“Data and Service Engineering for Automating Business Processes”

Motivation

Machine learning is changing the world rapidly and dramatically, every modern enterprise is now eyeing machine learning as one of the top instruments to improve business KPIs. Yet, behind any successful application of machine learning is a large chunk of work that is done by engineers, which includes Data Engineering functions such as data cleaning, wrangling, integration, etc. And the models must be deployed in production as reliable services. And finally, advanced analytics will need to take place in order to understand how the service is operating. In this course you will learn the basics of these engineering and analytic disciplines.

We won’t focus on machine learning algorithms in this course, its a prerequisite.

Prerequisites

- Databases
- Machine Learning
- Python Programming

Final Grade:

- 80% is homeworks
    - 2 programming tasks

    - 2 homeworks
- 20% final exam

You can receive full credit for the final automatically, if you do well on all the assignments

Course Material

Lecture 1: Introduction

Here we'll learn why its hard to train a machine learning model and quickly put it into production and embark on another project. What are the extra problems that creep up during this process? What extra risks appear when the model is transferred to production mode? We'll do an overview of general decision systems based on Data Science. We'll also dive into a specific business scenario, that will be the guiding example in our course: online credit business. We will go over the business model, major KPIs, the constraints the business places on possible machine learning solution and some fundamental problems.

Seminar 1: Detailed description of the credit business and how it impacts automatic decision making. The cost of the data, the cost of a mistake and other characteristics of online credit business.

Lecture 2: Data, Basic Data Types, Data Models

Data in modern businesses comes in a variety of different types, from basic textual and numeric data, to geographical data, images, videos, timeseries, etc. We will go over basic data types and show how their are best used in machine learning tasks. Then we'll move into data models.

Seminar 2: Overview of the data that is accumulated by an online credit business.

Lecture 3: Data Models in Detail: Relational, XML, JSON

We'll dive into detail into relational, XML and JSON data models. We'll go over dangling pointers, referential integrity, 3rd normal form, XML and JSON schemas.

Seminar 3: The overview of SQL tools: PostgreSQL advanced features (window functions, complicated queries, optimization). XML tools: lxml lib for Python, XML parsing approaches on the real example (DOM and iterator parsers, XPath queries).

Lecture 4: Event-based data models. Kappa and Lambda architectures. Process mining.

Typical business can be described as a set of business processes, and the event-based data model captures all important events, generated by these processes. Log of such events is at the core of modern real-time architectures such as Lambda and Kappa. We'll study how to recover all the needed data from the event log, how to test hypothesis on top of such a log. We'll create usable data marts on top of event logs for analytics. We'll study advanced analytics techniques such as process mining and cohort analysis.

Seminar 4: Event log for an online credit business. Data marts for analytical queries. Ad-hoc queries. Process Mining typical use-cases. Querying event log with XQuery or PythonQL. Overview of the first programming assignment.

Lecture 5: Data Integration and Cleaning

What are the typical problems with data quality? How can we increase data quality? Data integration problem: semantic data integration, virtual data integration.

Seminar 5: Overview of Programming Assignment 2

Lecture 6: Data and model versioning

Data Science teams run into a problem of workflow and data versioning constantly. We will look at possible solutions to these problems.

Especially we want to achieve results reproducibility and correctness when working in teams.

Seminar 6: Bi-temporal data models.

Lecture 7: Building an automated decision engine

Feature engineering for the models. Feature marts. Workflows for creating features. Choosing the right loss function. Reward functions. Various problems of optimizing the KPIs of companies.

Seminar 7: Detailed overview of optimizing the economy of an online credit business. Models A/B testing. A/B testing versus multi-arm bandits.

Lecture 8: Collective machine learning

Feature engineering on a flow of events, instead of individual examples.

Lecture 9: Deployment of models, Anomaly detection

Engineering tasks in deployment of machine learning models. Monitoring of models and automatic anomaly detection. Advanced anomaly detection methods.

Seminar 8: Anomalies for objects with complex structure.