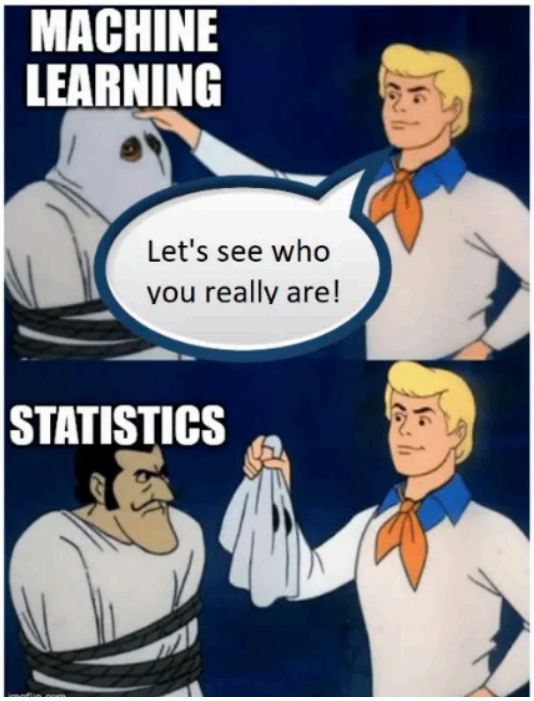
Klassische Statistik - Unterschätzte Basis der Data Science

by Woche 2

Der Begriff “Data Science” ruft häufig Bilder komplexer Algorithmen, Deep Learning und künstlicher Intelligenz hervor - all die “coolen” Elemente, die in Tech-Blogs und LinkedIn-Profilen glänzen. Im Gegensatz dazu wird klassische Statistik oft als die langweilige Schwester betrachtet: trocken, kompliziert und irgendwie altmodisch.

In der Branche gibt es sogar den Witz:

> Bei der Präsentation für Investoren spricht man von KI.
> Im Projektbericht spricht man von Machine Learning.
> Im eigentlichen Code wird lineare Regression verwendet.



Dieser Witz enthält mehr Wahrheit als viele zugeben möchten. Die Realität sieht nämlich so aus: Selbst die fortschrittlichsten Machine-Learning-Algorithmen bauen auf statistischen Grundprinzipien auf, die seit Jahrzehnten oder sogar Jahrhunderten bekannt sind. Und in vielen praktischen Anwendungsfällen liefern “einfache” statistische Methoden überraschend gute Ergebnisse - nicht selten genau so gut oder besser als komplexere Ansätze.

Dies soll natürlich nicht den Eindruck erwecken, dass fortgeschrittene Machine-Learning-Methoden keine Berechtigung hätten oder allen Data Scientists die nötigen Kompetenzen fehlen. Im Gegenteil - diese Technologien haben beeindruckende Fähigkeiten, die klassische Statistik allein nicht bieten kann. Der Punkt ist vielmehr, dass ohne ein solides statistisches Fundament selbst die fortschrittlichsten Algorithmen auf wackeligem Boden stehen.

Warum klassische Statistik in Data Science unverzichtbar ist

Es gibt mehrere Gründe, warum ein solides Verständnis klassischer Statistik für jeden Data Scientist unerlässlich ist:

1. Fundament für komplexe Methoden

Viele moderne Machine-Learning-Algorithmen sind im Kern Erweiterungen statistischer Konzepte:

- **Random Forests** sind Ensembles von Entscheidungsbäumen, die auf Konzepten wie Varianz und Bias basieren
- **Deep Learning** baut auf statistischen Konzepten wie Maximum-Likelihood-Schätzung auf
- **Regression** ist nicht nur ein statistisches Verfahren, sondern auch die Grundlage für viele ML-Modelle, von linearen Modellen bis hin zu neuronalen Netzen

Die mathematischen Grundlagen moderner KI-Technologien entstammen der statistischen Tradition: Bayes'sche Inferenz, Wahrscheinlichkeitstheorie und statistische Verteilungen sind die Bausteine, auf denen die meisten algorithmischen Innovationen aufbauen.

2. Explorative Datenanalyse

Bevor wir überhaupt an komplexe Modelle denken können, müssen wir unsere Daten verstehen. Hier sind statistische Methoden unverzichtbar:

- Zusammenhänge zwischen Variablen identifizieren
- Ausreißer erkennen
- Hypothesen zu den Daten formulieren und testen

Diese grundlegenden statistischen Konzepte sind die Voraussetzung für jedes erfolgreiche Data-Science-Projekt - egal wie fortschrittlich die später verwendeten Algorithmen sein mögen.

3. Modellevaluation und Hypothesentests

Woher wissen wir, ob unser Modell gut ist oder ob ein bestimmter Effekt in den Daten tatsächlich relevant ist? Klassische statistische Tests geben uns diese Antworten:

- t-Tests und ANOVA für Gruppenunterschiede
- Konfidenzintervalle zur Einschätzung der Unsicherheit
- p-Werte zur Bewertung statistischer Signifikanz
- Validierungsmetriken wie R^2, RMSE, AUC etc.

Ohne diese Grundlagen ist es schwierig, die Qualität und Zuverlässigkeit unserer Modelle und Ergebnisse zu beurteilen.

4. Interpretierbarkeit und Kommunikation

Ein häufig unterschätzter Aspekt: Die Ergebnisse müssen Stakeholdern vermittelt werden. Hier glänzen statistische Methoden:

- Einfache lineare Modelle sind leichter zu erklären als Deep Learning
- Statistische Kennzahlen liefern klare Anhaltspunkte für Entscheidungen
- Konfidenzintervalle kommunizieren die Unsicherheit transparent

In einer Geschäftswelt, die zunehmend evidenzbasierte Entscheidungen fordert, ist die Fähigkeit, Ergebnisse klar zu kommunizieren, mindestens genauso wichtig wie die Fähigkeit, komplexe Modelle zu erstellen.

Praxis vs. Theorie in Data Science

Im realen Projektalltag zeigt sich immer wieder ein interessantes Phänomen: Trotz des Hypes um KI und Deep Learning kommen oft einfachere statistische Methoden zum Einsatz. Die Gründe hierfür sind vielfältig:

- **Datenmenge**: Viele Unternehmen haben nicht die Millionen von Datenpunkten, die Deep Learning bräuchte
- **Interpretierbarkeit**: Gerade in regulierten Branchen müssen Entscheidungen nachvollziehbar sein
- **Effizienz**: Einfachere Modelle sind schneller zu entwickeln, zu trainieren und zu deployen
- **Wartbarkeit**: Komplexere Modelle bedeuten mehr Aufwand in der Pflege und Überwachung

Die richtige Balance finden

Als Data Scientist ist es nicht die Frage, ob man sich mit klassischer Statistik oder mit modernen ML-Methoden beschäftigen sollte - sondern wie man beides sinnvoll kombiniert. Ein guter Data Scientist:

1. Beginnt mit statistischer Exploration, um die Daten zu verstehen
2. Wählt Methoden basierend auf dem Problem, nicht auf dem aktuellen Hype
3. Validiert komplexere Modelle gegen einfachere statistische Baselines
4. Kommuniziert Ergebnisse klar und mit angemessener Darstellung der Unsicherheit

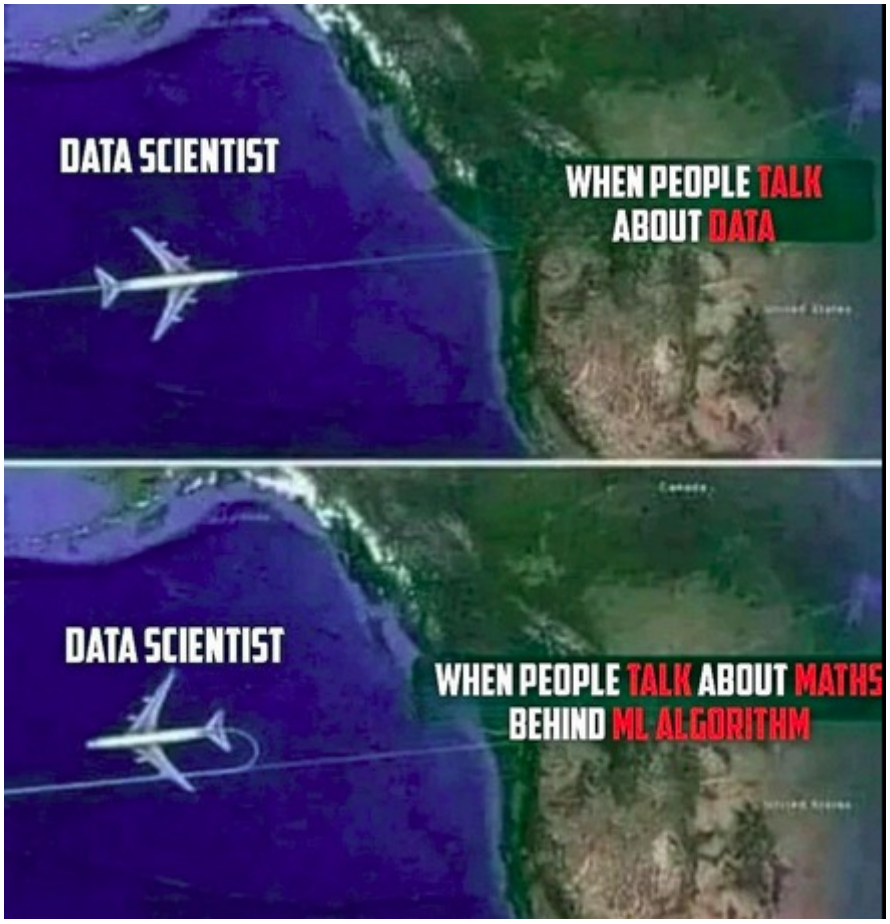
Es gibt einen Grund, warum selbst Unternehmen an der Spitze der KI-Innovation wie Google und Meta ihre Datenteams mit klassisch ausgebildeten Statistikern besetzen: Das Fundament muss stimmen, bevor das Hochhaus gebaut werden kann.

Ausblick auf die folgenden Kapitel

In den kommenden Kapiteln werden wir uns mit einigen der wichtigsten Konzepte der klassischen Statistik beschäftigen:

- **Korrelation** als Maß für Zusammenhänge zwischen Variablen
- **Regression** und ihre vielfältigen Anwendungen
- **Statistische Tests** zur Überprüfung von Hypothesen

Die Zeit, die wir in ihr Verständnis investieren, wird sich mehrfach auszahlen - sowohl in der Qualität unserer Analysen als auch in unserer Effektivität als Data Scientists. Und am Ende trifft dies hoffentlich nicht auf euch zu:


DATA SCIENTIST
WHEN PEOPLE TALK ABOUT DATA
DATA SCIENTIST
WHEN PEOPLE TALK ABOUT MATHS BEHIND ML ALGORITHM