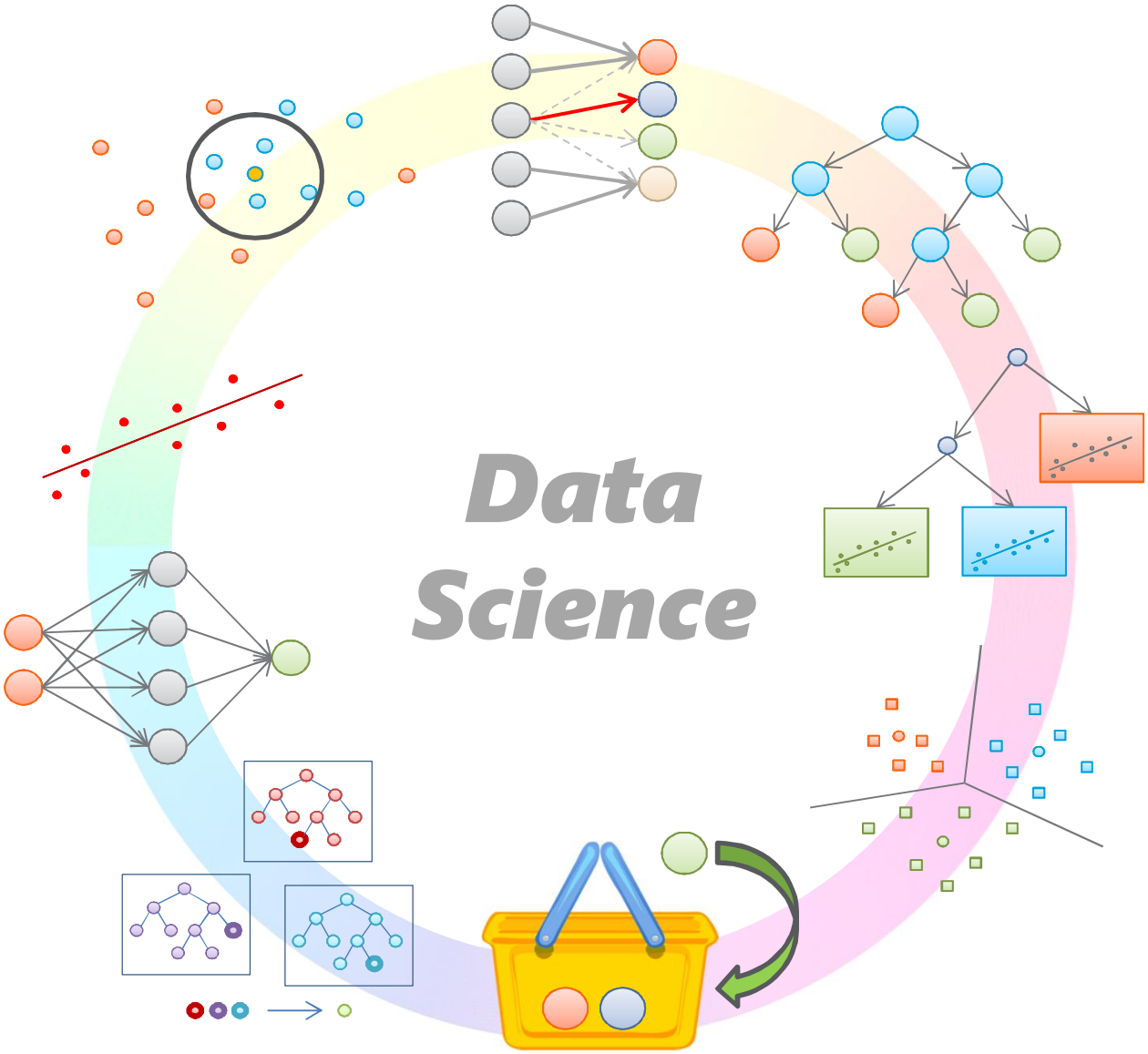
Seminarreihe 2017

Big Data, Statistik, Data Mining, maschinelles Lernen, R und Python

USEDATA



BIG DATA

R

Python

Grundlagen

Unsere allgemeinen Seminare behandeln Themen, welche grundlegend sind, um Aufgaben aus den Bereichen *Statistik*, *Data Mining* und *maschinelles Lernen* kompetent umzusetzen.

BIG DATA – EINE EINFÜHRUNG

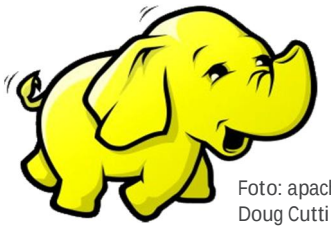


Foto: apache.org
Doug Cutting

In diesem Seminar erfahren die Teilnehmer, woher dieser Begriff stammt und mit welchen Techniken unstrukturierte Massendaten rasch eingelesen und bearbeitet werden können. Außerdem werden die Themen *Data Mining* und *maschinelles Lernen* behandelt, ohne denen *Big Data* gar nicht möglich wäre. Eine Vorstellung diverser Tools rundet dieses Seminar ab.

EINFÜHRUNG IN *R*

Die Open Source Software *R-Language* existiert bereits einige Jahre. Eine unglaubliche Anzahl an Packages (rund 10.000 mit Ende 2016) macht die Vielfalt dieser Sprache aus. In diesem Seminar lernen die Teilnehmer die Entwicklungsumgebung *R-Studio* und die wesentlichen Sprachelemente von *R* kennen. Es wird gezeigt, wie Daten (Textdateien, Serverdatenbanken, SAS, SPSS ...) importiert werden und eine einfache Analyse durchgeführt wird.

GRAFIK MIT *R*

Eine der Stärken von *R* ist die Visualisierung (= grafische Darstellung) von Daten und Modellen. In diesem Workshop werden die Grundbefehle, die benötigt werden, um eine ansprechende Grafik zu erstellen, erklärt. Anhand von Beispielen zeige ich auch die Vielfalt an Darstellungsmöglichkeiten. Neben dem Basispackage für *R* Grafiken werden auch neuere Packages vorgestellt, welche die Möglichkeiten der Visualisierung immer weiter vorantreiben.

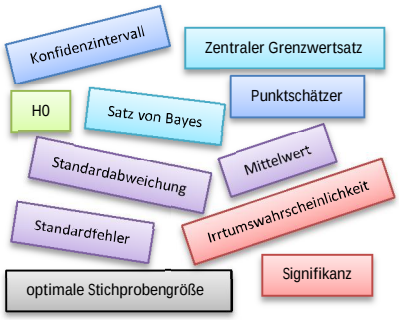
Python

R und *Python* sind populäre Sprachen für statistische Anwendungen. Worin liegen nun die Vorteile von *Python*?
Python lässt sich einfach in Web-Anwendungen integrieren und ist eine vollwertige Programmierumgebung, um sie in der Produktion einzusetzen. Folgende Bibliotheken werden von Data Science-Mitarbeitern hauptsächlich genutzt: NumPy/SciPy (science computing), pandas (Datenmanipulation), matplotlib (Grafiken) und scikit-learn (maschinelles Lernen).

Statistik

Diese Seminare behandeln ein konkretes Gebiet der Statistik unter Zuhilfenahme von *R*. Dazu habe ich eine eigene Software entwickelt - damit können Daten generiert werden, die den Anforderungen des jeweiligen Seminars entsprechen. Dabei sind die Voreinstellungen (Parameter) zunächst nur mir bekannt. Die Teilnehmer erhalten im Rahmen des Seminars eine Textdatei mit Daten, welche sie in *R* einlesen können. Die Aufgabe besteht nun darin, diese Parameter auf Grund der Daten zu ermitteln. Am Ende wird das Geheimnis gelüftet und die Teilnehmer können ihr Ergebnis mit dem tatsächlichen Resultat vergleichen. Durch wiederholtes Anwenden dieser Technik wird das Vertrauen in statistische Methoden gestärkt und gleichzeitig die notwendige Sicherheit für den Gebrauch im Alltag gebildet.

STATISTISCHE GRUNDBEGRIFFE



Dieser Kurs erläutert die wichtigsten statistischen Grundbegriffe.

Auch hier kommt die von mir entwickelte Software zum Einsatz, welche es mir erlaubt, auch zunächst scheinbar komplexe Begriffe anschaulich darzustellen.

So manchem Teilnehmer wird hier ein Aha-Erlebnis beschert.

KLASSISCHE TESTS MIT *R*

In diesem Seminar wird die Umsetzung unterschiedlicher Tests (Ein-, Zweistichprobentest) mit Hilfe der *R Syntax* dargestellt. Es werden Mittelwerte, Varianzen und Anteile verglichen. Auch der Einsatz von nicht-parametrischen Tests wird behandelt. Zur Sprache kommen z.B. der *Wilcoxon Rangsummentest* und der *Vorzeichentest*. Wie überprüft man eine Verteilungsannahme? Für Kontingenztabellen wird der *Chi-Quadrattest* vorgestellt. Abschließend wird das Thema *Korrelation* behandelt.

VARIANZANLYSE MIT *R*

Dieses Seminar behandelt das Thema *Varianzanalyse* unter Zuhilfenahme von *R.* Gibt es Unterschiede in den Klassen und mit welchen Effekten? Sind die Testergebnisse signifikant? Wurde ein erfolgreicher Test auf Grund einer zu geringen Fallzahl abgewiesen (Fehler 2. Art)? Was versteht man unter optimaler Fallzahl? Wie muss man Experimente aufbauen (Design), um konkrete Fragestellungen zu beantworten? Behandelt werden Begriffe wie *Randomisierung* und *Blockbildung*.

REGRESSION MIT *R*

Zunächst werden anhand der einfachen linearen Regression die wesentlichen Elemente beschrieben. In der Folge werden für alle Regressionsparameter die Konfidenzintervalle ermittelt und unterschiedliche Schlüsse gezogen. Dazu gehören die Vorhersage des Mittelwerts bzw. das Prognoseintervall für eine Einzelbeobachtung. Neben der klassischen linearen Regression werden zunehmend auch nichtlineare Ansätze erfolgreich eingesetzt. Der Nutzen ist vielfältig. So wird die lineare Regression bei der Beschreibung, Kontrolle und Vorhersage angewendet.

Data Mining und maschinelles Lernen unter Nutzung der *Analysis Services*

In diesen Seminaren geht es um die Einbindung der Microsoft Analysis Services Dienste, um Data Mining-Algorithmen bereit zu stellen und maschinelles Lernen zu automatisieren.

DATA MINING MIT EXCEL

Den meisten Excel-Anwendern ist nicht bekannt, dass sie Bereiche in *Excel* in Tabellen konvertieren können und so einen Zugang zu den Data Mining Diensten der Microsoft Analysis Services haben. In diesem Seminar wird zunächst diese Verbindung hergestellt. Anschließend geht es Schlag auf Schlag: Welche Variable wird durch welche Variablen beeinflusst? Wie bilde ich Cluster? Wie bestimme ich, ob ein Kunde kauft? Wie lokalisiere ich Ausreißer bzw. Extremwerte? Wie setzte ich eine Zeitreihe in die Zukunft fort? Wie teste ich Szenarien auf statistischer Basis? Wie erstelle ich eine Scorekarte? Wie führe ich eine Warenkorbanalyse durch? Antworten auf solche Fragen finden die Teilnehmer in diesem Seminar.

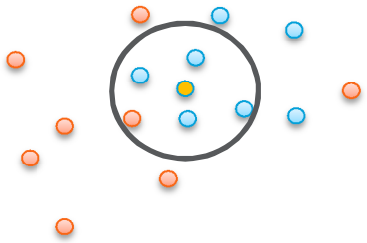
DATA MINING EXTENSION

In vielen Firmen werden die Microsoft Analysis Services eingesetzt. In den meisten Fällen werden dabei jedoch die Dienste nur zum Erstellen von OLAP Würfel genutzt. Doch die Analysis Services können mehr: Sie enthalten Dienste zum maschinellen Lernen. In diesem Seminar erhalten Sie das Grundwissen, wie solche Dienste programmatisch umgesetzt werden. Der Vorteil ist, dass die Daten ausschließlich im persönlichen Besitz bleiben. Sie müssen also nicht zwingend in die Cloud, um diese Techniken zu nutzen. Dieses Seminar zeigt Ihnen mögliche Anwendungen, die sich daraus ergeben.

Data Mining Algorithmen

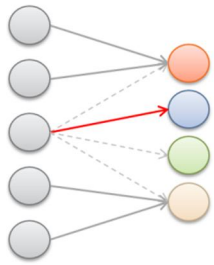
Die folgenden Seminare beinhalten jeweils einen klassischen Data Mining Algorithmus. An Hand eines Datenbestands werden Trainings- und Testdaten erstellt. Beim *Data Mining* geht es vor allem um Klassifikation, Vorhersage oder Clusterbildung. Man unterscheidet zwischen *supervised* und *unsupervised Learning*. Ersteres nutzt eine Zielvariable zur Steuerung des Algorithmus; mit Hilfe der Trainingsdaten wird nun ein Data Mining-Algorithmus ausgeführt und trainiert. Das Ergebnis ist ein Modell, das mit Hilfe der Testdaten geprüft werden kann. Alle angebotenen Data Mining Algorithmus-Seminare zeigen den vollständigen Prozess von der Datenbereitstellung, der Datenaufbereitung, der Zerlegung in Trainings- und Testdaten bis hin zur Anwendung des jeweiligen Data Mining Algorithmus'. Dabei wird die Sprache *R* zur Umsetzung verwendet. Als Ergebnis erhalten alle Teilnehmer ein *R*-Skript, welches sie später bei der Arbeit verwenden können.

K-NÄCHSTE NACHBARN



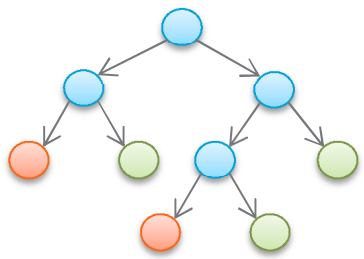
Die Cluster sind zunächst vorgegeben. Eine neue Beobachtung überprüft nun die Abstände zu den nächsten Nachbarn. Sobald diese bekannt sind, wird ausgezählt. Das Objekt wird jenem Cluster mit den meisten Treffern zugeordnet.

NAÏVE BAYES KLASSIFIKATION



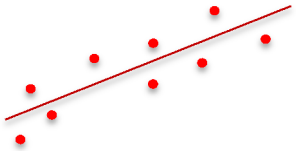
Hier geht es um Wahrscheinlichkeiten: Für jedes Objekt wird die Wahrscheinlichkeit ermittelt, mit der es einer bestimmen Klasse zugeordnet wird. Die Berechnung der Wahrscheinlichkeit erfolgt mit Hilfe des Satzes von Bayes (Thomas Bayes, englischer Theologe und Mathematiker aus dem 18. Jahrhundert). Um diese Rechnung zu vereinfachen, wird eine Unabhängigkeit der Attribute eingeführt – darum wird der Begriff *naïve* in diesem Zusammenhang verwendet.

ENTSCHEIDUNGSBÄUME



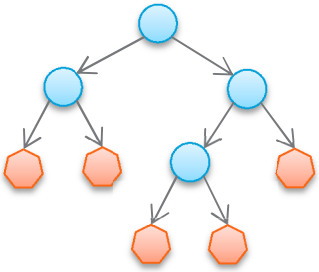
Entscheidungsbäume werden wegen ihrer Einfachheit und Verständlichkeit sehr gerne eingesetzt. Ich nutze sie ebenfalls gerne zur Modellfindung. Letztendlich besteht ein Baum (*tree*) aus Knoten (*nodes*) und Verbindungen (*branch*). An jedem Knoten wird eine Entscheidung gefällt (z. B. *Umsatz < 200 €*) und je nach Ausgang wird ein neuer Knoten betrachtet. Irgendwann kommt man zum Ende der Verzweigungen und landet auf einem *Blatt (leave)*. Dieses Blatt enthält nun die konkrete Entscheidung (*kauft/kauft nicht*).

LINEARE REGRESSION



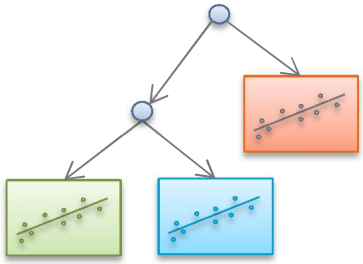
Regressionsmodelle können in der Regel sehr komplex ausfallen. Möchte man konkrete Aussagen über Parameter erhalten, werden bevorzugt klassische Techniken angewendet. Bei Klassifikationen sind moderne Data Mining Algorithmen den klassischen Methoden überlegen. Das Problem besteht in der Modellfindung, um Überanpassungen zu vermeiden. Dies ist ein zentrales Thema dieses Seminars.

REGRESSIONSBÄUME



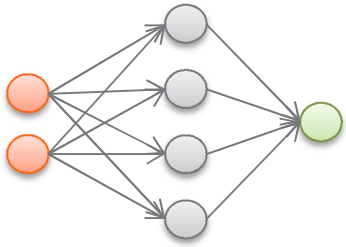
Regressionsbäume sind ebenfalls Entscheidungsbäume. Während Entscheidungsbäume zur Klassifikation (Käufer/Nichtkäufer) dienen, erlauben Regressionsbäume eine noch feinere Aussage. So kann z. B. das Ergebnis eines Regressionsbaums die Angabe der Bestellwahrscheinlichkeit oder ein Score sein. Mit Hilfe dieser Angaben können nun feiner abgestimmte Entscheidungen getroffen werden (z. B. Selektion von Kunden)

MODELLBÄUME



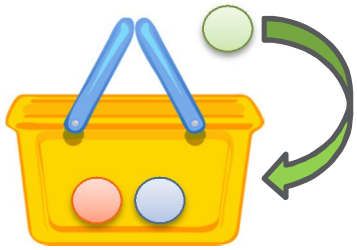
Modelbäume sind Entscheidungsbäume zur Auswahl eines Modells. An jedem Blattende befindet sich ein Regressionsmodell. Je nach Ausprägung der Attribute wird ein Regressionsmodell ausgewählt und die Vorhersage getroffen. Abhängig von der Fragestellung kann ein Modellbaum einige wenige Modelle (<10) aber auch viele Modelle (>100) enthalten.

NEURONALE NETZE



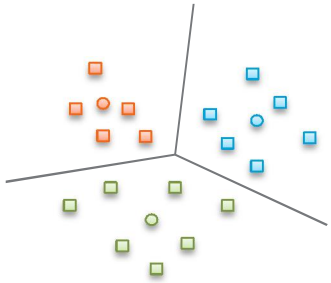
Das ist ein Reizthema! Heute werden *neuronale Netze* beim autonomen Fahren und bei Übersetzungen mit Erfolg eingesetzt. Man muss sich jedoch darüber im Klaren sein, dass hinter diesen Anwendungen harte Fakten stehen und z. B. beim autonomen Fahren ständig Sensoren die unmittelbare Umgebung scannen. Dies gilt nicht für sozialwissenschaftliche Fragestellungen - manchmal funktioniert‘s und manchmal nicht. Neuronale Netzte stellen sich für den Anwender als *Blackbox* dar. In diesem Seminar wird das Geheimnis des neuronalen Netzes gelüftet. So wird ein *Backpropagation Netz* erstellt und die Funktionsweise im Detail erläutert. Wenn man weiß, wie es geht, verliert das neuronale Netz zumindest einen Teil seiner Mystik.

WARENKORBANALYSE – ASSOZIIERUNGSREGELN



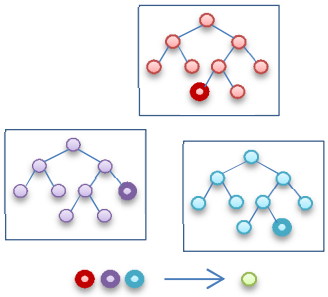
Die *Warenkorbanalyse* liefert Assoziierungsregeln, welche den meisten Anwendern bekannt sind, auch wenn ihnen nicht bewusst ist, wie sie funktionieren. Bekannt ist das Beispiel von Amazon: *Kunden, die Produkt A gekauft haben, können sich auch für Produkt XY interessieren.* In einem Warenkorb befinden sich häufig Produkt *A* und *B* gemeinsam. Hier lernen die Teilnehmer, wie sie bestimmte Produktkombinationen finden und dann möglicherweise gemeinsam präsentieren. Auf Webseiten können dann bei bestimmten Produkten auch weitere Produkte empfohlen werden, die gerne zusammen gekauft werden.

K-MITTELWERT CLUSTER



Bei diesem Verfahren wird zunächst die gewünschte Anzahl an Klassen und optional eine Kostenfunktion vorgegeben. Ausgangspunkt sind hier *k zufällig gewählte Mittelwerte*. Davon ausgehend wird jedes Objekt nun demjenigen Cluster zugeordnet, das die Summe der Abweichungsquadrate zum jeweiligen Mittelwert minimiert. Aus der erhaltenen Lösung wird erneut der Mittelpunkt des Clusters (Centroid) ermittelt und der Prozess so lange wiederholt, bis das Abbruchkriterium erreicht wurde. Auf Grund der zuvor gewählten Startwerte müssen nicht unbedingt die besten Lösungen gefunden werden. Auch hier erleichtern moderne PC das Auffinden guter Lösungen. Das Ergebnis kann zu differenzierten Werbezwecken oder für andere Aufgaben genutzt werden.

RANDOM FOREST



Der Ausgangspunkt bei der Modellerstellung ist der Begriff *Bootstrapping*. Dabei werden aus den Trainingsdaten Stichproben der Größe *n* mit Zurücklegen gezogen. Aus einer Stichprobe mit *m*- Attributen wird eine Teilmenge zufällig (*random*) gewählt. Auf Basis dieser Teilmenge wird dann der Entscheidungsbaum gebildet. Für die Zuordnung eines neuen Elements wird für jede abgebildete Attribut-Kombination der entsprechende Entscheidungsbaum ausgewertet und jene Klasse gewählt, welche die meisten Stimmen erhalten hat. Verwendet man anstelle von Entscheidungsbäumen Regressionsbäume, ergibt sich der Vorhersagewert durch Gewichtung.

Data Science ist eine junge Disziplin. Diese Wissenschaft nutzt Technolgien aus den Gebieten Data Mining und maschinelles Lernen, um Muster in Ihren Daten zu entdecken, die zu neuem Wissen führen und in der Folge dem Unternehmen neue Möglichkeiten eröffnen.

"Data Science ist nur was für Google, Facebook, Amazon ..."

Stimmt nicht!

Auch *KMU* können von diesen Technologien profitieren.
Open Source Software steht jedem *kostenfrei* zur Verfügung.
Es sind dazu keine Cloud-Dienste erforderlich.
Das *Knowhow* stellen wir für Sie bereit.

Wir machen Sie fit für die Nutzung Ihrer Daten!

Externe Seminare und Workshops:
- 140 € je Teilnehmer
- Teilnahmezertifikat
- ab 6 Teilnehmer

Inhouse Seminare und Workshops:
- 400 € (bis zu 3 Teilnehmer)
- jeder weitere Teilnehmer 80 €
- Teilnahmezertifikat
- zuzügl. Fahrtkosten und Nächtigung

Die angeführten Preise sind Netto-Preise zuzügl. 20% MWSt.

USEDATA

USEDATA
Mag. Helmut Grillenberger
Georg Rendl Weg 31b
5111 Bürmoos

www.usedata.com
+43 6274 20804