
Practical Machine Learning, AI and Data Science on SQL Server 2017, Microsoft ML Server, R, and Azure ML

Varighed: 5 Days Kursus Kode: M028

Beskrivelse:

This live classroom course is new for 2018! It focuses on the newest technologies of Microsoft Machine Learning Server and SQL Server 2017. By popular demand, Part B of this course teaches programming in R, however most of the course is also applicable to Python programmers, as the key libraries are the same.

This course has two parts. You start with 2-day Part A: Fundamentals of Machine Learning followed by 3-day Part B: Immersion into Machine Learning in R, SQL Server 2017, and Microsoft ML Server. The first part introduces the most important concepts and tools, while the second part teaches you R and how to use it for machine learning on the Microsoft platform.

Målgruppe:

Part A:[M030](#) Analysts, budding data scientists, database and BI developers, programmers, power users, DBAs, predictive modellers, forecasters, consultants. Part B:[M031](#) Data scientists and attendees of Part A.

If you have attended a prior course on Machine Learning, like Rafal's week-long class Practical Data Science that was offered in 2015–2017, and if you are versed in model validity, accuracy, and reliability, consider attending Part B only. Ask yourself these questions: can I explain the difference between cross-validation and hold-out testing, do I know which business metrics correspond to precision and which to recall, is model accuracy more important than reliability, and how does a boosted decision tree work. If in doubt, please attend both Parts 1 and 2.

Agenda:

- Why attend this class?
- Because of Rafal's 10+ years of real-world machine learning experience.
- You will not only learn all the concepts and tools that you need to know from a great teacher who has trained almost 500 data scientists world-wide, a highly-respected presenter, capable of holding your attention, but, above all, from a practitioner of machine learning. Rafal Lukawiecki has been delivering ML, data mining, and data science projects for customers in retail, banking, entertainment, healthcare, manufacturing, education, and government sectors for over ten years. Because of that, you will learn:
 - How to avoid common pitfalls,
 - How to get ahead of your competition by working faster,
 - What is really useful and practical,
- What is more theoretical but still important,
- What hype you should be wary of.
- You will be able to ask any questions related to your industry and you will get relevant, pragmatic, no-nonsense answers, helping you get ahead with your own projects.
- Learn from Rafal who has done it all, not from those who just teach it—this is why it is called Practical Machine Learning.
- 50% lectures, 30% demos, 20% tutorials.
- You are encouraged to follow the demos on your machine, and you will be challenged to find answers to 3 larger problems during the tutorials. While they are a hands-on part of the course, if you prefer not to practice, you are welcome to use that time for additional Q&A, or to analyse your own data. We will provide you with all the necessary data sets, and we will explain what free or evaluation edition software needs to be installed to follow the course on your own laptop. In some training centres we are able to provide pre-built machines which you can use instead of your own—please enquire. You will need an Azure account (even a free one) during the course. You can copy course experiments and data into your workspace for learning and for future reference after the course.

Indhold:

Part A: Introduction to Machine Learning, AI ; Data Science with Azure ML and SQL Server (Mon-Tue).

To deliver the best possible training we follow the industry. The agenda and course content are subject to continuous improvement and revision without further notice.

Machine Learning Fundamentals.

We begin with a thorough introduction of all of the key concepts, terminology, components, and tools. Topics include:

- Machine learning vs. data mining vs. artificial intelligence.
- Tool landscape: open source R vs. Microsoft R, Python, SQL Server, ML Server, Azure ML.
- Teamwork.

Algorithms

There are hundreds of machine learning algorithms, yet they belong to just a dozen of groups, of which 5 are in very common use. We will introduce those algorithm classes, and we will discuss some of the most often used examples in each class, while explaining which technology tools (Azure ML, SQL, or R) provide their most convenient implementation. You will also learn how to find more algorithms on the Internet and how to figure out if they are any good for real use. Topics include:

- What do algorithms do?
- Algorithm classes in R, Python, ML Server, Azure ML, and SSAS Data Mining.
- Supervised vs. unsupervised learning.
- Classifiers.
- Clustering.
- Regressions.
- Similarity Matching.
- Recommenders.

Data

Machine learning requires you to prepare your data into a rather unique, flat, denormalised format. While features (inputs) are always necessary, and you may need to engineer thousands of them, we do not need labels (predictive outputs) in all cases. Topics include:

- Cases, observations, signatures.
- Inputs and outputs, features, labels, regressors, independent and dependent variables, factors.

Introduction to Model Validation

The most important aspect of any data science, artificial intelligence, and machine learning project is the iterative validation and improvement of the models. Without validation, your models cannot be reliably used. There are several tests of model validity, most importantly those that check accuracy and reliability. Topics include:

- Testing accuracy.
- False positives vs. false negatives.
- Classification (confusion) matrix.
- Precision and recall.
- Balancing precision with recall vs. business goals and constraints.
- Introduction to lift charts and ROC curves.
- Testing reliability.
- Testing usefulness.

Part B: Intermediate Machine Learning in R on SQL Server and Microsoft ML Server (Wed-Fri).

To deliver the best possible training we follow the industry. The agenda and course content are subject to continuous improvement and revision without further notice.

Working with R

There is a large number of tools that you can use with R, and we begin the day focusing on the essential ones. You will also learn how to organise your workflow. Topics include:

- RStudio vs. R Tools for Visual Studio.
- Rattle.
- Microsoft Machine Learning Server vs SQL Server Machine Learning Services.
- Projects, files, scripts, history, version control.
- Notebooks and RMarkdown.

Data Preparation in R

R uses data frames, data tables, and tibbles, amongst others, while ML Server adds XDFs and the ability to work with data stored natively in Hadoop, Spark, and SQL Server. While most data preparation should be done as close to source, preferably using SQL, you will need to learn how to perform some transformations in R. Topics include:

- Data frames, tables, tibbles.
- Reading files and ODBC data.

also learn how to apply this technique for anomaly (outlier) detection and data preprocessing. Topics include:

- Introduction to segmentation.
- Clustering algorithms (k-means, EM, hierarchical, and others).
- Interpreting clusters.
- Anomaly detection with clustering, PCA and SVMs.

Classification

Without doubt, classifiers are the most important, and the most often used category of machine learning algorithms, and the foundation of algorithmic data science, and of most of today's Artificial Intelligence. We will focus on several variants of the most important classification algorithm—decision tree—while progressively interpreting the results, and improving its performance. After introducing neural networks and logistic regression we will also compare the performance of all of these classifiers on our test dataset. Topics include:

- Introduction to classifiers.
- Two-class (binary) vs multi-class.
- Decision trees, forests, and boosting.
- Neural networks and logistic regression.
- Overfitting (overtraining) concerns.

Classifier Validation

Validation of classifiers will be your key concern, because classifiers are used so often, and because their accuracy is not easy to balance with business requirements, such as restricted resources, or a required level of business performance. Building on your understanding of model validity (introduced in Part A of this course), you will learn how to balance an acceptable number of false positives with false negatives by using classification (confusion) matrices, metrics of precision and recall, by plotting ROC (Receiver Operating Characteristic) curves, and by measuring their business impact using profit and cost charts. Attendees have commented in the past that this is the most important module of the entire course. Topics include:

- Testing classifiers.
- Charting precision-recall and sensitivity-specificity.
- ROC curves and lift charts in detail.
- Other measures of accuracy, including AUC, and F1 scores.
- What exactly does cross-validation tell us?
- Measuring quality of cross-validation.

<ul style="list-style-type: none"> ■ Data formats, discretization/quantizing vs. continuous. ■ Indicator columns. ■ Feature engineering. ■ Azure ML data preparation and manipulation modules. ■ Moving data around and its storage, SQL vs. NoSQL, files, data lakes, BLOBS, and Hadoop. 	<ul style="list-style-type: none"> ■ XDFs and connecting to data in ML Server. ■ Tidyverse. ■ dplyr. <p>Plots and Visualisations in R</p> <p>One of the strengths of R is the ease of creating accurate (and good looking!) plots. As a bare minimum you need to understand how to use the most popular visualisation package, ggplot2, and some of the built-in base functions. Topics include:</p> <ul style="list-style-type: none"> ■ Summarising data. ■ Base boxplots, histograms, scatter plots. ■ ggplot2: grammar of graphics. ■ Combining visualisations into layers. ■ Density plots. ■ Surfacing R graphics in Power BI and SQL Server. 	<ul style="list-style-type: none"> ■ Optimising binary classifier prediction probability thresholds for a given business target. ■ Refining models to improve accuracy and reliability. ■ Hyperparameter tuning. ■ Class imbalance problem (fraud analytics and rare event prediction). <p>Regressions</p> <p>Considered by some as the numerical equivalent of classifiers, regression is a large subject of its own. We will introduce its simple but a very popular form, linear regression, and the more precise, but also prone-to-overfitting, decision tree variants. Topics include:</p> <ul style="list-style-type: none"> ■ Introduction to simple regressions in R. ■ Linear regression (classic). ■ Regression decision trees and other ensemble regression algorithms. ■ Regression as a building block of other algorithms.
<p>Process of Data Science</p> <p>The process consists of problem formulation, data preparation, modelling, validation, and deployment—in an iterative fashion. You will briefly learn about the CRISP-DM industry-standard approach but the key subject of this module will teach you how to apply the scientific method of reasoning to solve real-world business problems with machine learning and statistics. Notably, you will learn how to start projects by expressing needs as hypotheses, and how to test them. Topics include:</p> <ul style="list-style-type: none"> ■ CRISP-DM. ■ Stating business question in data science term. ■ Hypothesis testing and experiments. ■ Student's t-test. ■ Pearson chi-squared test. ■ Iterative hypothesis refinement. 	<p>Clustering, Segmentation, Anomaly Detection</p> <p>Segmentation is the main application of unsupervised learning using clustering algorithms. You will</p>	<p>Regression Validation</p> <p>Unlike classifiers, regressions are easier to assess. You will learn about basic tests of classical linear regressions that are easy to perform in R, and about measuring quality of machine learning, non-linear regressions. Topics include:</p> <ul style="list-style-type: none"> ■ Measuring linear regression quality. ■ Homoscedasticity, multicollinearity and other concerns. ■ Measuring machine learning regression quality. ■ R-squared (Coefficient of Determination), RMSE, MAE, RAE, RSE.
<p>Introduction to Model Building</p> <p>At the heart of every project we build machine learning models! The process is simple and it follows a well-trodden path. In this module you will build your first decision tree and get it ready for validation in the next module. Topics include:</p> <ul style="list-style-type: none"> ■ Connecting to data. ■ Splitting data to create a holdout. ■ Training a decision tree. ■ Scoring the holdout. ■ Plotting accuracy. 		<p>Deployment to Production</p> <p>If you plan on using your models for prediction, rather than just for the exploration of data, or if you want to embed them as Artificial Intelligence in your applications, you need to deploy your models to production and maintain them on an on-going basis. Since we focus on the Microsoft ML Server and SQL Server ML Services, you will learn about the PREDICT T-SQL statement, and other supported mechanisms for deploying your models. We will also discuss how to deploy models as a web service, using these, and other Microsoft and non-Microsoft techniques. Topics include:</p> <ul style="list-style-type: none"> ■ What needs to be deployed, and when? ■ PREDICT T-SQL statement. ■ Using sp_execute_external_script.

■ Web service deployment with and without Azure ML.

■ On-going maintenance and model updates.

Please note: we reserve the right to amend the order of the modules to best suit the dynamic character of the class and to answer questions as they arise. Some subjects will only be covered if time allows, but your satisfaction is guaranteed.

Flere Informationer:

For yderligere informationer eller booking af kursus, kontakt os på tlf.nr.: 44 88 18 00

training@globalknowledge.dk

www.globalknowledge.com/da-dk/

Global Knowledge, Stamholmen 110, 2650 Hvidovre