RocketScience.ai

TENSOR.BY

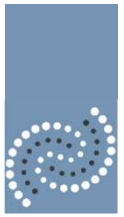
Курсы по машинному обучению

Продолжительность: 9 недель

Темы:

1. ***Введение в курс. Подготовка данных для моделирования в Python с использованием пакетов Numpy и Pandas***
2. ***Обучение с учителем - Регрессия. Метрики оценки качества моделей. Валидация моделей.***
3. ***Обучение с учителем - Классификация. Модели классификации: линейная модель и модели на основе ансамблей разрешающих деревьев. Метрики.***
4. ***Анализ и прогнозирование временных рядов. Метрики. "Наивные" модели. Семейство моделей ARIMA.***
5. ***Обучение без учителя. Кластерный анализ. Задачи и оценка качества кластеризации.***
6. ***Моделирование и анализ текстовой информации. Задачи Natural Language Processing. Способы представления текста в моделировании.***
7. ***Нейронные сети. Виды нейронных сетей. Глубокое обучение. Алгоритм обратного распространения ошибки. Фреймворк Keras.***
8. ***Рекуррентные нейронные сети и их применение. LSTM. Сверточные нейронные сети. Операции “свертка” и “пулинг”. Эмбеддинги.***
9. ***Защита выпускных проектов.***

Проекты



Датасеты :

- *https://www.kaggle.com/datasets*
- *Google Dataset Search https://toolbox.google.com/datasetsearch*

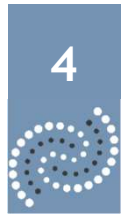
Примеры:

- *https://www.kaggle.com/zynicide/wine-reviews*
- *https://www.kaggle.com/secareanualin/football-events*
- *https://www.kaggle.com/wosaku/crime-in-vancouver/kernels*
- *https://www.kaggle.com/rounakbanik/the-movies-dataset*

Примерные требования к датасету:

- *подходят для задачи обучения с учителем,*
- *есть текстовые данные,*
- *большой объем, но не огромный*

Защита проектов



Примерный план презентации проекта:

1) постановка задачи;
2) подготовка данных, создание переменных;
3) моделирование (модели, гридсерч гиперпараметров, валидация);
4) сравнение рез-тов моделирования (метрик);
5) для разных моделей, выбор лучшей модели;
6) применение модели для прогнозирования.

RocketScience.ai

TENSOR.BY

ML-course

1. Preparing data for modeling in Python using Numpy and Pandas packages

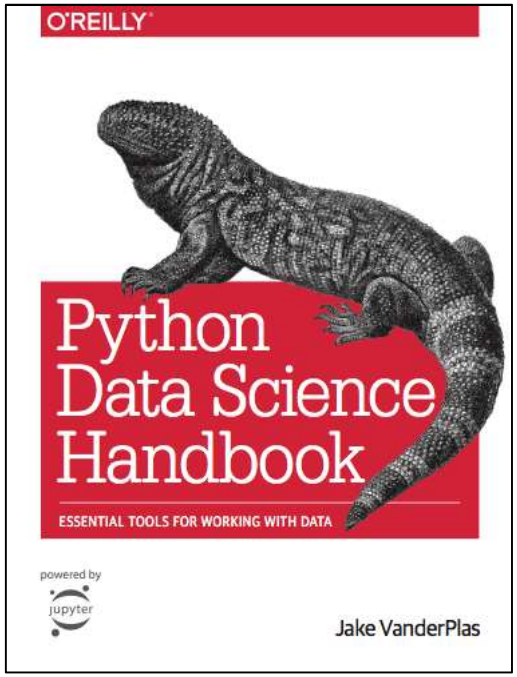
Kate Miniukovich (Data Scientist),
miniukovich@rocketscience.ai

Reference

Jake VanderPlas
Python Data Science Handbook

https://jakevdp.github.io/PythonDataScienceHandbook/

(available online for free)



Packages

NumPy

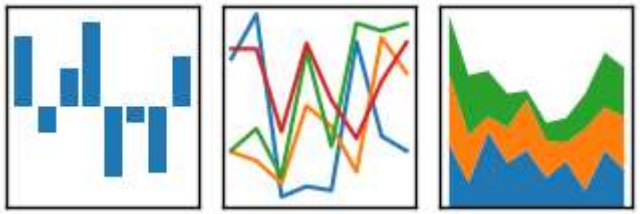
http://www.numpy.org/

http://www.numpy.org/devdocs/user/quickstart.html

Pandas

https://pandas.pydata.org/

http://pandas-docs.github.io/pandas-docs-travis/api.html



Why NumPy & Pandas?

	area	population	density
California	423967.0	38332521	90.413926
Florida	170312.0	19552860	114.806121
Illinois	149995.0	12882135	85.883763
New York	141297.0	19651127	139.076746
Texas	NaN	26448193	NaN

```
1 L3 = [True, "2", 3.0, 4]
2 [type(item) for item in L3]
```

```
[bool, str, float, int]
```

	title	year	name	type	character	n
1813168	12 Years a Slave	2013	Brad Pitt	actor	Bass	53.0
1813169	A River Runs Through It	1992	Brad Pitt	actor	Paul Maclean	2.0
1813170	Abby Singer	2003	Brad Pitt	actor	Himself	51.0
1813171	Across the Tracks	1990	Brad Pitt	actor	Joe Maloney	2.0
1813174	Babel	2006	Brad Pitt	actor	Richard	1.0
1813175	Being John Malkovich	1999	Brad Pitt	actor	Brad Pitt	NaN
1813176	Burn After Reading	2008	Brad Pitt	actor	Chad Feldheimer	3.0
1813177	By the Sea	2015	Brad Pitt	actor	Roland	1.0
1813178	Confessions of a Dangerous Mind	2002	Brad Pitt	actor	Brad, Bachelor #1	37.0
1813179	Cool World	1992	Brad Pitt	actor	Detective Frank Harris	3.0

Objects

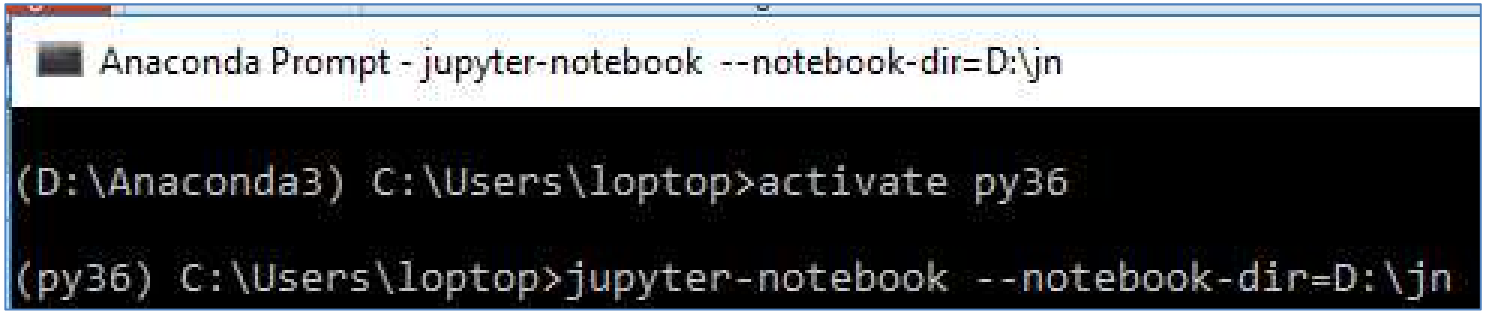
Ndarray (numpy.ndarray)

Series (pandas.core.series.Series)

DataFrame (pandas.core.frame.DataFrame)

- *.attribute*
- *.method()*
- *function()*

Practical part



Files: jn\

- data*.csv
- part1_numpy.jpynb, part2_pandas.jpynb
- Exercises-1.ipynb,…, Exercises-5.ipynb

Q & A

Thank you!