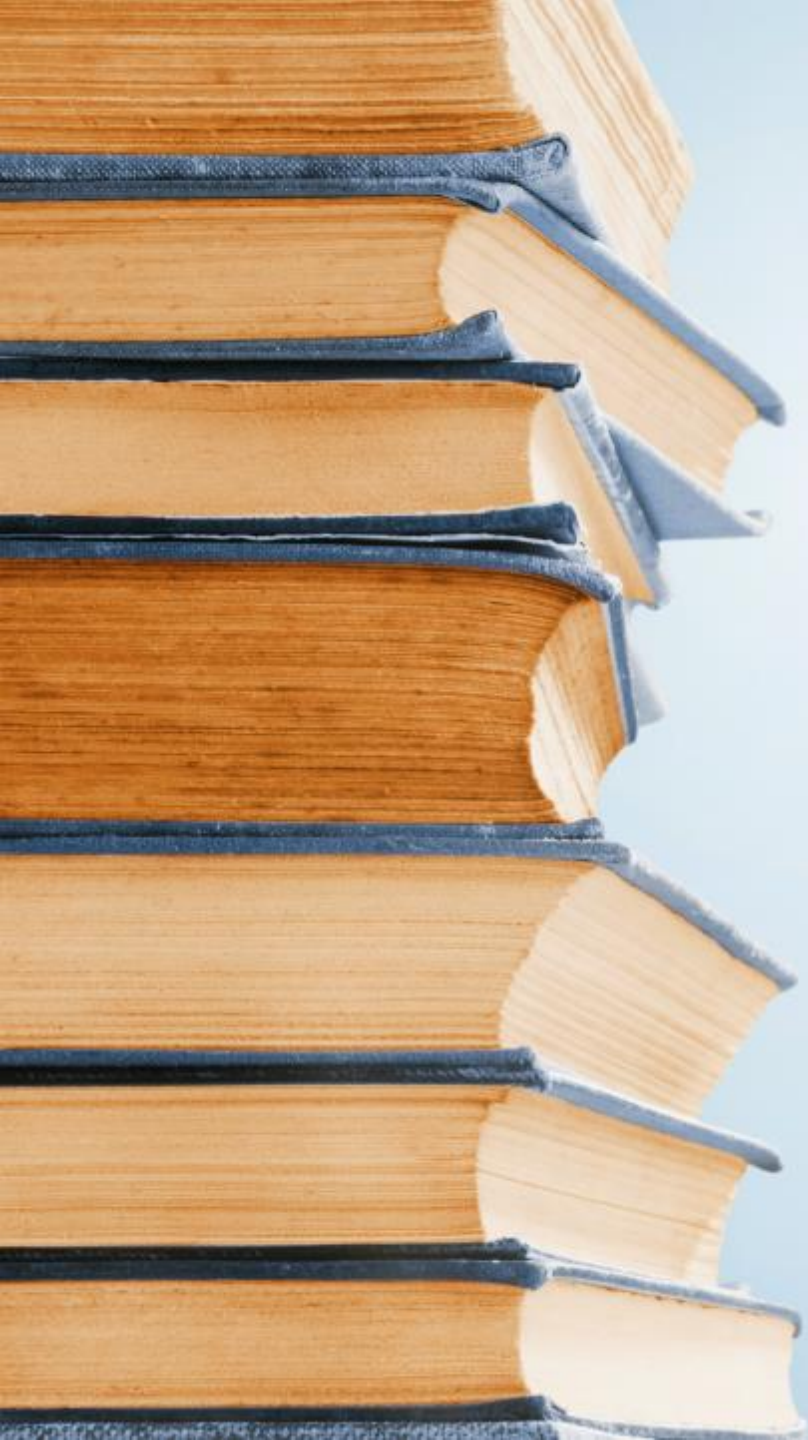


Библиотеки обработки данных для языка Python

ИУ-5

Краткий план лекции

- Структуры данных в машинном обучении
- Библиотеки для обработки данных
  - NumPy
  - Разреженные матрицы
  - Pandas
  - PandaSQL
  - Dask

Структуры данных в машинном обучении

Тензоры

- Тензор – основная структура данных в машинном обучении.
- Существует два определения тензора
  1. «Старое» определение. Объект линейной алгебры, линейно преобразующий элементы одного линейного пространства в элементы другого.
  2. **«Новое» определение, которое почти всегда используется в машинном обучении. Просто многомерная матрица.**
- Существует гипотеза, что большинство алгоритмов машинного обучения можно представить в виде последовательных преобразований тензоров. Гипотеза реализована в библиотеке TensorFlow.

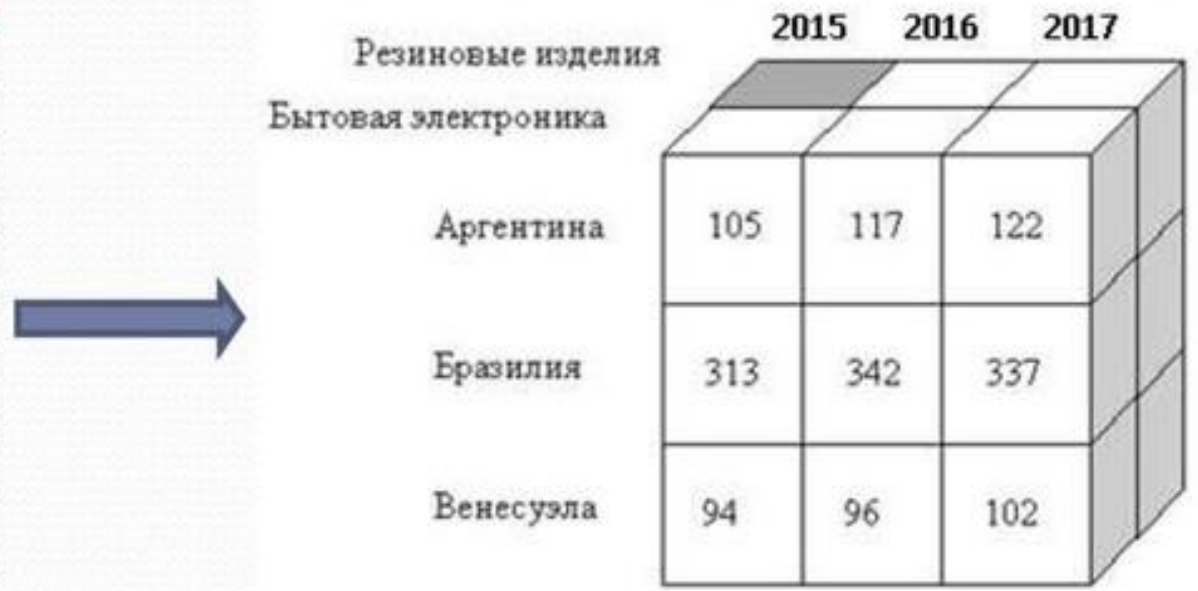
Факторизация тензоров

- Большинство тензоров очень разрежены (содержат много пустых значений).
- Факторизация тензора – его представление в виде произведения простых объектов (матрицы и тензоров меньшей размерности), которые не являются разреженными:
  - https://habr.com/ru/company/yandex/blog/313892/ (проф. Иван Оселедец, премия)
  - https://pdfs.semanticscholar.org/94cc/6daad548a03c6edb0351d686c2d4aa364634.pdf (проф. Andrzej Cichocki)
- Для реляционных БД аналогом операции факторизации является нормализация схемы БД.
  - В этом случае мы не всегда уменьшаем размер нормализованных таблицы, но всегда превращаем исходную денормализованную таблицу в набор «неразреженных» нормализованных таблиц.
  - Вместо произведения матриц (для факторизации) используется соединение на основе ключей (join).
- Таким образом, превращение денормализованной структуры в набор нормализованных составных частей (которые удобнее хранить и обрабатывать) является устойчивым «паттерном» анализа данных.
- Но для построения моделей машинного обучения удобнее использовать денормализованную таблицу (которая может быть результатом сборки факторизованного тензора или результатом соединения реляционных таблиц).

Тензоры и OLAP-кубы

- OLAP-куб является разновидностью тензора.
- Каждый OLAP-куб (справа) может быть представлен в виде денормализованной таблицы (слева). Дополнительный пример.

Страна	Товар	Год	Объем продаж
Аргентина	Бытовая электроника	1988	105
Аргентина	Бытовая электроника	1989	117
Аргентина	Бытовая электроника	1990	122
Аргентина	Резиновые изделия	1989	212
Аргентина	Резиновые изделия	1990	217
Бразилия	Бытовая электроника	1988	313
Бразилия	Бытовая электроника	1989	342
Бразилия	Бытовая электроника	1990	337
Бразилия	Резиновые изделия	1988	515
Бразилия	Резиновые изделия	1989	542
Бразилия	Резиновые изделия	1990	566
Венесуэла	Бытовая электроника	1988	94
Венесуэла	Бытовая электроника	1989	96
Венесуэла	Бытовая электроника	1990	102
Венесуэла	Резиновые изделия	1988	153
Венесуэла	Резиновые изделия	1989	147
Венесуэла	Резиновые изделия	1990	162



- Тензорное (кубическое) представление больше соответствует библиотеке NumPy, а денормализованное – библиотеке Pandas.

Тензоры и OLAP-кубы (2)

- Тензор является прежде всего математической моделью, в то время как OLAP-куб – инженерной.
- В отличие от OLAP-кубов для тензоров не определено естественных операций агрегирования.

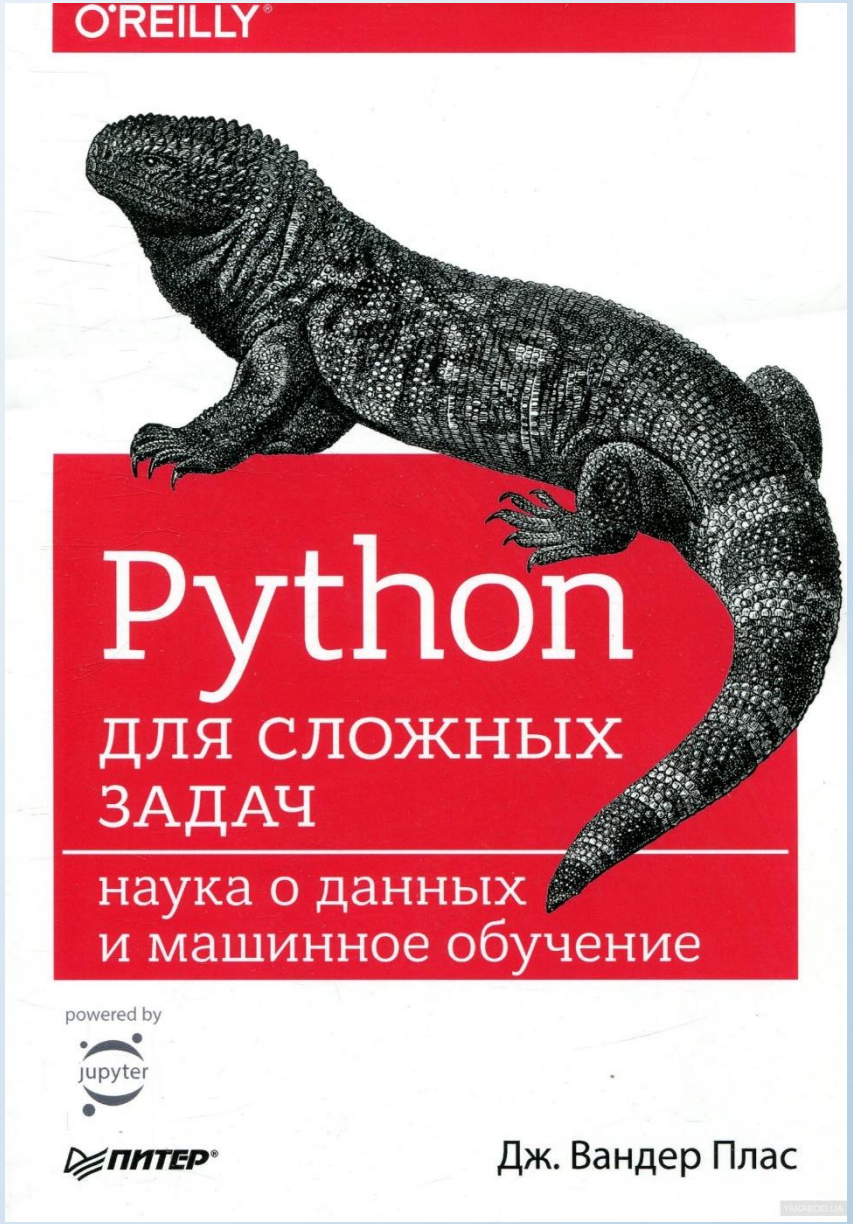
Выводы по разделу:

- Тензор – основная структура данных в машинном обучении, сейчас понимается как многомерная матрица.
- Тензор и OLAP-куб являются «родственными моделями». Для обоих моделей можно использовать срезы и уменьшение размерностей. Но в отличие от OLAP-кубов для тензоров не определено естественных операций агрегирования.
- Тензор и OLAP-куб могут быть представлены в форме денормализованной таблицы.
- Для построения моделей машинного обучения удобнее использовать денормализованную таблицу (которая может быть получена из тензора, OLAP-куба или как результат соединения реляционных таблиц).

Библиотеки для обработки данных

Рекомендуемая книга

- Библиотеки NumPy и Pandas хорошо описаны в книге Дж. Вандер Пласа.



Библиотека NumPy

- Библиотека NumPy предназначена для научных вычислений. Основным типом данных является тензор (многомерная матрица).
- Сайт библиотеки - http://www.numpy.org/
- Введение на русском языке - https://habr.com/ru/post/352678/
- Введение для начинающих (Kaggle) - https://www.kaggle.com/abdullahsahin/numpy-tutorial-for-beginner
- Введение - https://docs.scipy.org/doc/numpy/user/quickstart.html
- Документация - https://docs.scipy.org/doc/numpy/reference/index.html#reference
- Другие тьюториалы по NumPy:
  - Часть 1 - https://www.machinelearningplus.com/python/numpy-tutorial-part1-array-python-examples/
  - Часть 2 - https://www.machinelearningplus.com/python/numpy-tutorial-python-part2/
  - 101 упражнение по NumPy - https://www.machinelearningplus.com/python/101-numpy-exercises-python/
  - 100 упражнение по NumPy - http://www.labri.fr/perso/nrougier/teaching/numpy.100/
  - Сопряжение размерностей матриц (broadcasting) - https://docs.scipy.org/doc/numpy/user/basics.broadcasting.html

Разреженные матрицы

- Если набор данных слишком велик и содержит много пустых значений, то можно использовать разреженные матрицы:
  - https://docs.scipy.org/doc/scipy/reference/sparse.html
  - http://scipy-lectures.org/advanced/scipy_sparse/index.html
  - https://rushter.com/blog/scipy-sparse-matrices/ (простое введение)

Библиотека Pandas

- Pandas - это библиотека, предназначенная для чтения, хранения, записи и обработки наборов данных (датасетов).
- Основная структура данных в Pandas это денормализованная таблица данных. Pandas одновременно обладает некоторыми характеристиками электронной таблицы (Excel) и реляционной СУБД.
- В чем Pandas похож на электронную таблицу:
  - В Pandas основной структурой данных является одна таблица (а не схема связанных таблиц как в реляционной СУБД).
  - Таблица должна содержать полный набор данных, предназначенных для дальнейшего построения моделей машинного обучения.
  - Таблица данных, как правило, является денормализованной.
  - В данных могут быть пропущенные значения.
- В чем Pandas похож на реляционную СУБД:
  - Таблица данных (как и реляционная таблица) состоит из типизированных столбцов (атрибутов).
  - Строка таблицы соответствует записи в реляционной БД.
  - Над таблицами возможно выполнение операций реляционной алгебры – соединение (join), группировка и другие операции.

Библиотека Pandas

- Официальная документация - http://pandas.pydata.org/pandas-docs/stable/
- Русскоязычное руководство - https://khashtamov.com/ru/pandas-introduction/
- Введение для начинающих (Kaggle) - https://www.kaggle.com/abdullahsahin/step-by-step-pandas-tutorial-for-beginner
- Интерактивное руководство - https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python
- 101 упражнение по Pandas - https://www.machinelearningplus.com/python/101-pandas-exercises-python/
- 12 наиболее полезных техник Pandas - https://www.analyticsvidhya.com/blog/2016/01/12-pandas-techniques-python-data-manipulation/

Библиотека PandaSQL

- Предназначена для выполнения SQL-запросов над наборами данных Pandas.
- Официальный сайт - https://github.com/yhat/pandasql
- Библиотека использует синтаксис SQLite - https://www.sqlite.org/lang.html
- Примеры использования:
  - https://habr.com/ru/post/279213/
  - https://github.com/miptgirl/udacity_engagement_analysis/blob/master/pandasql_example.ipynb

Библиотека Dask

- Библиотека Dask предназначена для распараллеливания вычислений при обработке данных.
- Преимущества библиотеки - http://docs.dask.org/en/latest/why.html
- Сравнения с библиотеками для обработки больших данных:
  - http://docs.dask.org/en/latest/spark.html
  - https://matthewrocklin.com/blog//work/2018/08/28/dataframe-performance-high-level