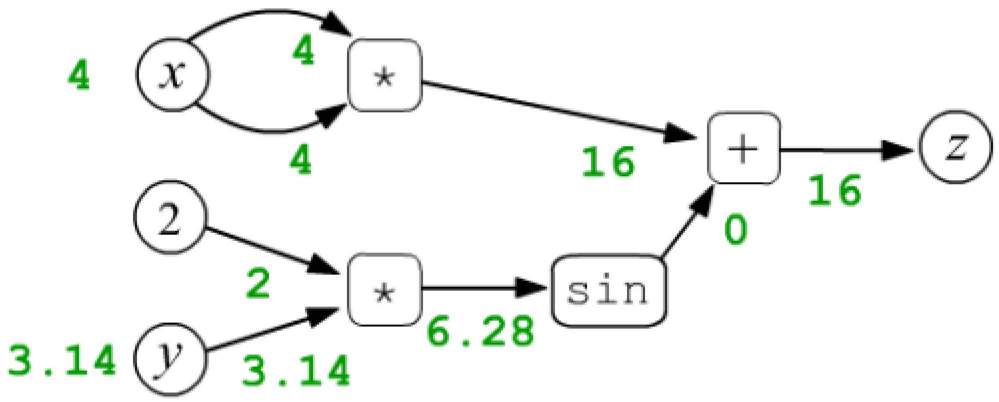



$$\mathbf{p} \mapsto \mathbf{p} - \lambda\,\mathbf{g}, \qquad \mathbf{g} = \frac{\partial L}{\partial \mathbf{p}}.$$

Вычисления проводятся в два этапа. Сначала при **прямом проходе** (слева-направо) вычисляется ошибка $L$ при заданных $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ и $\mathbf{p}$. При этом определённые значения принимают все узлы вычислительного графа.
Затем запускается процедура **обратного прохода**. Начальный градиент $g$ в скалярном узле ошибки $L$ является скаляром. Проходя через узлы графа от ошибки справа-налево, он превращается в вектор $\mathbf{g}$ (в общем случае в тензор). Изменения $\mathbf{g}$ на каждом узле происходят таким образом, что, когда он "добирается" до параметров $\mathbf{p}$, то оказывается равным частным производным: $\mathbf{g}^{(\mathbf{p})} = \partial L/\partial \mathbf{p}$.

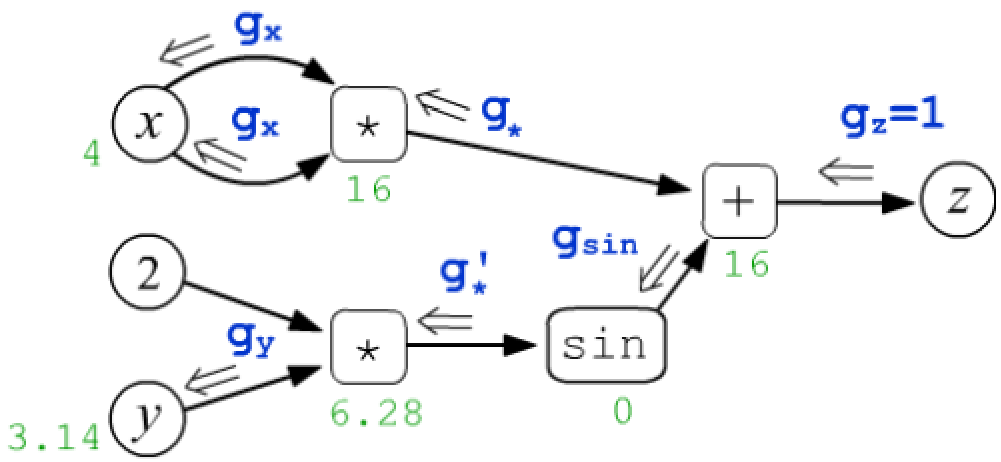
https://qudata.com/ml/ru/ML_Comp_Graph.html

$$z = x \cdot x + \sin(2 \cdot y)$$




$$g_x = \frac{\partial z}{\partial x} \quad = \quad 2 \cdot x,$$

$$g_y = \frac{\partial z}{\partial y} \quad = \quad 2\cos(2 \cdot y)$$



# МАТЕМАТИКА – ЭТО ИНТЕРЕСНО!!!