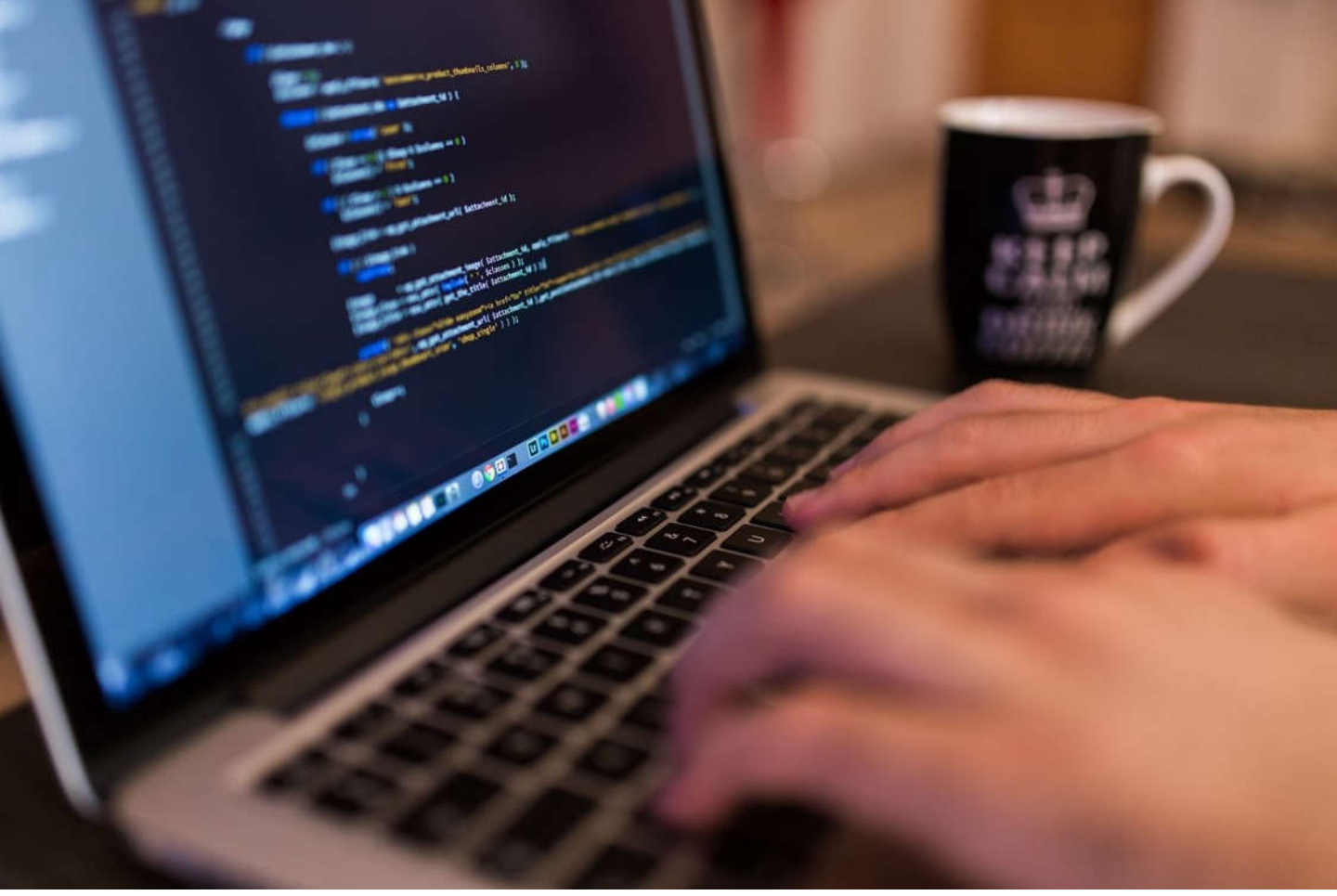

Data Science in R

Schulung der Grundlagen von Machine Learning und Datenanalyse mit data.table

ENABLE AI

MASTER YOUR DATA

Data Science in R

Grundlagen von Machine Learning und data.table

Über den Kurs

Dauer: 3 Tage

Gruppengröße: 3-10

Level: Anfänger mit Programmiererfahrung*

Anteil Coding: 60%

Sprache: R

Bibliotheken: data.table, ggplot2, caret, rpart, stats, randomForest,…

Auf einen Blick

- Wichtige Datenstrukturen
- data.table
- Statistiken berechnen
- Eine eigene Funktion schreiben
- Grafiken mit ggplot2 erstellen
- Daten einlesen und schreiben (csv, xls, .RData)
- Machine Learning Einführung
- Lineare Regression, Entscheidungsbaum, K-Means
- Train-Test-Split, Kreuzvalidierung

Überblick über das Seminar

Der dreitägige Einführungskurs in R und Data Science erklärt die Grundlagen von Data Mining/Data Science und die Verwendung von R. Das notwendige theoretische Verständnis wird geschult mit dem Ziel die Verfahren praktisch in R umsetzen zu können. Am Ende des Seminars können Sie selbstständig erste Daten Analysen umsetzen und Machine Learning Algorithmen für die Datenanalyse einsetzen.

Das Paket data.table ist ein Schwerpunkt des R Kurses, da dieses Paket speziell für Data Science entwickelt wurdet und aufgrund seiner Performance überzeugt. Die data.tables entsprechen den häufig in der statistischen Programmiersprache R verwendeten data.frames und ermöglichen eine elegante Analyse der Daten.

Das Paket ggplot2 wird für die Erstellung von Plots und Grafiken erläutert. ggplot2 ist eines der beliebtesten Pakete in R für die Visualisierung.

Nach den Grundlagen der Datenanalyse erhalten Sie einen Überblick über Machine Learning Algorithmen, welche wir in R selbst coden werden. Ein wichtiger Bestandteil des Seminards ist das eigenständige Arbeiten und Lösen von Übungsaufgaben, so dass Sie mit direkter Hilfe des Trainers das Besprochene direkt in der Praxis umsetzen und anwenden können.

Inhalte des Seminars

Sie erhalten einen Überblick über die in der Statistik, Data Science und Machine Learning viel benutzen Programmiersprache R. Als Entwicklungsumgebung/IDE wird RStudio verwendet, die am Meisten verwendete Umgebung für R. Nach dem Seminar können Sie Daten aus verschiedenen Formaten einlesen, Daten mit ggplot2 plotten bzw. Daten bereinigen (fehlende Werte ersetzen, Zeilen und Spalten anpassen, neue Spalten erzeugen).

Sie kennen die wichtigsten Datentypen in R, können eigene einfache Funktionen schreiben und kennen die Umsetzung von Control Flows (For-Schleife, If-Else) in R. Sie verstehen das Grundkonzept eines data.table und können damit Data Wrangling und Data Cleaning durchführen.

ENABLE AI

* Programmierkenntnisse in R sind nicht notwendig. Erste Erfahrung mit einer anderen Sprache (z.B. Python, VBA, C) ist ausreichend.

Die Einteilung von Machine Learning in supervised-unsupervised und Reinforcement Learning ist Ihnen bekannt und Sie können eigenständig Algorithmen in R trainieren, validieren, einen Train-Test Datensplit durchführen und Gütekriterien zur Beurteilung eines Algorithmus berechnen und interpretieren. Bekannte Algorithmen im Machine Learning werden erläutert und Sie können die verschiedenen Algorithmen verstehen und in R im Code schreiben. Das Seminar umfasst Algorithmen für die Regression, Klassifikation und dem Clustering: Lineare Regression, Logistische Regression, Entscheidungsbaum, Random Forest, k-means clustering.

Die Einstiegshürde für die Benutzung von R im Machine Learning und Data Science ist genommen, so dass Sie eigenständig Ihr Wissen nach dem Seminar erweitern können. Der Schwerpunkt liegt auf der eigenen Umsetzung auf Ihrem Laptop mit Unterstützung des Dozenten.

Wer sollte teilnehmen?

Dieser Kurs mit R richtet sich an data scientists, angehende Machine Learning engineers, Datenanalysten, Busines Intelligence Analysts, Data Analysts, o.ä. welche die Programmiersprache R für Data Science/Data Mining kennenlernen möchten, um Datenanalysen unter Verwendung von Machine Learning Algorithmen eigenständig umzusetzen

Voraussetzungen

Dieser R Kurs setzt keine grundlegenden Kenntnisse in R voraus. Notwendig ist jedoch Vorerfahrung mit einer Programmiersprache, damit Konzepte einer Variablen, Zuweisung von Werten zu einer Variablen, eine Funktion bzw. eine for-Schleife bekannt sind.

Notwendig sind außerdem grundlegende Vorkenntnisse im Bereich der Statistik (Begriffsdefinitionen wie bspw. Mittelwert, Median, Standardabweichung, Normalverteilung), Kenntnisse grundlegender mathematischer Symbole und Begriffe (Summenzeichen, Integral, Funktion, Ableitung, Exponentialfunktion) und Kenntnis der booleschen Algebra mit den logischen Operatoren (UND, ODER, NICHT) sind sehr empfohlen.

Die Teilnehmer sollten Vorerfahrung mit dem Umgang von Daten haben, z.B. in Excel oder einer BI-Software, damit Konzepte einer spaltenweisen Berechnung bzw. einfache Statistiken (Mittelwert, Varianz) bekannt sind.

Das Seminar wird auf Deutsch gehalten. Englischkenntnisse (lediglich im Verstehen von englischen Texten) sind sehr empfehlenswert, da die Programmiersprache, Fachbegriffe und die Dokumentationen im Internet auf Englisch sind. Aus diesem Grund sind auch die erstellten Folien in der Schulung auf Englisch.

Methode des Seminars

Dieses Seminar ist sehr praxisorientiert. Die Teilnehmer arbeiten direkt und selbstständig mit der Programmiersprache R in der Entwicklungsumgebung RStudio, so dass das Erlernte direkt geübt und vertieft werden kann. Der Trainer moderiert dabei verschiedene Aufgaben und begleitet die Teilnehmer durch die einzelnen Lehreinheiten.

Technische Voraussetzungen der Teilnehmer (Laptop, etc.)

- Die Teilnehmer benötigen für die Übungsaufgaben Laptops. Wir empfehlen, Ihren eigenen Laptop mit der vorab installierten Software mitzubringen. Eine genaue Installationsanleitung für die Software wird Ihnen vor dem Seminar per E-mail zugesandt. Auf Anfrage stellen wir auch Schulungslaptops zur Verfügung.
- Bitte prüfen Sie, ob Ihr Firmenlaptop Zugangsbeschränkungen im Internet hat. Die digitalen Unterlagen (Skript, Code, Dateien) werden im Seminar online zum Download zur Verfügung gestellt. Sie erhalten vor dem Seminar per E-Mail den Link zu einer Testdatei zum Download, um dies überprüfen zu können.
- Sie sollten sich in firmenfremde WLAN-Netze registrieren können.
- Als Backup Lösung ist es möglich, dass der USB Port bei Ihrem Laptop freigeschalten ist, um damit verwendete Dateien oder sonstige Unterlagen übertragen zu können.
- Im Seminar wird das Betriebssystem Windows verwendet. Der Umgang mit Ihrem verwendeten Betriebssystem und Laptop sollte bekannt sein. Insbesondere sollten Sie ohne Schwierigkeiten Sonderzeichen auf der Tastatur finden (insbesondere bei Apple Geräten werden auf manchen Tastaturen nicht immer runde, eckige bzw. geschweifte Klammern dargestellt).

Tagesübersicht

Tag 1

10.00-10.15
Begrüßung und Organisatorisches
- Vorstellungsrunde
- Erwartungen der Teilnehmer

10.15-11.45
Einführung in R & Wichtige Datenstrukturen
- R und RStudio
- Installation und Laden von Paketen
- Wie finde ich weiterführende Informationen?
- Welchen Unterschied hat R zu anderen Programmiersprachen?
- Grundlegende Datenstrukturen in R

11.45-12.00
Kaffeepause

12.00-13.30
Data.table Paket – Einführung
- Einführung in die Besonderheit vom data.table Paket und Ähnlichkeit zu SQL-Abfragen
- Erzeugen eines data.table
- Möglichkeiten, Zeilen und Spalten zu extrahieren
- Berechnungen direkt auf Spalten durchführen
- Berechnungen nach Variablen gruppieren

13.30-14.30
Mittagspause

14.30-16.00
Daten einlesen und schreiben
- Das Arbeitsverzeichnis in R und RStudio setzen
- Ein CSV, Excel bzw SPSS Datei einlesen / schreiben
- Überblick über nützliche Parameter
- Die fread() Funktion für große Datenmengen
- Abfragen aus einer Datenbank (SQLite)

16.00-16.15
Kaffeepause

16.15-18.00
Datenmanipulation
- Manipulationen auf einem data.table bei Zeilen und Spalten
- Zusammenfassen und Filtern von Daten
- Variablen erstellen, löschen, ändern
- Fehlende Werte ersetzen
- lapply() und die Anwendung in einem data.table (mit .SD und .SDcols)

18.00
Ende

Tagesübersicht

Tag 2

09.00-09.15
Rückblick und offene Fragen von Tag 1

09.15-10.45
Control Flows
- Eine eigene Funktion schreiben
- Default Parameter in einer Funktion setzen
- For Schleifen
- If-Else Bedingungen
- While Schleife

10.45-11.00
Kaffeepause

11.00-12.30
Datenvisualisierung mit ggplot2
- Das Konzept hinter der Grammar of Graphics
- Die Layer von ggplot2 zur Erstellung erster Plots und zum Plotten von Statistiken
- Darstellungen (Größe der Datenpunkte, Farbe, Gruppierung) mit einer Variable variieren oder festsetzen
- Mehrere Subplots erstellen, Plots anpassen und speichern

12.30-13.30
Mittagspause

13.30-15.15
Berechnen von Statistiken direkt im data.table
- Deskriptive Statistiken
- Zufallszahlen aus verschiedenen Verteilungen ziehen
- Korrelationen (Spearman, Pearson)

15.15-15.30
Kaffeepause

15.30-17.00
Überblick über Machine Learning
- Einführung in Machine Learning (ML)
- Anwendungsbeispiele von ML
- Künstliche Intelligenz – Machine Learning – Deep Learning
- Unterschied Supervised – Unsupervised Learning (überwachtes – unüberwachtes Lernen)
- Overfitting, Train-Test-Split und cross-validation (Kreuzvalidierung)

17.00
Ende

Tagesübersicht

Tag 3

09.00-09.15
Rückblick und offene Fragen von Tag 2

09.15-10.45
Lineare Regression
- Aufteilung der Daten in Test-Train, ein Model erstellen und validieren
- Grundlagen des Algorithmus
- Umsetzung in R
- Validieren der Ergebnisse (mean squared error)
- cross-validation (Kreuzvalidierung)

10.45-11.00
Kaffeepause

11.00-12.30
Entscheidungsbaum in R
- Aufteilung der Daten in Test-Train, ein Model erstellen und validieren
- Grundlagen des Algorithmus
- Umsetzung in R
- Validieren der Ergebnisse (u.a. confusion matrix, sensitivity, accuracy)
- Anpassen von Hyperparametern im Training
- cross-validation (Kreuzvalidierung)

12.30-13.30
Mittagspause

13.30-15.15
Logistische Regression in R
- Aufteilung der Daten in Test-Train, ein Model erstellen und validieren
- Grundlagen des Algorithmus
- Umsetzung in R
- Validieren der Ergebnisse (u.a. AIC; BIC; confusion matrix, sensitivity, accuracy)
- ROC curve und AUC
- Interpretation der Ergebnisse
- cross-validation (Kreuzvalidierung)

15.15-15.30
Kaffeepause

15.30-17.00
Weitere Machine Learning Algorithmen
- Überblick der Algorithmen Support Vector Machine (SVM), Random Forest und K-means
- Umsetzung der Algorithmen in R
- Validieren der Ergebnisse

17.00
Ende

Ihre Dozenten

Unser folgender Experte leitet das Seminar

Jan Köhler
Vom Hintergrund Statistiker und Wirtschaftsingenieur hat er über 7 Jahre an den neuesten Technologien in Machine Learning, Deep Learning und Data Science im Bosch Center for Artificial Intelligence (BCAI) gearbeitet, hat in der Praxis bei über 25 Patentanmeldungen (meist als Haupterfinder) beigetragen und ist Mitautor bei Veröffentlichungen auf Machine Learning Konferenzen neben Veröffentlichungen im Bereich der Medizinstatistik bzw. des Operations Research. In vielen Praxisprojekten unterstützte er bisher als Data Scientist und hat verschiedene Teilnehmer, vom Projektmitarbeiter bis zum Konzern-Vorstand geschult.

Zusammenfassung

Preise

1710 € zzgl. MwSt.

Termin und Ort

Termine und Orte finden Sie unter https://enable-ai.de

Haben Sie Fragen? Wir helfen Ihnen. Versprochen.

Enable AI, Stuttgart
0711 96881553
info@enable-ai.de

ENABLE AI