Topic 6: Overfitting

Block 1: Data Science Grundlagen

Das Overfitting-Problem

Definition:

- Modell lernt Trainingsdaten *zu gut* – inklusive Rauschen
- Erfasst zufällige Muster statt echter Zusammenhänge

Symptome:

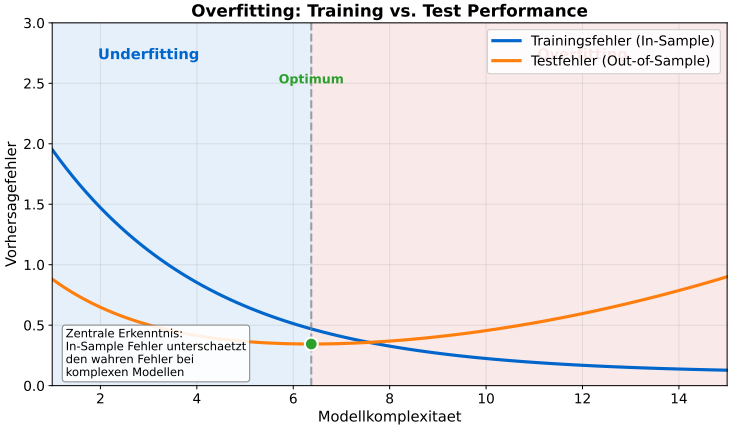
- Gute In-Sample Performance (hoher R^2)
- Schlechte Out-of-Sample Performance (niedriger R^2 auf neuen Daten)
- Instabile Modelle – kleine Datenänderungen führen zu großen Koeffizientenänderungen
- Extreme oder unplausible Koeffizienten

Ursachen:

- Zu viele Variablen im Verhältnis zur Datenmenge
- Zu komplexes Modell (z.B. Polynome hohen Grades)
- Zu wenig Trainingsdaten

Overfitting = Modell lernt Rauschen statt Signal

Overfitting Visualisierung



Training vs. Test Error:

- Training Error sinkt monoton mit Modellkomplexität
- Test Error hat U-Form – steigt bei zu hoher Komplexität
- Optimales Modell minimiert Test Error

Das beste Modell ist nicht das komplexeste – Bias-Variance Tradeoff

In-Sample vs. Out-of-Sample Performance

In-Sample:

- Performance auf Trainingsdaten
- Immer optimistisch
- R^2 steigt mit jeder Variable
- Nicht zur Modellwahl geeignet!

Beispiel:

- 5 Variablen: $R^2 = 0.65$
- 10 Variablen: $R^2 = 0.78$
- 20 Variablen: $R^2 = 0.92$

Out-of-Sample:

- Performance auf neuen Daten
- Realistischer Maßstab
- Kann bei Overfitting sinken
- Basis für Modellwahl

Methoden:

- Train/Test Split (z.B. 70/30)
- Kreuzvalidierung (Block 3)
- Hold-out Validierung

Modelle immer auf neuen Daten validieren – In-Sample Performance täuscht

Zusammenfassung Block 1 & Kernbotschaft

Block 1 – Zentrale Konzepte:

- Data Science = Statistik + Computer Science + Domain Expertise
- **CRISP-DM:** Strukturierter Prozess von Business Understanding bis Deployment
- **EDA vor Modellierung:** Daten verstehen, visualisieren, Ausreißer identifizieren
- **Lineare Regression:** Grundmodell für Vorhersagen – Interpretation, R^2, Residuen
- **Overfitting:** Gefahr bei zu komplexen Modellen – Out-of-Sample Validierung

Kernbotschaft:

Data Science ist modellbasiertes Denken mit dem Ziel der Generalisierung – nicht des perfekten Fits auf Trainingsdaten.

Ausblick Block 2:

- Statistische Inferenz – Unsicherheit quantifizieren
- Bootstrapping & Konfidenzintervalle
- Hypothesentests & Effektstärken

Block 1 abgeschlossen – nächste Woche: Inferenz und statistische Tests