Машинное обучение: вводный семинар

А.В. Зухба
a__l@mail.ru

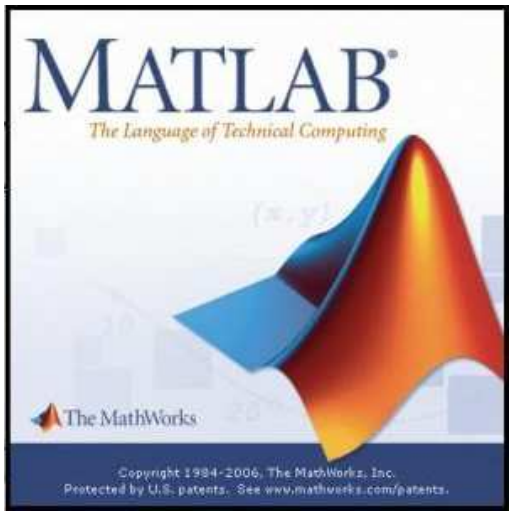
февраль 2016

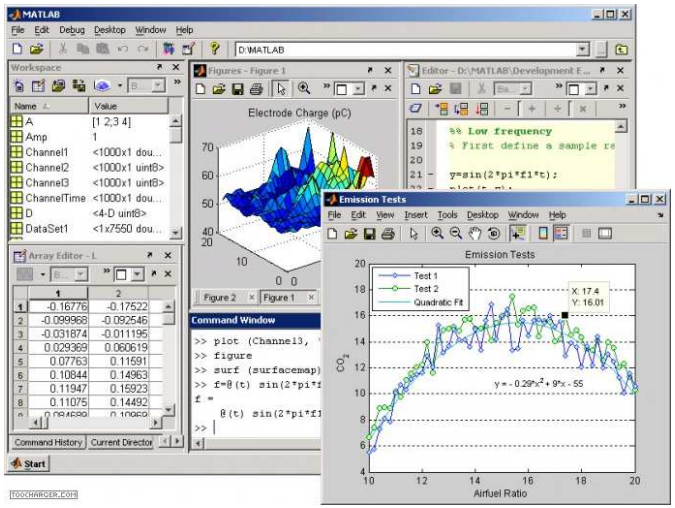
Что понадобится?

- Линейная алгебра
- Комбинаторика
- Методы оптимизации
- Теория вероятности и математическая статистика
- Оценки вычислительной сложности
- Теория графов
- Инструменты для работы с данными

MATLAB

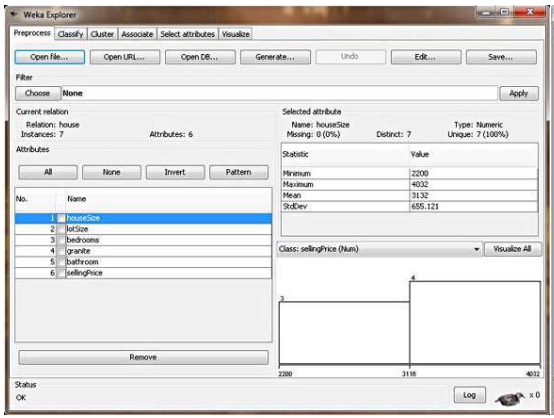
- Язык программирования и среда для матричных вычислений
- Огромный функционал вне машинного обучения
- Нет многих инструментов для машинного обучения

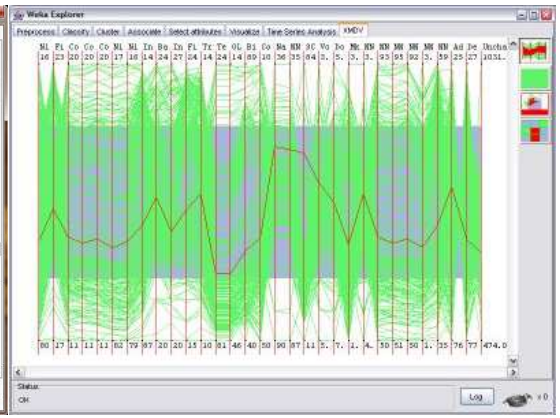




WEKA

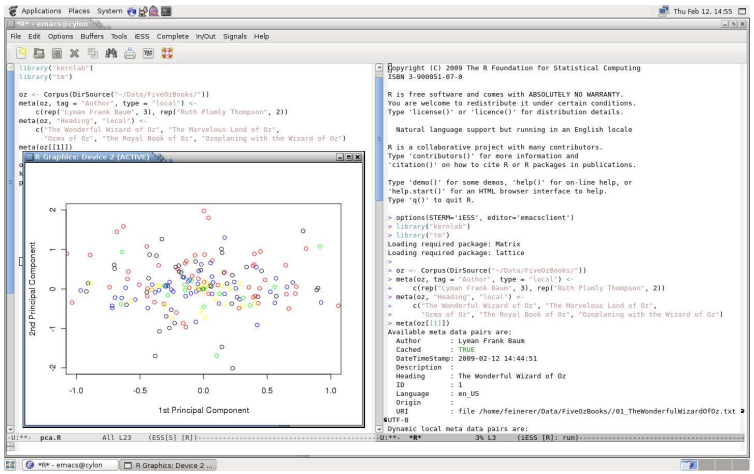
- Графический интерфейс
- Простота использования
- Библиотека алгоритмов на Java
- Не получится реализовать сложные алгоритмы

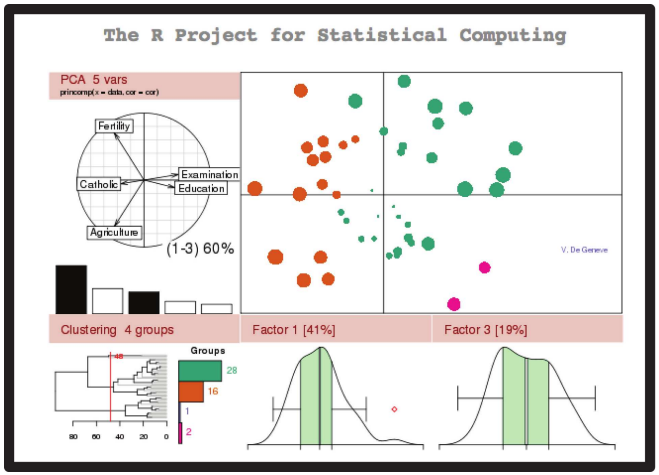




R

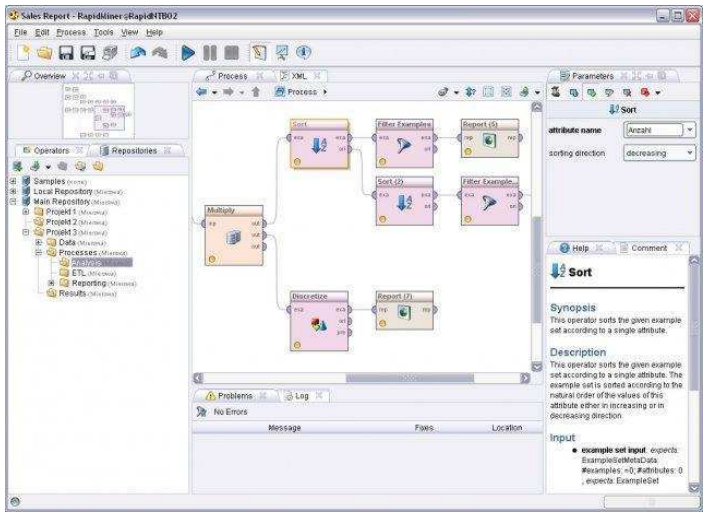
- Язык программирования, созданный для статистического анализа данных
- Богатая библиотека, большие возможности

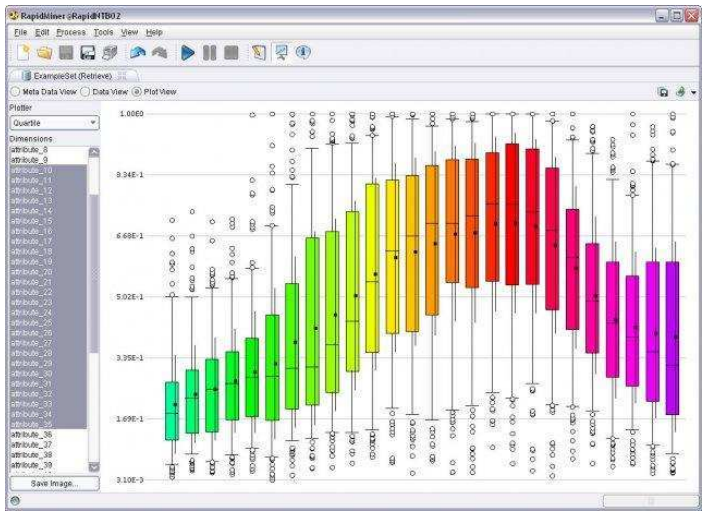




RAPIDMINER

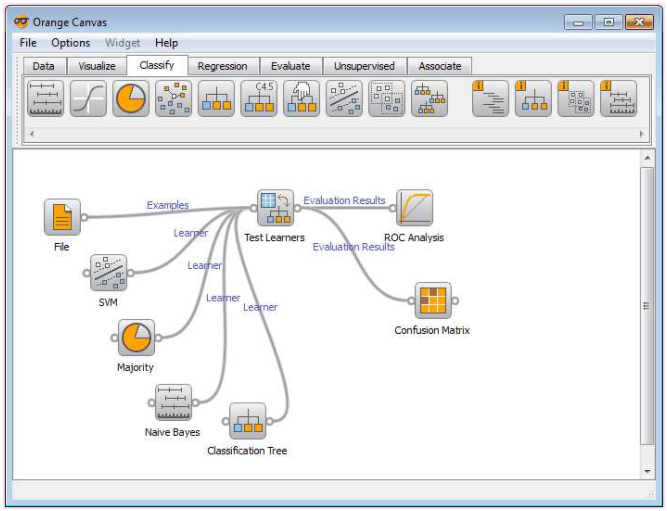
- Графический интерфейс
- На базе Weka и R
- Значительно больше возможностей чем в Weka

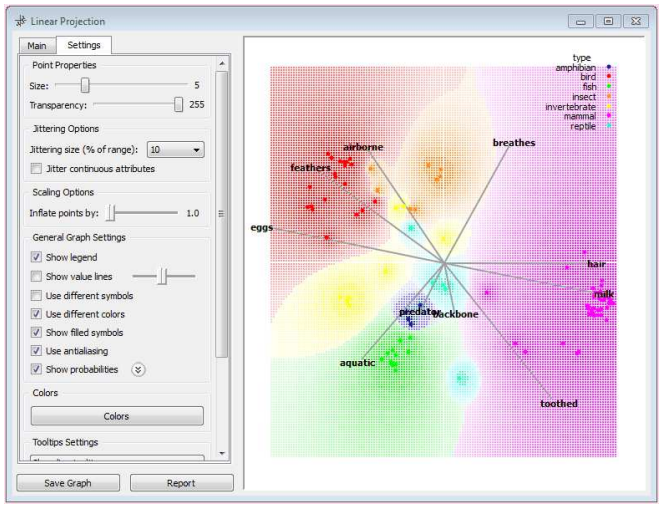




Orange

- Графический интерфейс
- Предоставляет библиотеку алгоритмов на Python





Python

- Удобные язык программирования общего назначения, т.е. можно решать широкий круг задач
- Очень быстро развивается и набирает популярность
- Много быстро развивающихся библиотек
- Пока есть не все, что есть, например, в R или Matlab
- Имеет интерфейсы почти ко всему
- High-level — Low-level programming
- Масштабируемость

julia

- Современный производительный язык для научных вычислений
- Имеет интерфейсы к некоторых библиотеками Python

	Fortran gcc 4.8.1	Julia 0.2	Python 2.7.3	R 3.0.2	Matlab R2012a	Octave 3.6.4	Mathematica 8.0	JavaScript V8 3.7.12.22	Go go1
fib	0.26	0.91	30.37	411.36	1992.00	3211.81	64.46	2.18	1.03
parse_int	5.03	1.60	13.95	59.40	1463.16	7109.85	29.54	2.43	4.79
quicksort	1.11	1.14	31.98	524.29	101.84	1132.04	35.74	3.51	1.25
mandel	0.86	0.85	14.19	106.97	64.58	316.95	6.07	3.49	2.36
pi_sum	0.80	1.00	16.33	15.42	1.29	237.41	1.32	0.84	1.41
rand_mat_stat	0.64	1.66	13.52	10.84	6.61	14.98	4.52	3.28	8.12
rand_mat_mul	0.96	1.01	3.41	3.98	1.10	3.41	1.16	14.60	8.51

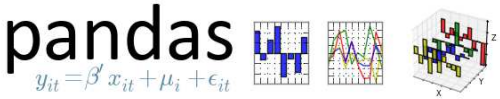
Базовые математические библиотеки

Базовые матричные и общие математические операции

Дополнительный математический функционал

Визуализация данных

Библиотеки для анализа данных



Чтение и предобработка данных

Машинное обучение

Статистический анализ

Расширения Python

IPython Interactive Computing

Упрощение научных исследований на Python. Включает в себя IPython Notebook для интерактивных отчетов в браузере

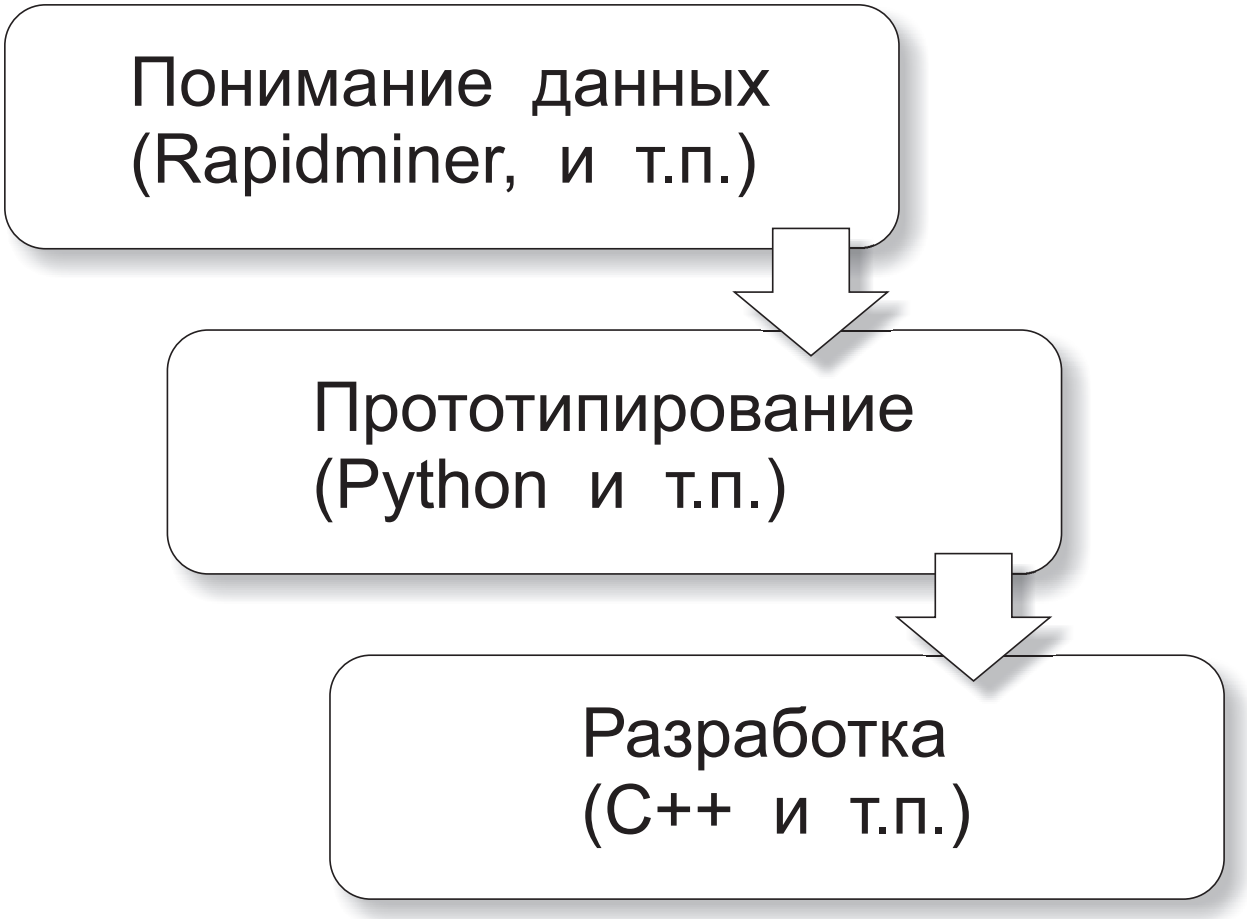
Cython

Язык программирования, упрощающий написание вычислительно затратных частей на C/C++

RPy

Использование R прямо в Python скриптах

Промышленная разработка



Python

- Интерпретируемый
- Строгая динамическая типизация
- Кроссплатформенный
- Мультипарадигменный
- Простой в понимании и написании
- Медленнее компилируемых языков
- Легко интегрируется с C/C++

Zen of Python

- Красивое лучше уродливого
- Явное лучше неявного
- Простое лучше сложного
- Сложное лучше усложненного
- Плоское лучше вложенного
- Разреженное лучше плотного
- Читаемость важна
- Частные случаи недостаточно частные, чтобы нарушать правила
- Хотя практичность превыше чистоты, ошибки не должны подавляться, если только не указано обратное
- В случае неопределенности не поддавайся соблазну угадывания
- Должен быть один, а лучше только один, очевидный путь решения, хотя этот путь может быть неочевиден с первого взгляда, если ты не обладаешь нестандартным мышлением
- Сейчас лучше, чем никогда, хотя лучше никогда, чем прямо сейчас
- Если реализацию идеи сложно объяснить, это плохая идея
- Если реализацию идеи легко объяснить, возможно, идея хорошая