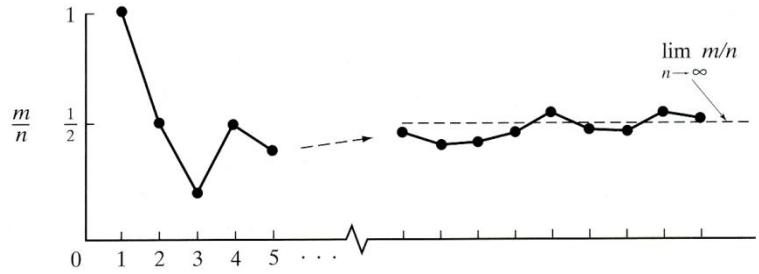
Ich starte mit diesem Beitrag meine R-Initiative im Web. Wohin die Reise geht, kann ich nicht vorhersagen. Ich lade jeden, der sich für R interessiert ein, an dieser Reise teilzunehmen – auch ohne Vorkenntnisse. Wöchentlich werde ich einen Beitrag auf meiner Homepage www.usedata.com bereitstellen und - mit etwas Verzögerung - auch auf **LinkedIn**.

Zunächst möchte ich zu den Begriffen Statistik, Informatik, Data Mining, maschinelles Lernen, Big Data und Data Science Stellung nehmen.

Die Statistik ist das Teilgebiet der Mathematik, welches sich mit dem Zufall beschäftigt. Werfen wir z. B. eine Münze n-mal und beobachten dabei m-mal das Ereignis, dass die Münze Kopf zeigt, dann können wir die Wahrscheinlichkeit für dieses Ereignis mit p = m/n schätzen. Wir würden erwarten, dass p sich dem Wert ½ annähert, sofern n groß genug gewählt wurde.



Wie oft muss man das Experiment ausführen, um sicher eine korrekte Antwort zu erhalten? Das ist nicht vorhersehbar. Wenn Kopf und Zahl die gleiche Wahrscheinlichkeit haben, einzutreten, dann kennen wir die Antwort: p = ½. Dazu benötigt man nicht zwingend ein Experiment. Doch nicht jede Aufgabe lässt sich so einfach lösen.

Trotzdem enthält diese experimentelle Vorgangsweise bereits die Ansätze des maschinellen Lernens.

Als man über solche Ansätze diskutierte, gab es noch keine Computer. Dafür gab es aber kluge Köpfe. Einer davon war der Russe Andrei Kolmogorov. Zu Beginn des 20 Jahrhunderts legte er mit nur vier einfachen Axiomen den Grundstein für die Wahrscheinlichkeitsrechnung und damit für die heutige Statistik. Ein Axiom wird nicht hinterfragt. Es handelt sich dabei um Aussagen, die dem Hausverstand nicht widersprechen. Statistik ist damit eine Wissenschaft, die auf klaren Regeln aufbaut.

Früher wurden solche Experimente tatsächlich von Hand ausgeführt. Heute würde ein moderner Computer in Bruchteilen einer Sekunde diese Experimente millionenfach ausführen und das Ergebnis zur Anzeige bringen. Trotzdem wäre das Ergebnis in der Regel mit einem Fehler behaftet. Die Frage ist, ob dieser Fehler für praktische Entscheidungen vernachlässigbar ist.

Damit komme ich zur Informatik. Diese hat in den letzten zehn Jahren enorme Fortschritte gemacht. Die Computer wurden immer schneller und der Speicherplatz immer größer und günstiger. Damit

war der Weg frei für die Anwendung heuristischer Methoden, um statistische Fragestellungen und Probleme zu lösen.

Heuristik kommt aus dem altgriechischen heurísko (ich finde) bzw. heurískein (Auffinden und Entdecken). Als Heuristik bezeichnet man (nach Wikipedia) also die Kunst, aus unvollständigen Informationen dennoch zu wahrscheinlichen Aussagen oder praktikablen Lösungen zu kommen.

Heuristiken führen immer zu einem Ergebnis. Man kann sich jedoch nicht sicher sein, ob die gefundene Lösung auch die optimale Lösung darstellt. In vielen Fällen werden optimale Lösungen gefunden. Für heuristische Verfahren bedarf es großer Datenmengen.

Ein Spezialgebiet der Informatik ist das maschinelle Lernen. Hier kommen Data Mining Algorithmen wie z. B.

- Naïve Bayes
- Entscheidungsbäume
- Regressionsmodelle
- Support Vektor Maschinen,
- Neuronale Netze
- Warenkorbanalysen
- Clusterverfahren etc.

zum Einsatz. Dabei werden die zu analysierenden Datenvolumen immer größer, die möglichen Strukturen immer vielfältiger und dies in einem nicht endenden Datenstrom - besser bekannt unter dem Begriff Big Data.

Data Science ist ein Sammelbegriff für Statistik, Data Mining, maschinelles Lernen und Big Data etc. Eine Person, die sich mit diesem Themenbereich beschäftigt, bezeichnet man als Data Scientist.

Ein Data Scientist muss also vielfältige Erfahrungen mit sich bringen. Er sollte über statistisches Knowhow verfügen, jedoch auch Erfahrung im IT-Bereich mitbringen. Gerade der Umgang mit Datenbanken, BI-System, der Cloud und natürlich Big Data etc. gehört zum Werkzeugkasten. Hinzu kommt, dass sich diese Themenlandschaft kontinuierlich im Wandel befindet. Ein Data Scientist muss immer aufgeschlossen für Neues sein und sich dem Wandel stetig anpassen.

Neben den technischen Fähigkeiten zur Datenhaltung und Datennutzung benötigt der Data Scientist auch entsprechende Analysewerkzeuge. Hier kommen die Sprachen R und Python zum Zug. Bei beiden Sprachen handelt es sich um Open Source Software.

Ich werde mich in diesem Blog hauptsächlich mit R beschäftigen, jedoch auch Ausflüge in andere Themenbereiche machen!

Als ich Statistik studierte, kamen die ersten PCs auf den Markt und waren für die meisten unerschwinglich. Die Informatik beschäftigte sich damit mit Edgar F. Codd's Regeln für eine relationale Datenbank. Alles war noch überschaubar.

Heute hat sich die Statistik, dank der Fortschritte in der Informatik, grundlegend verändert.

Was sind die Unterschiede zwischen Statistik und Informatik? Eine Frage, die mich im Rahmen meines Berufslebens immer wieder beschäftigt hat. Ich habe diese Thematik mit verschiedenen

Fachleuten, auch aus dem universitären Bereich, diskutiert, um mir so eine Meinung bilden zu können. Heute sehe ich die Situation wesentlich entspannter. Beide Bereiche haben ihre Berechtigung; es hängt alles von der Art der Fragestellung und der Sichtweise ab. Manche Fragestellungen bevorzugen die Statistik und andere die Informatik. Einen eindeutigen ‚Gewinner' gibt es hier nicht.

Ein Beispiel aus der Praxis.

Für einen Versender wird ein Scoringmodell erstellt. Ziel ist es, jene Kunden zu selektieren, die auf einen Katalog mit hoher Wahrscheinlichkeit bestellen werden. Für den Selektionsmanager geht es also um eine *kauft/kauft nicht* Entscheidung. Eine Antwort auf diese Frage gibt folgendes chinesisches Sprichwort:
Egal, ob die Katze weiß oder schwarz ist, Hauptsache ist, sie fängt Mäuse.

Mit Hilfe des maschinellen Lernens und unter Einsatz eines Data Mining Algorithmus' wird die Entscheidung getroffen, ob ein Kunden kaufen wird oder nicht. Die Erklärung für den Kauf spielt keine Rolle - solange der Algorithmus optimale Ergebnisse liefert.

Maschinelles Lernen hat den Vorteil, dass es in vielen Fällen auf nichtlinearen Modellen aufbaut. Oft sind die Lösungen jedoch in einer Black Box versteckt. Wenn es nur um eine *Ja/Nein* Entscheidung geht, spielt dies eine untergeordnete Rolle.

Das Management möchte jedoch zu Planungszwecken mehr über das Kundenverhalten erfahren. Auch wenn die Statistik – durch einfachere Modellannahmen – geringfügig schlechtere Ergebnisse im Vergleich zum maschinellen Lernen liefert, kann sie entscheidend zur Klärung des Kundenverhaltens beitragen. Statistik unterstützt das Management bei der Frage, warum etwas so ist, wie es ist.

Im nächsten Beitrag werde ich – aus der Vogelperspektive – diese Unterschiede an Hand eines Kreditscoring-Modells verdeutlichen.

Mit statistischen Grüßen

Helmut Grillenberger

www.usedata.com
www.pulsmagic.com