

Machine Learning on Big Data

Шугаепов Ильнур
VK.com
ilnur.shug@gmail.com

январь 2020 г.

О чем курс?

Моделирование интересов пользователей

Yahoo!¹

Данные: поисковые сессии и посещенные старницы для десятков млн пользователей, логи показов рекламы

¹Amr Ahmed и др. “Scalable distributed inference of dynamic user interests for behavioral targeting”. В: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, с. 114—122.

Предсказание кликов по рекламе

Criteo²,
<https://www.kaggle.com/c/criteo-display-ad-challenge>
Данные: (анонимизированные) логи показов рекламы

²Alekh Agarwal и др. "A reliable effective terascale linear learning system". В: *The Journal of Machine Learning Research* 15.1 (2014), с. 1111—1133.

Рекомендации фильмов

Netflix³,
<https://www.kaggle.com/netflix-inc/netflix-prize-data>
Данные: множество четверок - (пользователь, фильм, рейтинг, дата)

³Yehuda Koren. "The bellkor solution to the netflix grand prize". В: *Netflix prize documentation* 81.2009 (2009), с. 1—10.

Особенности задач

- Большое количество объектов и признаков ($> 10^6$)
- Данные сильно разрежены (маленькое число ненулевых признаков)
- Большие категориальные признаки

Подвыборка?

Кратко: качество хуже (Пример от Facebook⁴)

Следствие

Нужны инструменты для работы с большими данными и быстрые алгоритмы обучения

⁴Xinran He и др. “Practical lessons from predicting clicks on ads at facebook”. В: *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM. 2014, с. 1—9.

Часть I

Методы и системы обработки больших данных

- 1 Introduction to Hadoop and MapReduce
- 2 Apache Spark
- 3 Spark SQL

Часть II

Машинное обучение на больших данных

- 1 Spark MLLib Overview
- 2 Stochastic Gradient Descent, Linear Models, Neural Networks
- 3 Hyperparameters Optimization
- 4 Gradient Boosting Decision Tree
- 5 Word2Vec, k-Nearest Neighbors (LSH)
- 6 Collaborative Filtering (ALS)
- 7 Latent Dirichlet Allocation
- 8 Dimensionality Reduction
- 9 Online Learning
- 10 Algorithms on Graphs

Часть III

Проведение онлайн экспериментов

- 1 How to conduct AB Tests (Experiment Design, Execution, Analysis)
- 2 Results Analysis ((Multiple) Hypothesis testing, Sensitivity, Power)
- 3 Heterogeneous Treatment Effect

Практики

Pull Requests with Jupyter notebooks

Домашние Задания

Kaggle

- 1 CTR-prediction (Criteo Ads)
- 2 Recommendations (MovieLens + TMDB)

В равных долях: тетрадки по практикам, контекст 1, контекст 2



Рис.: <https://vk.me/join/AJQ1d8u00RYxn/gKevTcK8VJ> — чат в VK

- <https://github.com/ishugaepov/MLBD> — презентации, jupyter тетрадки и пр.;
- <https://github.com/ishugaepov/MLBD-notes> — конспект.