УПРАВЛЕНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
экосистема Python, математика

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

1 марта 2018 г.

ИПУ РАН, Москва, Россия

Outline

Outline section

Общие положения

Jupyter Notebook – интерактивная оболочка для ряда языков программирования (Julia, Python, R), позволяющая объединить код, текст, изображения, графики, в один документ и распространять его для других пользователей с сохранением возможности интерактивных «перевычислений».

Общие положения

Jupyter Notebook – интерактивная оболочка для ряда языков программирования (Julia, Python, R), позволяющая объединить код, текст, изображения, графики, в один документ и распространять его для других пользователей с сохранением возможности интерактивных «перевычислений».

Jupyter Notebook является развитием *IPython Notebook*, поддерживает подмножество MarkDown – для форматирования текста и LaTeX – для вывода математических формул. Основной элемент это Ячейка (допускает независимое «перевычисление»), их совокупность составляет документ.

Примечание. На подобном принципе отображения информации в виде последовательности взаимосвязанных ячеек построен также Wolfram Mathematica Notebook.

Общие положения

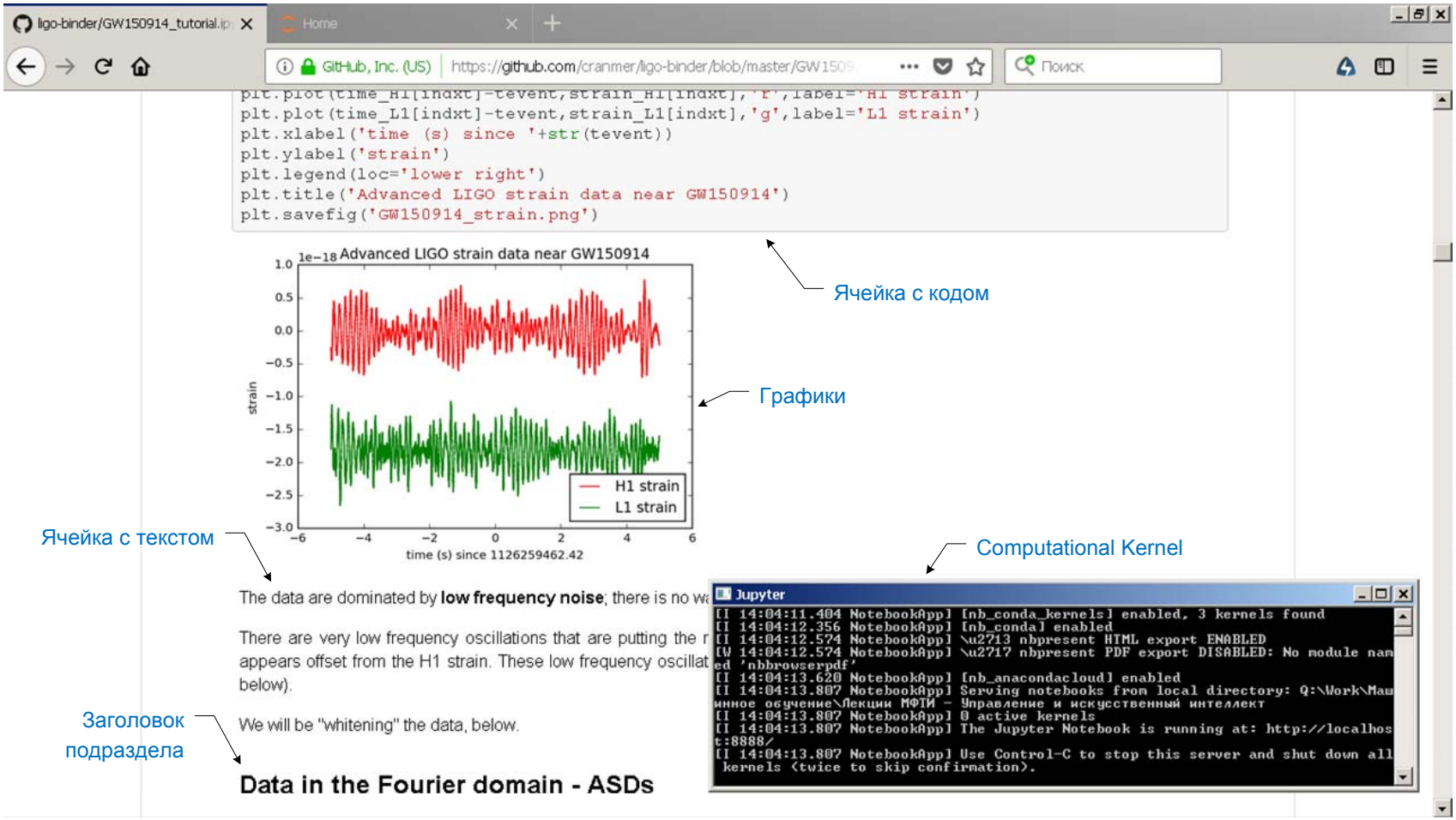
Jupyter Notebook – интерактивная оболочка для ряда языков программирования (Julia, Python, R), позволяющая объединить код, текст, изображения, графики, в один документ и распространять его для других пользователей с сохранением возможности интерактивных «перевычислений».

Jupyter Notebook является развитием *IPython Notebook*, поддерживает подмножество MarkDown – для форматирования текста и LaTeX – для вывода математических формул. Основной элемент это Ячейка (допускает независимое «перевычисление»), их совокупность составляет документ.

Примечание. На подобном принципе отображения информации в виде последовательности взаимосвязанных ячеек построен также Wolfram Mathematica Notebook.

Jupyter Notebook построен по клиент-серверной архитектуре: документ редактируется и отображается в web браузере, обработка ведётся в вычислительном ядре, которое может быть как локальным, так и удалённым.

Пример рабочего окна



Пример: GW150914_tutorial

Дополнительные моменты

- Есть возможность выбора web браузера, через редактирование конфигурационного файла jupyter_notebook_config.py

  import webbrowser

  webbrowser.register('firefox', None, webbrowser.GenericBrowser('firefox.exe'))

  c.NotebookApp.browser = 'firefox'

- Есть возможность задания рабочей директории для файлов проекта через редактирование ярлыка (пример для MS Win 7):

  Объект: "C:/Anaconda3/Scripts/jupyter.exe notebook"

  Рабочая папка: "G:/Projects/DL/Test_1"

- Для Jupyter Notebook доступны расширения Jupyter notebook extensions

- Расширение RISE – презентации в Jupyter Notebook. См. пример.

- Модуль nbconvert – преобразование Jupyter Notebook в PDF (отчёты), HTML (посты).

Общие положения

NumPy – это библиотека с открытым исходным кодом для высокоэффективных операций («cycle free») и математических вычислений над многомерными массивами (объект **ndarray**). Дополнительно поддерживаются: файловый ввод-вывод, вызов C/C++ функций.

Документация: NumPy Manual.

Соглашение об импорте: `import numpy as np`.

Особенности:

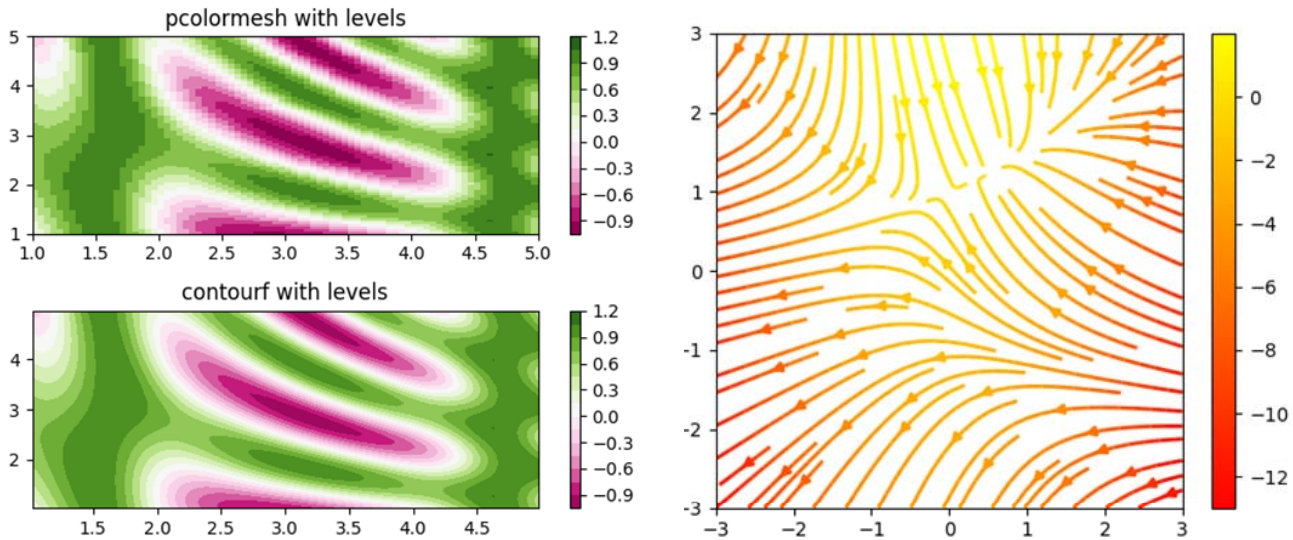
- В отличии от нативных списков Python, массивы NumPy имеют фиксированный (выровненный) размер, элементы массива имеют фиксированный тип.
- Основная парадигма – индексация и слайсинг: `x[:, 1]`, `x[::-1, :]`.
- Поддерживается конвейер (pipeline):
  `x = np.arange(9).reshape(3, 3).sum(axis = 1)`.

Общие положения

MatPlotLib – это библиотека для визуализации данных посредством построения 2D графиков. При построении графиков используется объектно-ориентированная нотация. Получаемые изображения являются векторными и могут быть экспортированы в ряд форматов (SVG, EPS, PDF, TIFF, PNG, и т.д.).

Документация: MatPlotLib Overview.

Соглашение об импорте: `import matplotlib.pyplot as plt`.



Источник: MatPlotLib Gallery

Общие положения

SciPy – это библиотека с открытым исходным кодом, предназначенная для выполнения научных и инженерных расчётов. Построена по модульному принципу. Основная структура данных – массив **ndarray**. Имеет развитые возможности ввода-вывода данных в различные форматы файлов.

Документация: SciPy Reference Guide.

Соглашение об импорте: `import scipy as sc`.

Основные модули, востребованные в программе курса:

- **fftpack** – Fourier Transforms.
- **signal** – Signal Processing.
- **linalg** – Linear Algebra.
- **stats** – Statistics.
- **io** – File IO.

Общие положения

Pandas – библиотека для обработки и анализа данных, функционирует поверх NumPy и предоставляет две специализированные структуры данных верхнего уровня: **`Series`** – 1D временные ряды и **`DataFrame`** – 2D таблицы (аналог **`data.frame`** из языка R). Основное назначение библиотеки: индексация (в том числе иерархическая) и манипулирование (переформатирование, вставка, удаление, выборка, срез и т.п.) многомерными массивами данных. Имеет развитые возможности ввода-вывода данных в различные форматы файлов и SQL СУБД.

Документация: Pandas documentation.

Соглашение об импорте: `import pandas as pd`.

Особенности:

- В отличии от массивов NumPy **`DataFrame`** допускает для столбцов разнородный тип.
- Доступ к ячейке **`DataFrame`** через именования:
  `A["name_col"].loc("name_row")`.
- Для индексов поддерживаются временные метки с дискретой в 1 нс (тип данных NumPy **`datetime64`**).

Общие положения

Scikit-learn – это библиотека алгоритмов машинного обучения и интеллектуального анализа данных, функционирует поверх NumPy и SciPy. Построена по модульному принципу. Основная структура данных – массив **ndarray**. Имеет развитые возможности ввода-вывода данных в различные форматы файлов.

Документация: User Guide.

Соглашение об импорте: `import sklearn as sk`.

Основные модули, востребованные в программе курса:

- **datasets** – Embedded Dataset.
- **metrics** – Model Evaluation.
- **linear_model**, **naive_bayes**, **neighbors**, **svm**, **tree** – Sup. Alg.
- **cluster** – Unsupervised Clustering Algorithms.
- **decomposition** – Matrix Decomposition.
- **manifold** – Manifold Learning.

Общие положения

Эффективная работа в экосистеме Python, в части интеллектуального анализа данных и машинного обучения, предполагает также хорошее знание следующих дополнительных модулей:

- H5Py – ввод-вывод данных в файлы формата HDF5. Документация. Соглашение об импорте: `import h5py as h5`.
- SQLite – простая и эффективная SQL СУБД. Документация. Соглашение об импорте: `import sqlite as sq`.
- OpenCV – мощная библиотека машинного зрения. Документация. Соглашение об импорте: `import cv2 as cv`.

Outline section

Краткий перечень датасетов

MNIST – коллекция рукописных цифр (размер изображения 28x28 пикселей) состоит из тренировочного (60K образцов) и тестового (10K образцов) наборов. Web-сайт.

AudioSet – коллекция вручную размеченных YouTube видеороликов (2.1 млн.) с выделением временных интервалов и звуковых событий (527 классов). Всего 5.8K часов аннотированного аудио-контента. Web-сайт.

CSTR VCTK Corpus – коллекция речевых данных, произнесённых 109 носителями английского языка с различными акцентами. Каждый из дикторов читает около 400 предложений. Web-сайт.

HolStep – коллекция текстовых фрагментов (2 013 046 тренировочные, 196 030 тестовые) предназначенных для разработки алгоритмов машинного обучения направленных на доказательство формальных утверждений (теорем). Web-сайт.

Outline section

Обозначения и соглашения – объекты

скаляр: a – число, $\dim a = 0$.

вектор: $\mathbf{a}$ – одномерный массив, $\dim \mathbf{a} = N$.

единичный вектор: $\mathbf{1}$ – одномерный массив, $\dim \mathbf{1} = N$.

вектор-столбец, - строка: $\dim \mathbf{a}_| = N \times 1$, $\dim \mathbf{a}^| = 1 \times N$,

$$\mathbf{a}_| = \begin{bmatrix} a_1 \\ a_2 \\ \dots \end{bmatrix}, \quad \mathbf{a}^| = \begin{bmatrix} a_1 & a_2 & \dots \end{bmatrix}.$$

матрица: $\mathbf{A}$ – двумерный массив, $\dim \mathbf{A} = N \times M$.

единичная матрица: $\mathbf{E}$ – двумерный массив, $\dim \mathbf{E} = N \times N$,

$$\mathbf{E} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix} = \operatorname{diag} \mathbf{1}.$$

«тензор»: $\mathbf{A}$ – N-мерный массив, $\dim \mathbf{A} = N_1 \times N_2 \times \dots$.

множество, пространство: S.

Обозначения и соглашения – операции

сложение/вычитание: $\mathbf{a} \pm \mathbf{b}$, $\mathbf{A} \pm \mathbf{B}$.

покомпонентное умножение: $\mathbf{a} * \mathbf{b}$, $\mathbf{A} * \mathbf{B}$.

скалярное умножение: $\mathbf{a} \cdot \mathbf{b}$.

матричное умножение: $\mathbf{A}\,\mathbf{B}$, $\mathbf{a}\,\mathbf{b}$.

внешнее умножение: $\mathbf{a} \otimes \mathbf{b}$, $\mathbf{A} \otimes \mathbf{B}$.

декартово произведение $\mathrm{S} \times \mathrm{T}$.

транспонирование: $\mathbf{a}^\top$, $\mathbf{B}^\top$.

комплексное сопряжение: $\mathbf{a}^*$, $\mathbf{B}^*$.

эрмитово сопряжение: $\mathbf{a}^{\dagger}$, $\mathbf{B}^\dagger$.

обратный вектор, матрица: $\mathbf{a}^{-1}$, $\mathbf{B}^{-1}$.

норма вектора, матрицы: $|\mathbf{a}|$, $\|\mathbf{A}\|$.

детерминант матрицы: $|\mathbf{A}|$.

мощность множества: $|\mathrm{V}|$.

Сложение, умножение

сложение/вычитание:

$$\mathbf{a} \pm \mathbf{b} = \begin{bmatrix} a_1 \pm b_1 \\ a_2 \pm a_2 \end{bmatrix}, \quad \mathbf{A} \pm \mathbf{B} = \begin{bmatrix} a_{11} \pm b_{11} & a_{12} \pm b_{12} \\ a_{21} \pm b_{21} & a_{22} \pm b_{22} \end{bmatrix}.$$

покомпонентное умножение:

$$\mathbf{a} * \mathbf{b} = \begin{bmatrix} a_1\, b_1 \\ a_2\, a_2 \end{bmatrix}, \quad \mathbf{A} * \mathbf{B} = \begin{bmatrix} a_{11}\, b_{11} & a_{12}\, b_{12} \\ a_{21}\, b_{21} & a_{22}\, b_{22} \end{bmatrix}.$$

скалярное умножение:

$$\mathbf{a} \cdot \mathbf{b} = a_1\, b_1 + a_2\, b_2, \quad \mathbf{a}^{|} \cdot \mathbf{b}_{|} = \begin{bmatrix} a_1\, b_1 + a_2\, b_2 \end{bmatrix}.$$

матричное умножение:

$$\mathbf{A}\,\mathbf{B} = \begin{bmatrix} a_{11}\, b_{11} + a_{12}\, b_{21} & a_{11}\, b_{12} + a_{12}\, b_{22} \\ a_{21}\, b_{11} + a_{22}\, b_{21} & a_{21}\, b_{12} + a_{22}\, b_{22} \end{bmatrix}, \quad \mathbf{A}\,\mathbf{B} \neq \mathbf{B}\,\mathbf{A}.$$

внешнее умножение:

$$\mathbf{a}_{|} \otimes \mathbf{b}^{|} = \begin{bmatrix} a_1\, b_1 & a_1\, b_2 \\ a_2\, b_1 & a_2\, b_2 \end{bmatrix}.$$

Транспонирование, комплексное сопряжение, обращение

транспонирование:

$$\mathbf{a}^\top = \begin{bmatrix} a_1 & a_2 \end{bmatrix}, \quad \mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}.$$

комплексное сопряжение:

$$\begin{bmatrix} a_{11} + \mathrm{i}\, b_{11} & a_{12} + \mathrm{i}\, b_{12} \\ a_{21} + \mathrm{i}\, b_{21} & a_{22} + \mathrm{i}\, b_{22} \end{bmatrix}^* = \begin{bmatrix} a_{11} - \mathrm{i}\, b_{11} & a_{12} - \mathrm{i}\, b_{12} \\ a_{21} - \mathrm{i}\, b_{21} & a_{22} - \mathrm{i}\, b_{22} \end{bmatrix}.$$

эрмитово сопряжение:

$$\begin{bmatrix} a_{11} + \mathrm{i}\, b_{11} & a_{12} + \mathrm{i}\, b_{12} \\ a_{21} + \mathrm{i}\, b_{21} & a_{22} + \mathrm{i}\, b_{22} \end{bmatrix}^\dagger = \begin{bmatrix} a_{11} - \mathrm{i}\, b_{11} & a_{21} - \mathrm{i}\, b_{21} \\ a_{12} - \mathrm{i}\, b_{12} & a_{22} - \mathrm{i}\, b_{22} \end{bmatrix}.$$

симметрические, ортогональные, эрмитовы и унитарные матрицы:

$$\mathbf{A} = \mathbf{A}^\top, \quad \mathbf{B}^\top \mathbf{B} = \mathbf{E}, \quad \mathbf{C} = \mathbf{C}^\dagger, \quad \mathbf{C}^\dagger \mathbf{C} = \mathbf{E}.$$

обратный вектор, матрица:

$$\mathbf{a}^{-1} \cdot \mathbf{a} = 1, \quad \mathbf{A}^{-1} \mathbf{A} = \mathbf{E}.$$

Собственные значения, базис, SVD-разложение, нормы

собственное значение:

$$\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}.$$

ортонормированный набор базисных векторов:

$$\text{если}\ \mathbf{A}^{-1} = \mathbf{A}^{\dagger},$$

тогда столбцы (или строки) матрицы $\mathbf{A}$ образуют базис.

SVD-декомпозиция:

$$\text{если}\ \mathbf{B} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{\dagger}, \quad \dim \mathbf{B} = M \times N,$$

тогда унитарные матрицы $\mathbf{U}$ ($\dim \mathbf{U} = M \times M$) и $\mathbf{V}$ ($\dim \mathbf{V} = N \times N$) содержат левые и правые сингулярные вектора $\mathbf{B}$, соответственно, а $\operatorname{diag} \mathbf{S}$ – сингулярные числа $\mathbf{B}$. SVD-разложение имеет широкую применимость в задачах анализа данных и машинного обучения.

нормы векторов и матриц:

$$L_1, \quad L_2, \quad L_\infty.$$

Outline section

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 $y = X(\mathrm{A})$.

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:

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

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:

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

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 $X \leqslant x$, рассматриваемая как функция параметра x:

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

Базовые определения III

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

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

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:

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

Базовые определения III

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

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

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:

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

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

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

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

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

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:

$$P(N,\, n) = \mathrm{C}_N^n\, p^n\, (1-p)^{N-n}.$$

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 касаемо 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 касаемо 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 $0 \leqslant p \leqslant 1$. 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:

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

Найти:

Оценки моментов случайной величины.

Оценивание эмпирической функции распределения с.в.

Дано:

$\mathrm{X}^* = \{x_1, x_2, \ldots, x_N\}$ – 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:

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

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

Найти:

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

Проверка статистических гипотез

Дано:

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

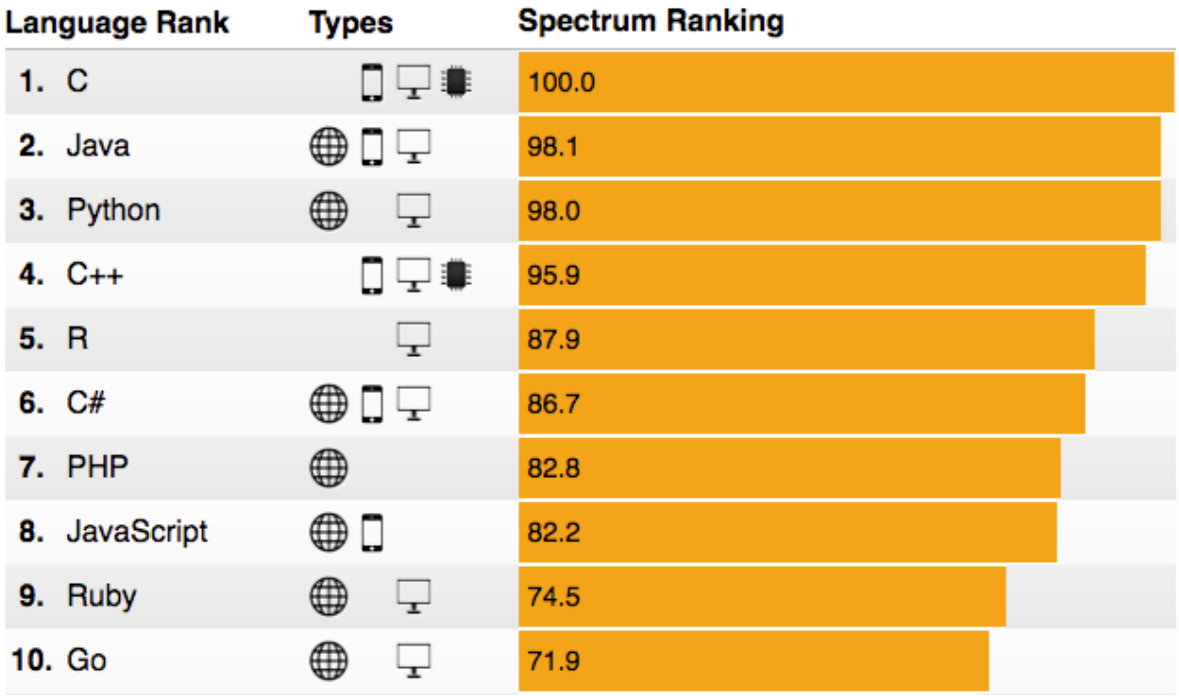
$F(x)$ – теоретическая функции распределения случайной величины x.

Найти:

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

Почему R?

IEEE 2016 Top Programming Languages



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

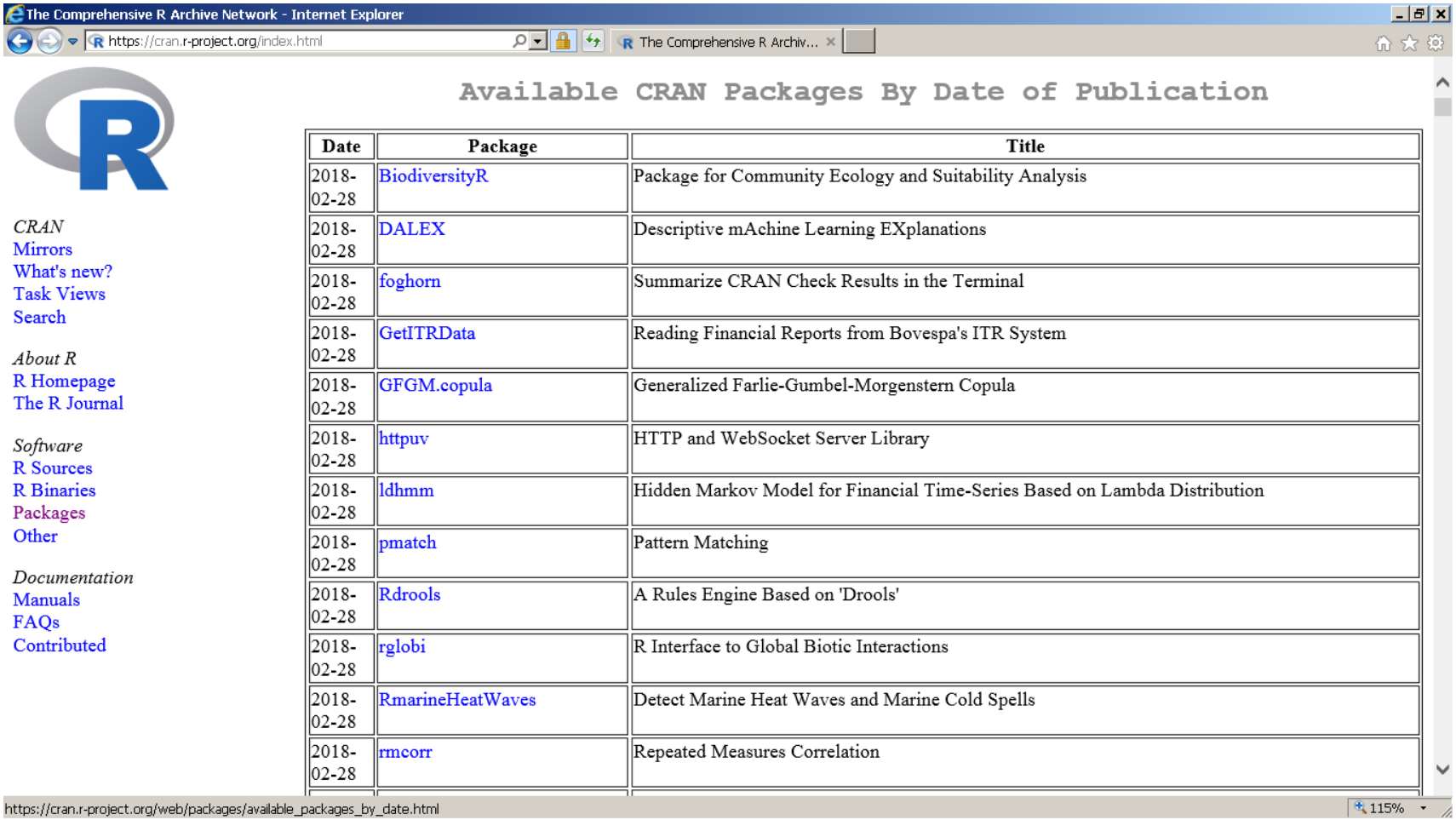
Парадигмы программирования:

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

Особенности языка:

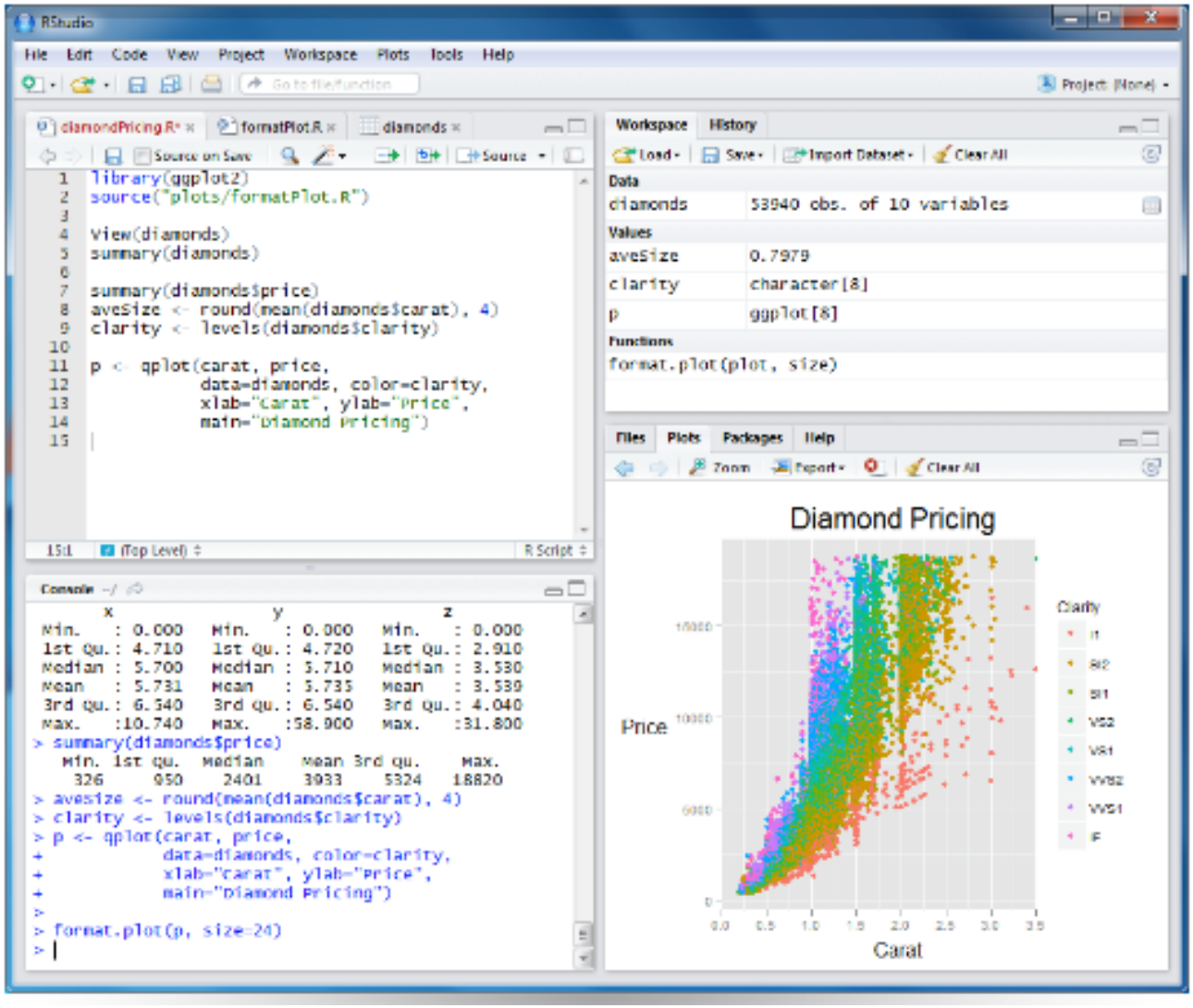
- Язык предметной области (DSL) для обработки данных, высокоуровневый со встроенными высокоуровневыми структурами данных.
- Реализован поверх классической архитектуры интерпретатора-компилятора Scheme.
- Интерпретируемый (поддерживает REPL среду).
- Динамическая типизация, автоматическое управление памятью.
- 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



Официальные пакеты расширения: 12 220

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 $\mathbf{T}$, и перед их подачей на вход моделей машинного обучения, как правило, требуется проведение ряда преобразований:

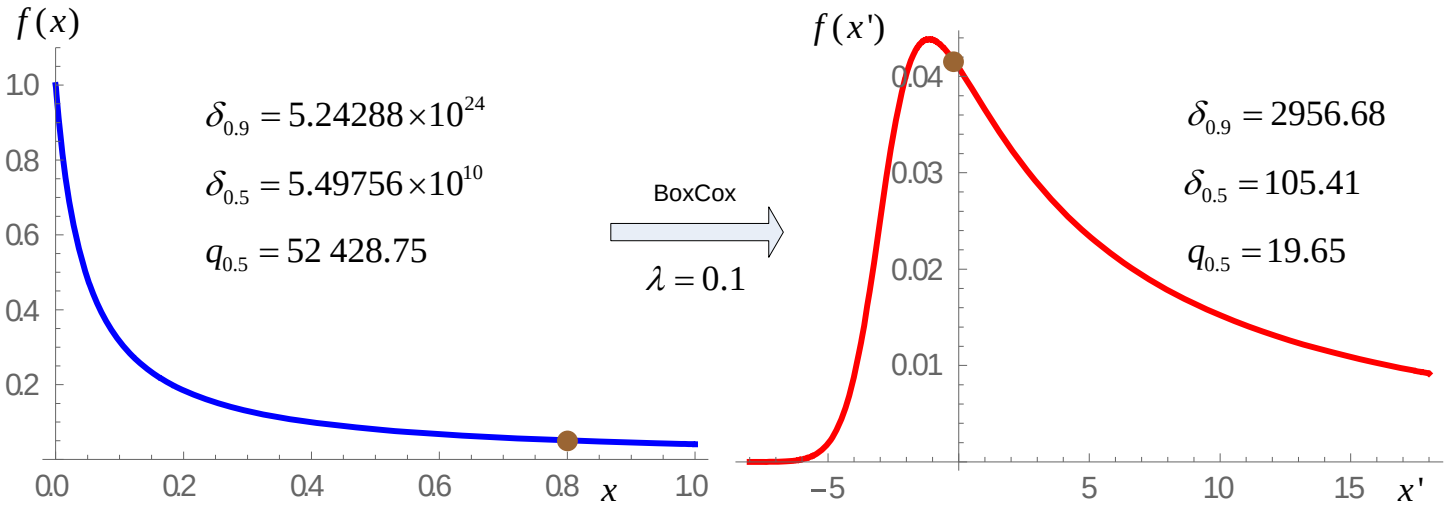
Формирование признаков

При формировании из «сырых» данных массива информативных признаков $\mathbf{T}$, 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:

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

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



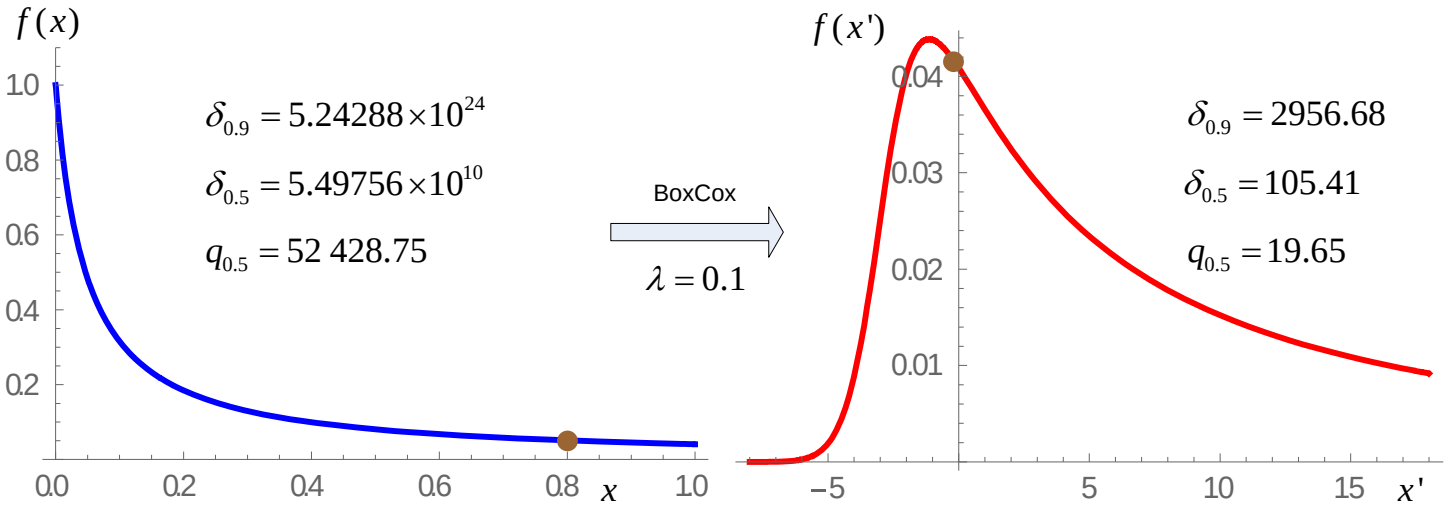
Формирование признаков

При формировании из «сырых» данных массива информативных признаков $\mathbf{T}$, 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:

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

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



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

Outline section

Контрольная работа

Задание для слушателей:

1. Скачать датасет MNIST. Оформить Jupyter Noteebook с кодом на Python: (i) – функция чтения файлов, через `np.fromfile()`, с цифрами и метками; (ii) – функция визуализации, через `plt.imshow()`, трёх случайных цифр. 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 цифр должен иметь структуру `A[k,i,j]`, где `k` – индекс изображения цифры, `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 Noteebook: (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 Noteebook: (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.