

УПРАВЛЕНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

матстат, искусственный интеллект, машинное обучение I

А.В. Макаренко

avm@rdcn.ru

Научно-исследовательская группа «Конструктивная Кибернетика»
Москва, Россия, www.rdcn.ru

Институт проблем управления РАН
Москва, Россия

Учебный курс – Лекция 2

20 февраля 2020 г.

ИПУ РАН, Москва, Россия

- ① Математическая статистика
- ② Искусственный интеллект
- ③ Машинное обучение-I
- ④ Заключение

Outline section

- ① Математическая статистика
 - Общие положения
 - Основные классы решаемых задач
 - Байесов подход
 - Язык R
 - Задачи анализа данных

② Искусственный интеллект

③ Машинное обучение-I

④ Заключение

Базовые определения I

СОБЫТИЕ – (кибернетика, физика) – это то, что происходит в конкретный момент времени, в конкретном месте пространства, и изменяет состояние системы.

Базовые определения I

СОБЫТИЕ – (кибернетика, физика) – это то, что происходит в конкретный момент времени, в конкретном месте пространства, и изменяет состояние системы.

СЛУЧАЙНОЕ СОБЫТИЕ – это событие, появление которого невозможно заранее предсказать. Является A подмножеством Ω – пространства элементарных событий ω .

Базовые определения I

СОБЫТИЕ – (кибернетика, физика) – это то, что происходит в конкретный момент времени, в конкретном месте пространства, и изменяет состояние системы.

СЛУЧАЙНОЕ СОБЫТИЕ – это событие, появление которого невозможно заранее предсказать. Является A подмножеством Ω – пространства элементарных событий ω .

СЛУЧАЙНЫЙ ЭКСПЕРИМЕНТ – это математическая модель соответствующего реального эксперимента, результат которого невозможно точно предсказать.

Базовые определения I

СОБЫТИЕ – (кибернетика, физика) – это то, что происходит в конкретный момент времени, в конкретном месте пространства, и изменяет состояние системы.

СЛУЧАЙНОЕ СОБЫТИЕ – это событие, появление которого невозможно заранее предсказать. Является A подмножеством Ω – пространства элементарных событий ω .

СЛУЧАЙНЫЙ ЭКСПЕРИМЕНТ – это математическая модель соответствующего реального эксперимента, результат которого невозможно точно предсказать.

ЭЛЕМЕНТАРНОЕ СЛУЧАЙНОЕ СОБЫТИЕ – это конкретный исход ω случайного эксперимента.

Базовые определения I

СОБЫТИЕ – (кибернетика, физика) – это то, что происходит в конкретный момент времени, в конкретном месте пространства, и изменяет состояние системы.

СЛУЧАЙНОЕ СОБЫТИЕ – это событие, появление которого невозможно заранее предсказать. Является A подмножеством Ω – пространства элементарных событий ω .

СЛУЧАЙНЫЙ ЭКСПЕРИМЕНТ – это математическая модель соответствующего реального эксперимента, результат которого невозможно точно предсказать.

ЭЛЕМЕНТАРНОЕ СЛУЧАЙНОЕ СОБЫТИЕ – это конкретный исход ω случайного эксперимента.

ПРОСТРАНСТВО ЭЛЕМЕНТАРНЫХ СОБЫТИЙ – это множество Ω всех различных исходов случайного эксперимента.

Базовые определения I

СОБЫТИЕ – (кибернетика, физика) – это то, что происходит в конкретный момент времени, в конкретном месте пространства, и изменяет состояние системы.

СЛУЧАЙНОЕ СОБЫТИЕ – это событие, появление которого невозможно заранее предсказать. Является A подмножеством Ω – пространства элементарных событий ω .

СЛУЧАЙНЫЙ ЭКСПЕРИМЕНТ – это математическая модель соответствующего реального эксперимента, результат которого невозможно точно предсказать.

ЭЛЕМЕНТАРНОЕ СЛУЧАЙНОЕ СОБЫТИЕ – это конкретный исход ω случайного эксперимента.

ПРОСТРАНСТВО ЭЛЕМЕНТАРНЫХ СОБЫТИЙ – это множество Ω всех различных исходов случайного эксперимента.

СЛУЧАЙНАЯ ВЕЛИЧИНА – это функция $y = X(\omega)$, которая ставит в соответствие исходу ω численное значение y . Возможен также вариант $y = X(A)$.

Базовые определения II

ВЕРОЯТНОСТЬ – степень (относительная мера, количественная оценка) возможности наступления некоторого события. Исследование вероятности с математической точки зрения составляет особую дисциплину – теорию вероятностей. В теории вероятностей и математической статистике понятие вероятности формализуется как числовая характеристика события – вероятностная мера (или её значение) – мера на множестве событий (подмножеств множества элементарных событий), принимающая значения от 0 (*Невозможное событие*) до 1 (*Достоверное событие*).

Базовые определения II

ВЕРОЯТНОСТЬ – степень (относительная мера, количественная оценка) возможности наступления некоторого события. Исследование вероятности с математической точки зрения составляет особую дисциплину – теорию вероятностей. В теории вероятностей и математической статистике понятие вероятности формализуется как числовая характеристика события – вероятностная мера (или её значение) – мера на множестве событий (подмножеств множества элементарных событий), принимающая значения от 0 (*Невозможное событие*) до 1 (*Достоверное событие*).

Эмпирическое «определение» вероятности связано с частотой наступления события исходя из того, что при достаточно большом числе испытаний частота должна стремиться к объективной степени возможности этого события:

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N},$$

где N – количество наблюдений (кол-во случайных экспериментов), n – количество наступлений события A .

Базовые определения II

ВЕРОЯТНОСТЬ – степень (относительная мера, количественная оценка) возможности наступления некоторого события. Исследование вероятности с математической точки зрения составляет особую дисциплину – теорию вероятностей. В теории вероятностей и математической статистике понятие вероятности формализуется как числовая характеристика события – вероятностная мера (или её значение) – мера на множестве событий (подмножеств множества элементарных событий), принимающая значения от 0 (*Невозможное событие*) до 1 (*Достоверное событие*).

Эмпирическое «определение» вероятности связано с частотой наступления события исходя из того, что при достаточно большом числе испытаний частота должна стремиться к объективной степени возможности этого события:

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N},$$

где N – количество наблюдений (кол-во случайных экспериментов), n – количество наступлений события A .

В современном изложении теории вероятностей вероятность определяется аксиоматически, как частный случай абстрактной теории меры множества. Тем не менее, связующим звеном между абстрактной мерой и вероятностью, выражающей степень возможности наступления события, является именно частота его наблюдения.

Базовые определения III

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ случайной величины X называется вероятность неравенства $X \leq x$, рассматриваемая как функция параметра x :

$$F(x) = P(X \leq x).$$

Базовые определения III

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ случайной величины X называется вероятность неравенства $X \leq x$, рассматриваемая как функция параметра x :

$$F(x) = P(X \leq x).$$

Если случайная величина X дискретна, то есть её распределение однозначно задаётся функцией вероятности

$$p(x_i) = P(X = x_i) = p_i,$$

то функция распределения $F(x)$ этой случайной величины кусочно-постоянна и может быть записана в виде:

$$F(x) = \sum_{i: x_i \leq x} p_i.$$

Базовые определения III

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ случайной величины X называется вероятность неравенства $X \leq x$, рассматриваемая как функция параметра x :

$$F(x) = P(X \leq x).$$

Если случайная величина X дискретна, то есть её распределение однозначно задаётся функцией вероятности

$$p(x_i) = P(X = x_i) = p_i,$$

то функция распределения $F(x)$ этой случайной величины кусочно-постоянна и может быть записана в виде:

$$F(x) = \sum_{i: x_i \leq x} p_i.$$

Если случайная величина X непрерывна, то функция распределения $F(x)$ этой случайной величины есть интеграл

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

где $f(x)$ – плотность распределения с.в. X :

$$f(x) \geq 0, \forall x \in \mathbb{R}; \quad \int_{-\infty}^{+\infty} f(x) dx \equiv 1.$$

Замечания к базовым определениям

- Пространство элементарных событий может быть как дискретным – тогда говорят об элементарных событиях ω , так и непрерывным – тогда говорят об элементарных измерениях ω .
- Совокупность всех ω – элементарных событий случайного эксперимента составляет полную группу событий. Т.е. в результате произведённого случайного эксперимента непременно произойдет одно и только одно из них. Сумма вероятностей всех событий в группе всегда равна 1.
- Ни X , ни y не являются вероятностью наступления исхода ω или события A .

Пример трактовки базовых определений

Понятие	Обозн.	Пример
Случайный эксперимент	—	однократное бросание игральной кости
Элементарное случайное событие	ω	«выпала единица», «выпала двойка», ...
Пространство элементарных событий	Ω	вся совокупность: «выпала единица» ... «выпала шестёрка»
Случайное событие	A	выпало чётное число; или выпало число более трёх
Случайная величина	x	1 – «выпала единица», 2 – «выпала двойка», ...; или 0/1 – выпало нечётное/чётное число

Теория вероятностей vs Математическая статистика I

ТЕОРИЯ ВЕРОЯТНОСТЕЙ. Одной из задач т.в. является разработка методов нахождения вероятностей сложных событий и/или законов распределения составных случайных величин, исходя из известных вероятностей более простых событий и/или законов распределения элементарных случайных величин. Таким образом, в прикладном аспекте, т.в. занимается разработкой и исследованием вероятностных моделей систем/процессов, подверженных случайным факторам. Для т.в., как раздела чистой математики, характерен главным образом дедуктивный метод, при котором исследователь отталкивается от аксиом и утверждений, и вычисляет те или иные интересующие характеристики изучаемого явления.

Теория вероятностей vs Математическая статистика I

ТЕОРИЯ ВЕРОЯТНОСТЕЙ. Одной из задач т.в. является разработка методов нахождения вероятностей сложных событий и/или законов распределения составных случайных величин, исходя из известных вероятностей более простых событий и/или законов распределения элементарных случайных величин. Таким образом, в прикладном аспекте, т.в. занимается разработкой и исследованием вероятностных моделей систем/процессов, подверженных случайным факторам. Для т.в., как раздела чистой математики, характерен главным образом дедуктивный метод, при котором исследователь отталкивается от аксиом и утверждений, и вычисляет те или иные интересующие характеристики изучаемого явления.

Задача т.в.

При подбрасывании исследуемой монеты, с вероятностью p выпадает «орёл» и с вероятностью $(1 - p)$ – «решка». Какова вероятность того, что в результате N подбрасываний «орёл» выпадет ровно n раз?

Теория вероятностей vs Математическая статистика I

ТЕОРИЯ ВЕРОЯТНОСТЕЙ. Одной из задач т.в. является разработка методов нахождения вероятностей сложных событий и/или законов распределения составных случайных величин, исходя из известных вероятностей более простых событий и/или законов распределения элементарных случайных величин. Таким образом, в прикладном аспекте, т.в. занимается разработкой и исследованием вероятностных моделей систем/процессов, подверженных случайным факторам. Для т.в., как раздела чистой математики, характерен главным образом дедуктивный метод, при котором исследователь отталкивается от аксиом и утверждений, и вычисляет те или иные интересующие характеристики изучаемого явления.

Задача т.в.

При подбрасывании исследуемой монеты, с вероятностью p выпадает «орёл» и с вероятностью $(1 - p)$ – «решка». Какова вероятность того, что в результате N подбрасываний «орёл» выпадет ровно n раз?

Решение

На основе биномиального распределения, решение задачи формулируется в виде:

$$P(N, n) = C_N^n p^n (1 - p)^{N-n}.$$

Теория вероятностей vs Математическая статистика II

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. Одной из задач м.с является восстановление закона распределения исследуемой случайной величины, используя информацию, полученную из эксперимента (статистические данные). Таким образом, в прикладном аспекте, м.с. занимается уточнением (отбором) вероятностно-статистических моделей систем/процессов, подверженных случайным факторам. Для м.с., как раздела прикладной математики, характерно главным образом индуктивное построение, так как в этом случае исследователь идёт от наблюдения событий (систем, процессов) к гипотезам касаясь теоретического устройства изучаемых явлений.

Теория вероятностей vs Математическая статистика II

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. Одной из задач м.с является восстановление закона распределения исследуемой случайной величины, используя информацию, полученную из эксперимента (статистические данные). Таким образом, в прикладном аспекте, м.с. занимается уточнением (отбором) вероятностно-статистических моделей систем/процессов, подверженных случайным факторам. Для м.с., как раздела прикладной математики, характерно главным образом индуктивное построение, так как в этом случае исследователь идёт от наблюдения событий (систем, процессов) к гипотезам касаясь теоретического устройства изучаемых явлений.

В определённом смысле, математическая статистика решает задачи, обратные теории вероятностей, но при этом полностью базируется на понятийном и инструментальном аппарате т.в.

Теория вероятностей vs Математическая статистика III

Задача м.с.

Монета подбрасывается N раз, при этом «орёл» выпадает n раз. Что можно сказать о неизвестном параметре p ?

Теория вероятностей vs Математическая статистика III

Задача м.с.

Монета подбрасывается N раз, при этом «орёл» выпадает n раз. Что можно сказать о неизвестном параметре p ?

Схема решения

Исходно нам известно, что $0 \leq p \leq 1$. Кроме того, $p \neq 0$, если $n > 0$, и $p \neq 1$, если $n < N$. Далее вводится понятие наиболее правдоподобное значение p и малый интервал правдоподобных значений:

$$p_1 < \frac{n}{N} < p_2,$$

который содержит истинное значение p . Пусть $\delta = p_2 - p_1$, тогда чем больше δ , тем с большей достоверностью в интервал попадает истинное значение p , но при этом более широкий интервал даёт нам меньшую информацию об истинном значении p . Таким образом, в статистическом анализе всегда присутствует принципиальная неопределённость, которую с одной стороны необходимо принимать во внимание, а с другой – оценивать её значение.

Оценивание моментов с.в.

Дано:

$X^* = \{x_1, x_2, \dots, x_N\}$ – наблюдаемая выборка объёма N из генеральной совокупности X .

Найти:

Оценки моментов случайной величины.

Оценивание эмпирической функции распределения с.в.

Дано:

$X^* = \{x_1, x_2, \dots, x_N\}$ – наблюдаемая выборка объёма N из генеральной совокупности X .

Найти:

$F_N(x)$ – эмпирическую функцию распределения случайной величины, соответствующей выборке X^* .

Оценивание параметров функции распределения с.в.

Дано:

$X^* = \{x_1, x_2, \dots, x_N\}$ – наблюдаемая выборка объёма N из генеральной совокупности X .

$F(x, \mu_1, \mu_2, \dots, \mu_M)$ – теоретическая функции распределения случайной величины x , где $\mu_1, \mu_2, \dots, \mu_M$ – неизвестные параметры распределения.

Найти:

Оценки параметров $\mu_1^*, \mu_2^*, \dots, \mu_M^*$.

Проверка статистических гипотез

Дано:

$X^* = \{x_1, x_2, \dots, x_N\}$ – наблюдаемая выборка объёма N из генеральной совокупности X .

$F(x)$ – теоретическая функции распределения случайной величины x .

Найти:

Подтверждение совместимости значений X^* с гипотезой H_0 о том, что случайная величина x имеет распределение $F(x)$.

		Test	
		H_0	H_1
		N	P
Reference	H_0	TN	α
	H_1	β	TP

Теорема Байеса

$$p(H_j|A) = \frac{p(A|H_j) p(H_j)}{p(A)}$$

$p(H_j)$ – априорные ожидания (prior): насколько правдоподобна гипотеза H_j перед наблюдением события A .

$p(A|H_j)$ – правдоподобие (likelihood): насколько правдоподобно наступление события A при условии того, что гипотеза H_j верна.

$p(A)$ – маргинальная вероятность (marginal probability или evidence): вероятность наступления события A , суммированная по всем возможным гипотезам H_i :

$$p(A) = \sum_i p(A|H_i) p(H_i), \quad \sum_i p(H_i) \equiv 1.$$

$p(H_j|A)$ – апостериорное распределение (posterior): насколько правдоподобна гипотеза H_j при наблюдаемом событии A .

Смысл Байесова подхода

Замечания:

- Теорема Байеса позволяет «переставить местами причину (гипотезу) и следствие (событие)»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной.
- В интерпретации Байеса вероятность отражает уровень доверия. Теорема Байеса связывает воедино некие предположения до и после принятия во внимание очевидных (наблюдаемых измеримых) фактов.
- Разница между байесовской и классической интерпретацией вероятности достаточно фундаментальна. Классическая статистика оперирует порогом отвергания гипотезы H_0 , а байесовская – формирует апостериорные вероятности (уверенности) в гипотезах H_0 и H_j .

Смысл Байесова подхода

Замечания:











- Теорема Байеса позволяет «переставить местами причину (гипотезу) и следствие (событие)»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной.
- В интерпретации Байеса вероятность отражает уровень доверия. Теорема Байеса связывает воедино некие предположения до и после принятия во внимание очевидных (наблюдаемых измеримых) фактов.
- Разница между байесовской и классической интерпретацией вероятности достаточно фундаментальна. Классическая статистика оперирует порогом отвергания гипотезы H_0 , а байесовская – формирует апостериорные вероятности (уверенности) в гипотезах H_0 и H_j .

Полезно знать:

- Метод (понятие) максимального правдоподобия.
- Статистический последовательный анализ, критерий Вальда.
- Методы проверки сложных (составных, зависимых) гипотез.

Почему R?

IEEE 2016 Top Programming Languages

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

<https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016>

Макаренко А.В. Комплексный анализ данных и машинное обучение: 8 причин для миграции с Wolfram Mathematica на Python/R, www.rdcn.ru [Мнение, 2016].

История создания



- Автор языка: Росс Айхэка, Роберт Джентлмен.
- Мотивация названия: первая буква имён создателей языка.
- Дата первого релиза: 1993 г.
- Эталонная реализация: CRAN.
- Лицензия: GNU GPL.

Основные свойства языка

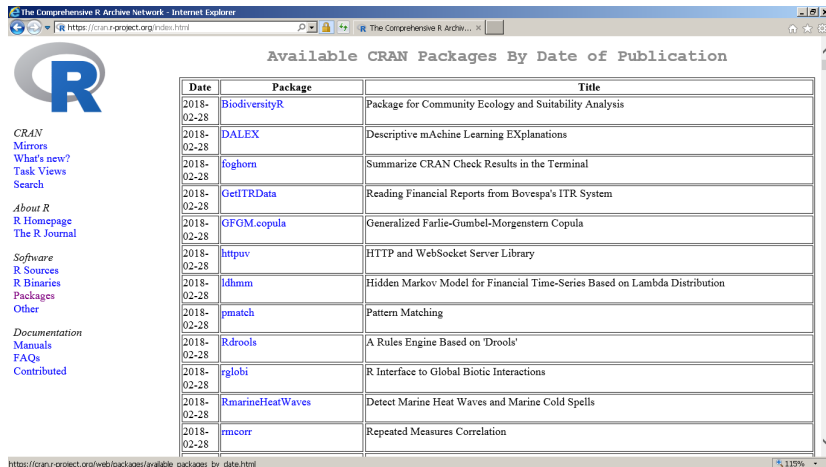
Парадигмы программирования:

- Императивная (процедурный, структурный, модульный подходы).
- Объектно-ориентированная.
- Функциональная.

Особенности языка:

- Язык предметной области (DSL) для обработки данных, высокоуровневый со встроенными высокоуровневыми структурами данных.
- Реализован поверх классической архитектуры интерпретатора-компилятора Scheme.
- Интерпретируемый (поддерживает REPL среду).
- Динамическая типизация, автоматическое управление памятью.
- Синтаксис ядра минималистичен, расширяется через пакеты.
- Код организовывается в функции и классы, которые могут объединяться в модули (они в свою очередь могут быть объединены в пакеты).
- Интегрируется с другими языками (C/C++, Python, Java, ...).

Экосистема R



The screenshot shows the CRAN website in Internet Explorer. The title bar reads "The Comprehensive R Archive Network - Internet Explorer". The address bar shows "https://cran.r-project.org/index.html". The page content includes the R logo, navigation links, and a table titled "Available CRAN Packages By Date of Publication".

Navigation Links:

- CRAN
- Mirrors
- What's new?
- Task Views
- Search
- About R
- R Homepage
- The R Journal
- Software
- R Sources
- R Binaries
- Packages
- Other
- Documentation
- Manuals
- FAQs
- Contributed

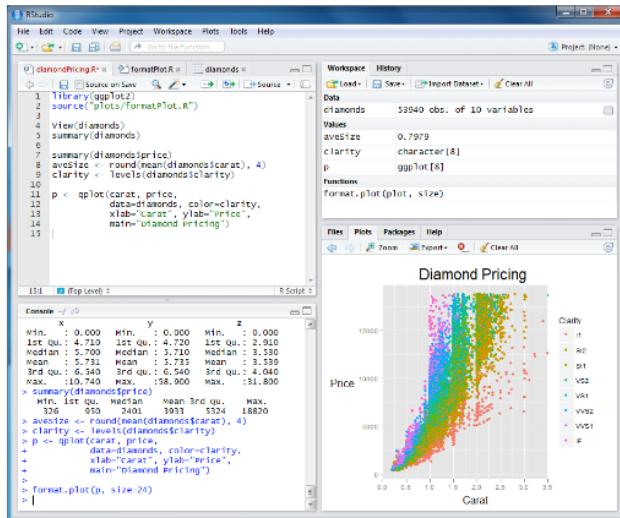
Available CRAN Packages By Date of Publication

Date	Package	Title
2018-02-28	BiodiversityR	Package for Community Ecology and Suitability Analysis
2018-02-28	DALEX	Descriptive mAchine Learning EXplanations
2018-02-28	foghorn	Summarize CRAN Check Results in the Terminal
2018-02-28	GetTRData	Reading Financial Reports from Bovespa's ITR System
2018-02-28	GFGM.copula	Generalized Farlie-Gumbel-Morgenstern Copula
2018-02-28	httpuv	HTTP and WebSocket Server Library
2018-02-28	ldhmm	Hidden Markov Model for Financial Time-Series Based on Lambda Distribution
2018-02-28	pmatch	Pattern Matching
2018-02-28	Rdrools	A Rules Engine Based on 'Drools'
2018-02-28	rglobi	R Interface to Global Biotic Interactions
2018-02-28	RmarineHeatWaves	Detect Marine Heat Waves and Marine Cold Spells
2018-02-28	rmcorr	Repeated Measures Correlation

Address bar: https://cran.r-project.org/web/packages/available_packages_by_date.html

Официальные пакеты расширения: 15 300

IDE RStudio



- Бесплатна
- Мультиплатформенна
- Автодополнение кода
- Навигация и формат кода
- Подсветка кода (слабовато)
- Работа с проектами
- Исполнение и отладка кода
- Доступ к R консоли
- Доступ к переменным
- Визуализация и интерактив

Скачать...

Разведочный анализ

РАЗВЕДОЧНЫЙ АНАЛИЗ – (РАД, Exploratory data analysis (EDA)) – анализ основных свойств набора данных, нахождение общих закономерностей, распределений и аномалий, построение начальных моделей (оценивание, предсказание, объяснение). Термин EDA был введен математиком Джоном Тьюки в 1961 г.

Заполнение пропусков

#	P1	P2	P3	P4	P5	P6
1		b	5	8.1	g	4
2	5	a	3	9.4	c	
3	9	c		5.7	k	1
4	1	b	4	1.3	k	
5		g	9	6.8	d	3
6	8	d	1	7.3		5
7	9	c		2.5	b	2
8	4		5	9.8	a	
9	6	a	4	6.2	0	1
10	8		3	3.4		5
11	7	d	1		f	3
12	11		9	2.6	k	6

Особенности реальных данных:

- Слабая структурированность
- Пропуски в данных
- Аномальные значения



Заполнение пропусков

#	P1	P2	P3	P4	P5	P6
1		b	5	8.1	g	4
2	5	a	3	9.4	c	
3	9	c		5.7	k	1
4	1	b	4	1.3	k	
5		g	9	6.8	d	3
6	8	d	1	7.3		5
7	9	c		2.5	b	2
8	4		5	9.8	a	
9	6	a	4	6.2	0	1
10	8		3	3.4		5
11	7	d	1		f	3
12	11		9	2.6	k	6

Причины пропусков в данных:

- Отсутствие данных
- Запрет на доступ к данным
- Проблемы с ПО



Заполнение пропусков

#	P1	P2	P3	P4	P5	P6
1		b	5	8.1	g	4
2	5	a	3	9.4	c	
3	9	c		5.7	k	1
4	1	b	4	1.3	k	
5		g	9	6.8	d	3
6	8	d	1	7.3		5
7	9	c		2.5	b	2
8	4		5	9.8	a	
9	6	a	4	6.2	0	1
10	8		3	3.4		5
11	7	d	1		f	3
12	11		9	2.6	k	6

Варианты борьбы с пропусками:

- Фильтрация набора данных
- Заполнение медианными значениями
- Заполнение на основе эмпирической ф.р.
- Заполнение на основе теоретической ф.р.



Заполнение пропусков

#	P1	P2	P3	P4	P5	P6
1		b	5	8.1	g	4
2	5	a	3	9.4	c	
3	9	c		5.7	k	1
4	1	b	4	1.3	k	
5		g	9	6.8	d	3
6	8	d	1	7.3		5
7	9	c		2.5	b	2
8	4		5	9.8	a	
9	6	a	4	6.2	0	1
10	8		3	3.4		5
11	7	d	1		f	3
12	11		9	2.6	k	6

Причины аномальных значений в данных:

- Искажение на уровне источника данных
- Проблемы с ПО считывания
- Неверные априорные представления



Формирование признаков

При формировании из «сырых» данных массива информативных признаков **T**, и перед их подачей на вход моделей машинного обучения, как правило, требуется проведение ряда преобразований:

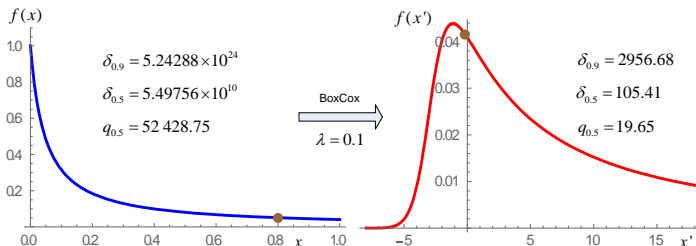
Формирование признаков

При формировании из «сырых» данных массива информативных признаков T , и перед их подачей на вход моделей машинного обучения, как правило, требуется проведение ряда преобразований:

Трансформация – нелинейное «выравнивание» функции распределения. Наиболее распространённый подход – это преобразование Бокса-Кокса:

Параметр λ выбирается через максимизацию логарифма правдоподобия. Второй способ: через поиск максимальной величины коэффициента корреляции между квантилями функции нормального распределения и отсортированной преобразованной последовательностью.

$$x' = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \ln x, & \lambda = 0, \end{cases}$$



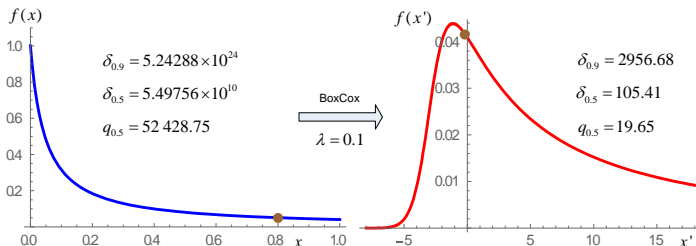
Формирование признаков

При формировании из «сырых» данных массива информативных признаков T , и перед их подачей на вход моделей машинного обучения, как правило, требуется проведение ряда преобразований:

Трансформация – нелинейное «выравнивание» функции распределения. Наиболее распространённый подход – это преобразование Бокса-Кокса:

Параметр λ выбирается через максимизацию логарифма правдоподобия. Второй способ: через поиск максимальной величины коэффициента корреляции между квантилями функции нормального распределения и отсортированной преобразованной последовательностью.

$$x' = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \ln x, & \lambda = 0, \end{cases}$$



Нормализация – линейный сдвиг и масштабирование величин в конкретный диапазон значений. Можно через с.к.о., но лучше через квантили.

Outline section

- 1 Математическая статистика
- 2 Искусственный интеллект
Общие положения
- 3 Машинное обучение-I
- 4 Заключение

Определение понятия

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ – это самостоятельное направление информатики, специализирующееся на разработке и исследовании искусственных интеллектуальных систем.

Определение понятия

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ – это самостоятельное направление информатики, специализирующееся на разработке и исследовании искусственных интеллектуальных систем.

ИСКУССТВЕННАЯ ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА – это аппаратно-программный комплекс, способный решать творческие задачи, традиционно считающиеся прерогативой человека.

Определение понятия

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ – это самостоятельное направление информатики, специализирующееся на разработке и исследовании искусственных интеллектуальных систем.

ИСКУССТВЕННАЯ ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА – это аппаратно-программный комплекс, способный решать творческие задачи, традиционно считающиеся прерогативой человека.

ТВОРЧЕСТВО – процесс деятельности, создающий качественно новые материальные и духовные ценности или итог создания объективно нового. Основным критерий, отличающий творчество от изготовления (производства) – уникальность его результата. Результат творчества невозможно прямо вывести из начальных условий. Именно этот факт придаёт продуктам творчества дополнительную ценность в сравнении с продуктами производства.

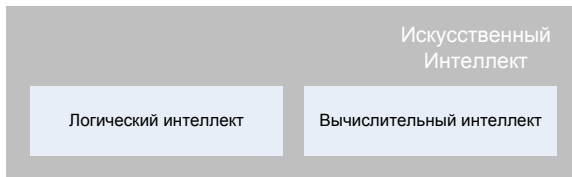
Классификация по возможностям

ИИ	узкий	широкий	
слабый	Распознавание образов, Сложные логические игры	Управление сложными системами	Адаптивное поведение
сильный	Формулирование и доказательство новых матем. теорем	42?	Разум
	конкретная предметная область	пересечение предметных областей	

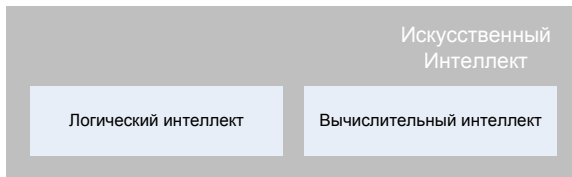
АДАПТИВНОЕ ПОВЕДЕНИЕ – (в кибернетике) – способность системы к целенаправленному приспособляющемуся поведению в сложных средах при изменении как внутренних, так и внешних условий. **Особенность:** не требуется понимание смысла оперируемой информации.

РАЗУМ – высший тип мыслительной (познавательной) деятельности, способность мыслить всеобще, способность анализа, абстрагирования и обобщения. **Особенность:** прохождение теста Тьюринга.

Классификация по методам I

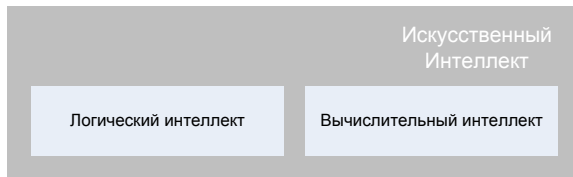


Классификация по методам I



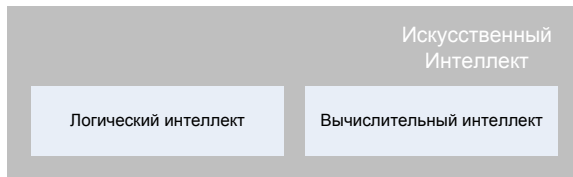
ЛОГИЧЕСКИЙ ИНТЕЛЛЕКТ – основан на строгом логическом выводе, в качестве основного математического инструментария применяются описательные логики.

Классификация по методам I



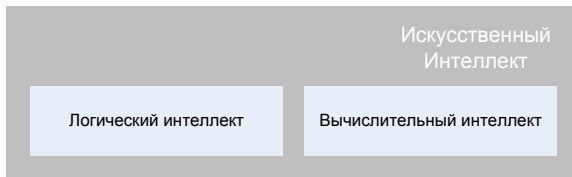
ОПИСАТЕЛЬНЫЕ ЛОГИКИ – семейство языков представления знаний, позволяющих описывать понятия предметной области в недвусмысленном, формализованном виде. Они сочетают в себе, с одной стороны, богатые выразительные возможности, а с другой – хорошие вычислительные свойства, такие как разрешимость и относительно невысокая вычислительная сложность основных логических проблем, что делает возможным их применение на практике. Являются основой для построения онтологий.

Классификация по методам I

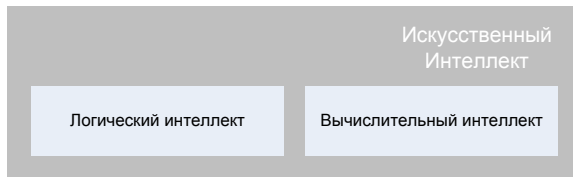


ОНТОЛОГИЯ – (в информатике) – это попытка всеобъемлющей и подробной формализации некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из структуры данных, содержащей все релевантные классы объектов, их связи и правила (теоремы, ограничения), принятые в этой предметной области.

Классификация по методам II

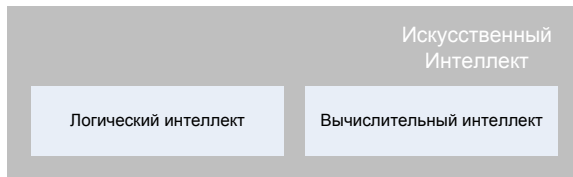


Классификация по методам II



ВЫЧИСЛИТЕЛЬНЫЙ ИНТЕЛЛЕКТ – опирается на эвристические алгоритмы, в качестве основного математического инструментария применяется машинное обучение.

Классификация по методам II



ВЫЧИСЛИТЕЛЬНЫЙ ИНТЕЛЛЕКТ – опирается на эвристические алгоритмы, в качестве основного математического инструментария применяется машинное обучение.

Тесно связан с концепцией мягких вычислений и кибернетикой.

Дополнительно: [IEEE Computational Intelligence Society](#).

Проблематика



Джон Маккарти
(04.09.1927 – 24.10.2011)

«Проблема состоит в том, что пока мы не можем в целом определить, какие вычислительные процедуры мы хотим называть интеллектуальными. Мы понимаем некоторые механизмы интеллекта и не понимаем остальные. Поэтому под интеллектом в пределах этой науки понимается только вычислительная составляющая способности достигать целей в мире.»

John McCarthy, WHAT IS ARTIFICIAL INTELLIGENCE? [Читать...](#)

Outline section

① Математическая статистика

② Искусственный интеллект

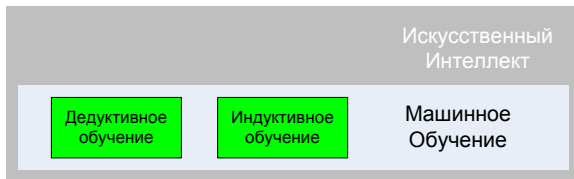
③ Машинное обучение-I

Общие положения

О данных

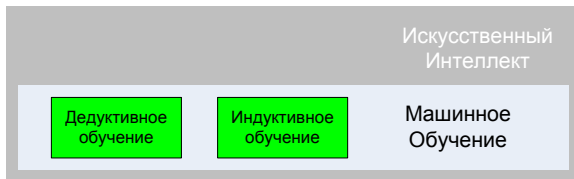
④ Заключение

Определение понятия



МАШИННОЕ ОБУЧЕНИЕ – обширный (центральный) подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Определение понятия



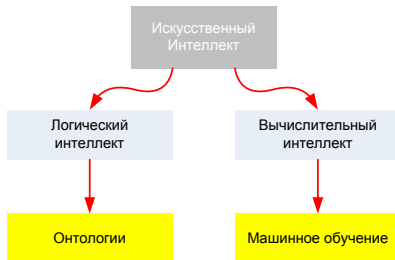
МАШИННОЕ ОБУЧЕНИЕ – обширный (центральный) подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Различают два типа обучения машин:

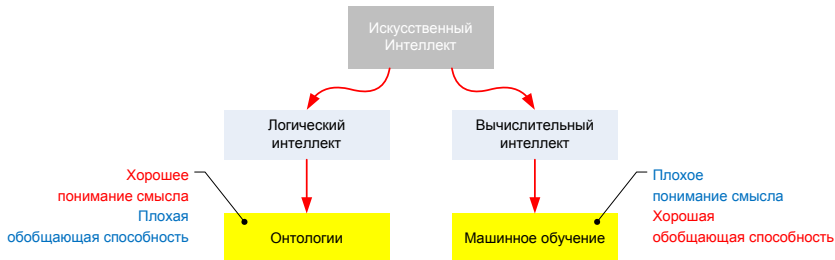
- Дедуктивное обучение – предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний (область экспертных систем).
- Индуктивное обучение – (обучение по прецедентам) – основано на выявлении общих закономерностей по частным эмпирическим (экспериментальным) данным.

Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. Наука, 1974.

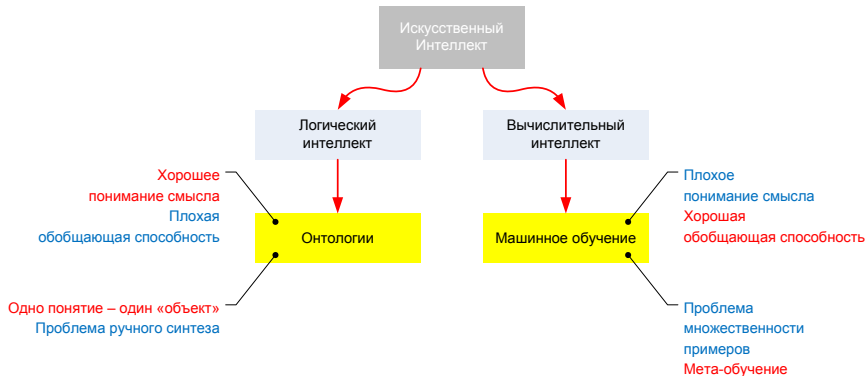
Проблематика ИИ 2-го порядка



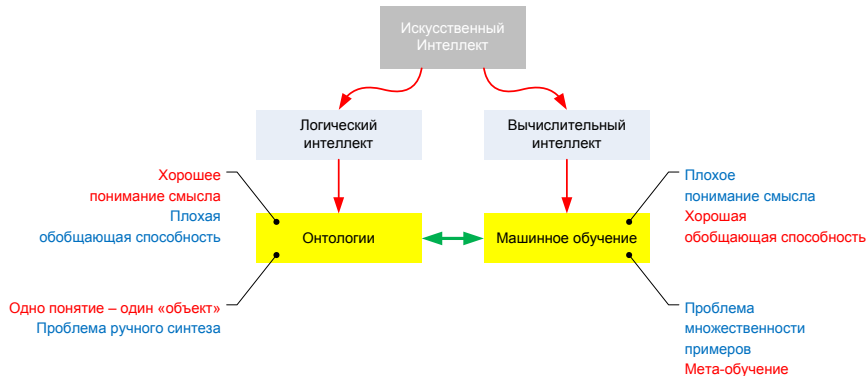
Проблематика ИИ 2-го порядка



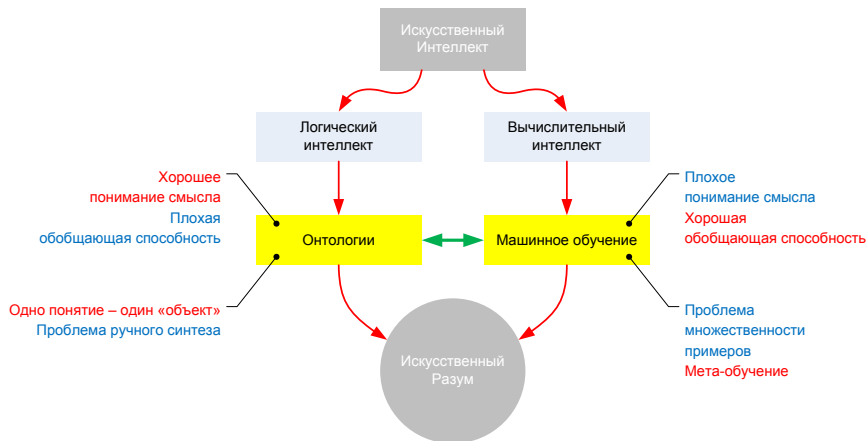
Проблематика ИИ 2-го порядка



Проблематика ИИ 2-го порядка



Проблематика ИИ 2-го порядка

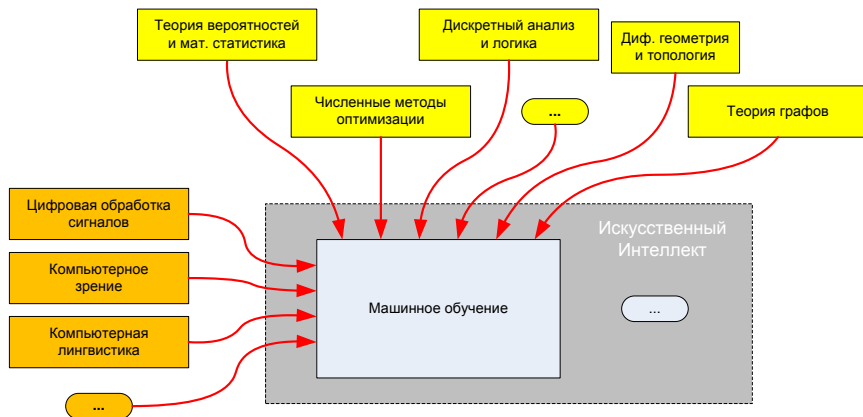


По теме: Besold T.R. et al. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. ArXiv: [1711.03902](https://arxiv.org/abs/1711.03902).

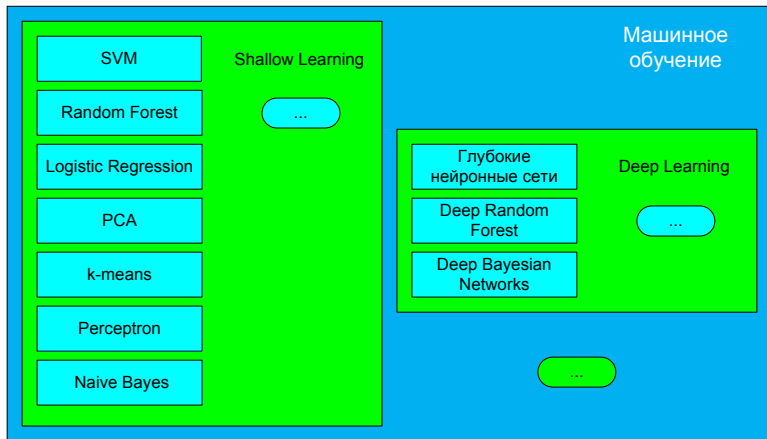
Замечание

Далее, если специально не будет оговорено иное, под искусственным интеллектом мы будем понимать вычислительный интеллект в его слабой форме, и машинное обучение в форме обучения по прецедентам.

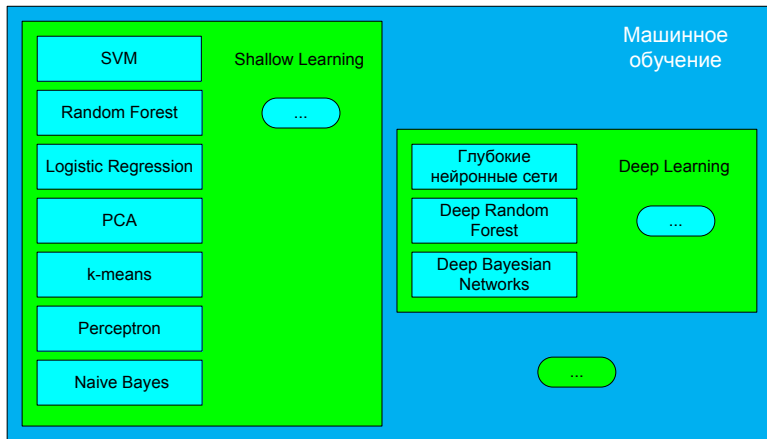
Внешние составляющие



Внутренняя структура



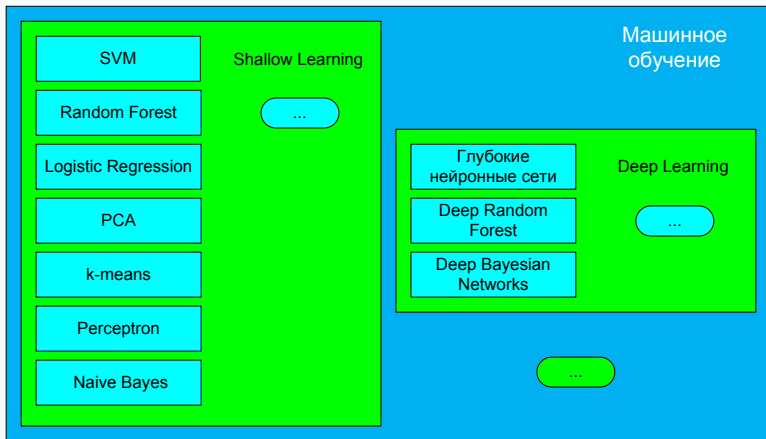
Внутренняя структура



Основная особенность **Shallow Learning** алгоритмов:

- Для их обучения требуются (как правило) вручную синтезируемые высокоуровневые признаки.

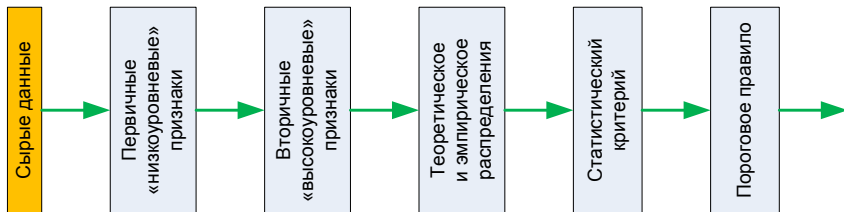
Внутренняя структура



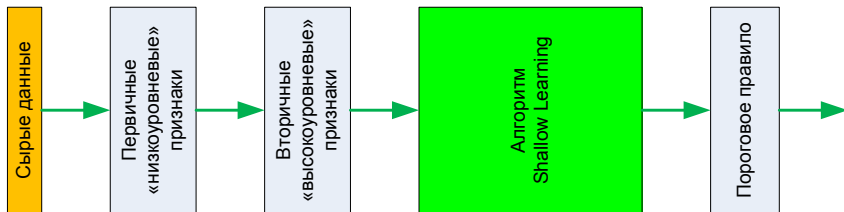
Основная особенность **Deep Learning** алгоритмов:

- Они работают с исходными данными (низкоуровневыми признаками) и самостоятельно извлекают (формируют) признаковое описание объектов.

Объёмы «ручного труда»



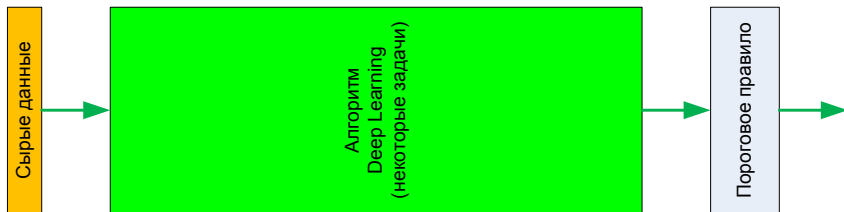
Объёмы «ручного труда»



Объёмы «ручного труда»



Объёмы «ручного труда»



Шкалы измерений I

ШКАЛА ИЗМЕРЕНИЙ – (согласно РМГ 83-2007)¹ – отображение множества различных проявлений количественного или качественного свойства [наблюдаемого объекта] на принятое по соглашению упорядоченное множество чисел или другую систему логически связанных знаков (обозначений).

¹Государственная система обеспечения единства измерений. Шкалы измерений. Термины и определения.

Шкалы измерений I

ШКАЛА ИЗМЕРЕНИЙ – (согласно РМГ 83-2007)¹ – отображение множества различных проявлений количественного или качественного свойства [наблюдаемого объекта] на принятое по соглашению упорядоченное множество чисел или другую систему логически связанных знаков (обозначений).

ИЗМЕРЕНИЕ – сравнение конкретного проявления измеряемого свойства (величины) [наблюдаемого объекта] со шкалой измерений этого свойства (величины) в целях получения результата измерений (оценки свойства или значения величины).

¹Государственная система обеспечения единства измерений. Шкалы измерений. Термины и определения.

Шкалы измерений I

ШКАЛА ИЗМЕРЕНИЙ – (согласно РМГ 83-2007)¹ – отображение множества различных проявлений количественного или качественного свойства [наблюдаемого объекта] на принятое по соглашению упорядоченное множество чисел или другую систему логически связанных знаков (обозначений).

ИЗМЕРЕНИЕ – сравнение конкретного проявления измеряемого свойства (величины) [наблюдаемого объекта] со шкалой измерений этого свойства (величины) в целях получения результата измерений (оценки свойства или значения величины).

СРАВНЕНИЕ – познавательная операция, заключающаяся в нахождении сходства и различия между предметами, явлениями, событиями и лежащая в основе суждений о сходстве или различии объектов. Один из главных способов познания окружающего мира. При сравнении устанавливают закономерности, присущие объектам, системам объектов и их характеристикам.

¹Государственная система обеспечения единства измерений. Шкалы измерений. Термины и определения.

Шкалы измерений I

ШКАЛА ИЗМЕРЕНИЙ – (согласно РМГ 83-2007)¹ – отображение множества различных проявлений количественного или качественного свойства [наблюдаемого объекта] на принятое по соглашению упорядоченное множество чисел или другую систему логически связанных знаков (обозначений).

ИЗМЕРЕНИЕ – сравнение конкретного проявления измеряемого свойства (величины) [наблюдаемого объекта] со шкалой измерений этого свойства (величины) в целях получения результата измерений (оценки свойства или значения величины).

СРАВНЕНИЕ – познавательная операция, заключающаяся в нахождении сходства и различия между предметами, явлениями, событиями и лежащая в основе суждений о сходстве или различии объектов. Один из главных способов познания окружающего мира. При сравнении устанавливают закономерности, присущие объектам, системам объектов и их характеристикам.

Два замечания:

- Различают пять основных типов шкал.
- Различают одномерные и многомерные шкалы измерений.

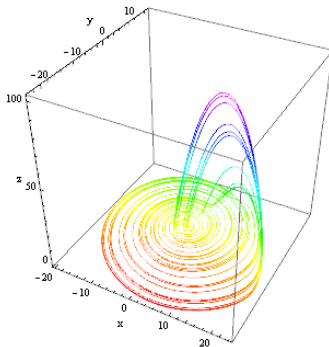
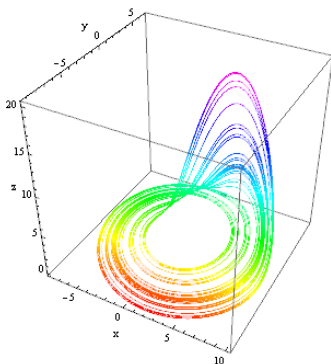
¹Государственная система обеспечения единства измерений. Шкалы измерений. Термины и определения.

Шкалы измерений II

Признак типа шкалы измерений	Тип шкалы измерений				
	Наименований	Порядка	Интервалов	Отношений	Абсолютные
Эквивалентность $A = B, A \neq B$	+	+	+	+	+
Порядок $A < B < C$		+	+	+	+
Метрика $d(A, B)$			+	+	+
Логарифмирование $\log A, \log B$			+	+	+
Единица измерения (масштаб)	Не применимо	Не применимо $\neg \exists A - B$	По соглашению $\neg \exists A / B$	По соглашению $\exists A / B$	Естественный $\exists A / B$
Наличие нуля	Не применимо	Необязательно	По соглашению	Естественный	Естественный
Допустимые преобразования $A' = f(A), B' = f(B)$	Изоморфное отображение	Монотонные преобразования	Линейные преобразования	Умножение (1, 2 р.) Сумма (2 р.)	Тожественные преобразования
Пример	Москва, Рубашка, Планета,, Холодно, Тепло, Жарко, ...	Шкала Времени Шкала Тем-ры, °C	Шкала Массы Шкала Тем-ры, К	Кол-во квантов Коэф. усиления

Формальное представление

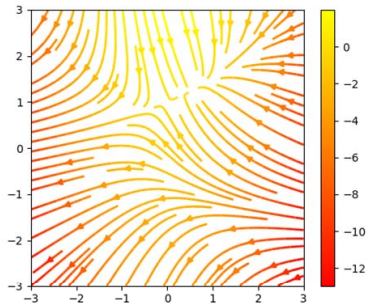
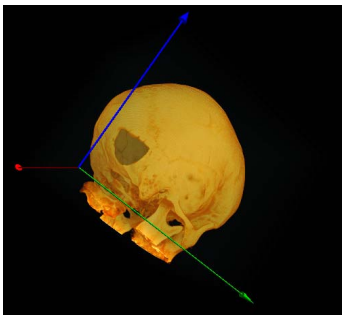
Временной ряд:



$$\text{TS} = \left\{ \{s_k\}_{k=1}^K, \{t_k\}_{k=1}^K \right\}, \quad s \in S \subset \mathbb{R}^N, \quad t \in T \subset \mathbb{R}, \quad t_k < t_{k+1}.$$

Формальное представление

Дискретное скалярное (векторное) поле:

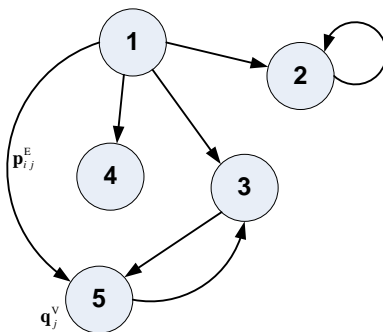


$$\mathbf{s}(\mathbf{r}_m, t_k) = \mathbf{S}\left(r_m^{(0)}, \dots, r_m^{(M)}, t_k\right), \quad \mathbf{s} \in \mathbf{S} \subset \mathbb{R}^N, \quad \mathbf{r} \in \mathbf{R} \subset \mathbb{R}^M,$$

$$t \in \mathbf{T} \subset \mathbb{R}, \quad t_k \prec t_{k+1}.$$

Формальное представление

Граф:



$\Gamma|t_k = \langle V, E \rangle|t_k$, V – множество вершин, E – множество рёбер,
 $t \in T \subset \mathbb{R}$, $t_k \prec t_{k+1}$.

Формальное представление

Мультимножество:



$$M|t_k = \{A_1, A_2, A_1, A_3, \dots\} | t_k \quad A_i - \text{множество},$$

$$t \in T \subset \mathbb{R}, \quad t_k \prec t_{k+1}.$$

Большие данные

БОЛЬШИЕ ДАННЫЕ – это набор стратегий, методов и технологий связанных со сбором, хранением и обработкой наборов данных, отвечающих следующим условиям:

- Большой объём данных, превосходящий возможности их сохранения «на одном сервере».
- Высокая скорость поступления данных (режим реального времени).
- Существенная неструктурированность и гетерогенность поступающих данных.

Большие данные

Условная граница Больших Данных:

- по объёму – один RAID контроллер с 256 SATA HDD – усреднённая оценка – это 1 ПБ (1024 ТБ).
- по скорости – примерно на уровне возможностей современных SATA HDD, что составляет около 200 МБ/с.

Большие данные

Три важных аспекта обработки Больших Данных:

- глубина анализа данных, позволяющая обнаружить, осмыслить и понять суть явления, а также спрогнозировать его последствия.
- проявление (статистически значимых только на больших объёмах выборок) тонких, порой контринтуитивных, эффектов.
- особые требования к алгоритмам и программному обеспечению:
 - (i) – производительность (массовая параллельность, алгоритмы типа $\log N$, N , $N \log N$);
 - (ii) – устойчивость (отсутствие рекурсивных вызовов, вспомогательная память не более $N^{3/4}$);
 - (iii) – организация (асинхронная среда с распределённой памятью, минимизация дисковых и коммуникационных операций).

Дополнительно: Макаренко А.В. Интеллектуальное управление. Введение / Теория управления (дополнительные главы). М.: ЛЕНАНД, ИПУ РАН, 2019. С. 359-367.

Outline section

- ① Математическая статистика
- ② Искусственный интеллект
- ③ Машинное обучение-I
- ④ **Заклучение**

Контрольная работа

Задание для слушателей:

- 1 Скачать датасет «Fashion MNIST». Оформить Jupyter Notebook с кодом на Python: (i) – функция чтения файлов, через `np.fromfile()`, с объектами и метками; (ii) – функция визуализации, через `plt.imshow()`, трёх случайных объектов. Чтение файлов должно выполняться за два вызова `np.fromfile()` – первый для заголовка, второй для тела файла. Чтение тела файла должно базироваться на константах, определяемых заголовком. На выходе функции массив изображений объектов должен иметь структуру `A[k,i,j]`, где `k` – индекс изображения объекта, `i, j` – строка и столбец 2D массива пикселей, соответственно. Преобразование в требуемый формат должно осуществляться средствами библиотеки NumPy, без применения циклов. С массивом меток – по аналогии.
- 2 Оформить Jupyter Notebook: (i) – описание метода SVD-разложения матриц; (ii) – области применения SVD-разложения; (iii) – пример решения какой-либо задачи с использованием SVD-разложения. При оформлении текста использовать разметку Markdown, формулы писать через команды LaTeX, код должен быть на языке Python.
- 3 Оформить Jupyter Notebook: (i) – изложить плюсы и минусы различных стратегий борьбы с пропусками; (ii) – сформировать демонстрационный набор данных, внести в него пропуски; (iii) – показать применение различных стратегий борьбы с пропусками. При оформлении текста использовать разметку Markdown, формулы писать через команды LaTeX, код должен быть на языках Python или R.