Case Study: Model Engineering

Course Description

The construction of data science models and applying the techniques to real-world problems requires a deep understanding of data science processes and techniques beyond the application of relevant algorithms. This course starts by introducing two commonly used data science methodologies: CRISP-DM and MS Team Data Science. Any data taken from real machines, systems, or processes will include some errors to varying degrees. This course discusses in detail how to detect and correct data quality issues, including the importance of domain knowledge in the determination of the veracity of the data. Many machine learning approaches require the creation and subsequent selection of model features that determine which parts of the data are used in which ways in the later modelling steps. This course discusses methods to engineer and build new features from raw data and outlines statistical methods to identify the most relevant features for the given task. Finally, the course outlines strategies to avoid common fallacies when building data science models, as well as approaches to automating workflows.

Contents

1. Data Science Methodologies
   1. CRISP-DM
   2. MS Team Data Science
2. Data Quality
   1. Evaluating Data Quality
   2. Using Low-Quality Data
   3. Data Duality and Domain Knowledge
3. Feature Engineering
   1. Building New Features
   2. Splitting Variables
   3. Feature Engineering Exploiting Domain Knowledge
4. Feature Selection
   1. Univariate Feature Selection
   2. Model-Based Feature Selection
5. Building a Predictive Model
   1. Establishing a Benchmark Model
   2. Prediction as Probabilities
   3. Interpretable Machine Learning and Results
6. Avoiding Common Fallacies
   1. Overtraining and Generalization
   2. Overfitting and Occam’s Razor
   3. Workflow Automation and Model Persistence