

2. Cross Validation

Grundlegender Ansatz:

Wir teilen einen Datensatz auf

Trainingsdaten

Hier wird ein statistisches Modell bestimmt

(Lin Reg, Log Reg, ...)

Testdaten

Hier wird das Modell angewandt

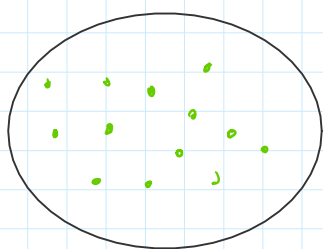


"Test Error Rate": Fehler bei der Prognose der Testdaten

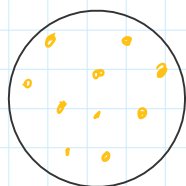
"Training Error Rate": Fehler bei der Prognose der Trainingsdaten

Gemessen werden beide Fehler durch den Mean Squared Error (MSE)

2.1 Validation Set Ansatz



Trainingsset N_T



Validation Set N_V

- In der Regel gilt: $N_V < N_T$
- Für verschiedene Samples von Validation Sets kann der MSE deutlich abweichen
- Anpassung wird nur an einem Trainingsset

→ Teile den Datensatz zufällig in Trainingsset und Validationset auf

→ Führe die statistische Methode (z.B. Lin Reg) auf dem Trainingsset durch

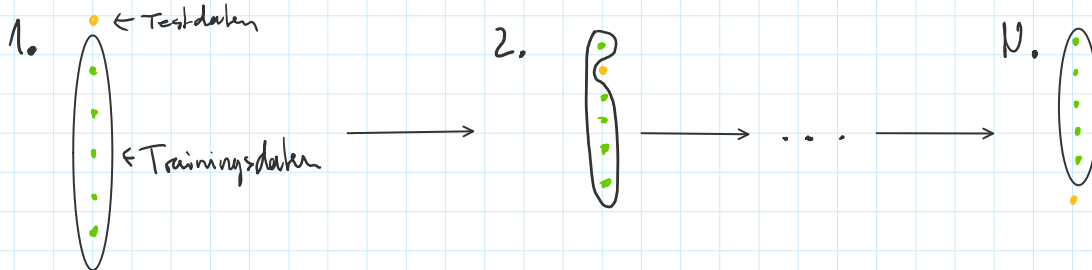
→ Erstelle mit Hilfe des bestimmten Modells eine Prognose \hat{y}_i für jeden y_i des Validation Sets, und bestimme $(y_i - \hat{y}_i)^2$

$$\Rightarrow \text{MSE} = \frac{1}{N_V} \sum_{i=1}^{N_V} (y_i - \hat{y}_i)^2$$

- Anpassung wird nur an einem Trainingsset vorgenommen

$$\Rightarrow MSE = \frac{1}{N} \sum_{i=1}^{N_v} (y_i - \hat{y}_i)^2$$

2.2 LOOCV ("Leave one out Cross Validation")



\rightarrow Für jede der N Durchführungen rechne $(y_i - \hat{y}_i)^2$ und bestimme am Ende $MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$

Vorteile: Modell wird an "mehr" Daten angepasst

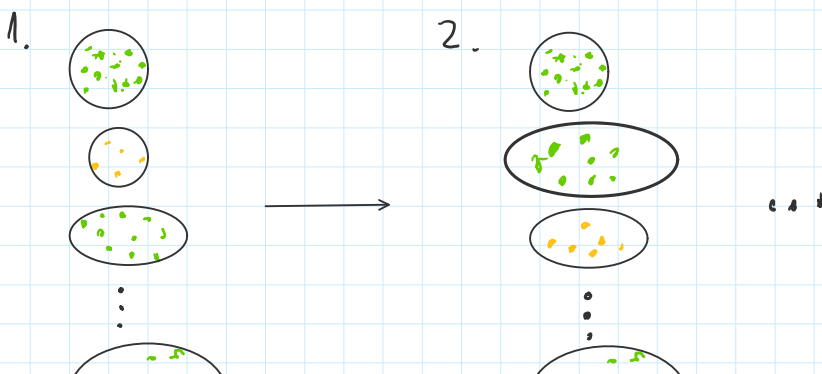
\Rightarrow Test Error Rate wird weniger überschätzt als im Validation Set Ansatz

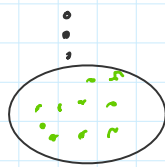
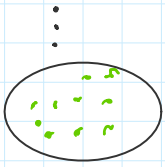
- Die Aufteilung in Trainings- und Testset ist schematisch (nicht mehr zufällig)

\Rightarrow MSE hängt nicht mehr von der zufälligen Aufteilung ab

Nachteil: N -malige Ausführung kann sehr rechenintensiv sein

2.3 k-fache Cross Validation





- Teile den Datensatz in h Teildatensätze auf
- jeder der h Teildatensätze wird in einem Schritt als Testdatensatz, die anderen $(h-1)$ Teildatensätze werden gemeinsam d/s Trainingsdatensatz verwendet

$$\Rightarrow MSE = \frac{1}{h} \sum_{i=1}^h MSE_i$$

MSE_i : MSE für die i -te Durchführung