Intermediate Machine Learning in R on SQL Server and Microsoft ML Server

Varighed: 3 Days   Kursus Kode: M031

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.

Målgruppe:

Part B: 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.

Indhold:

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.
- XDFs and connecting to data in ML Server.
- Tidyverse.
- dplyr.

Plots and Visualisations in R

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:

- 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.

Clustering, Segmentation, Anomaly Detection

Segmentation is the main application of unsupervised learning using clustering algorithms. You will

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:

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:

- 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.

Regression Validation

Unlike classifiers, regressions are easier to asses. 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:

- Measuring linear regression quality.
- Homoscedasticity, multicollinearity and other concerns.
- Masuring machine learning regression quality.
- R-squared (Coefficient of Determination), RMSE, MAE, RAE, RSE

Deployment to Production

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:

- 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

- 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.
- 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).

Regressions

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