

# Lernzettel

## Überwachtes Lernen: Klassifikation, Regression, Modellbewertung

**Universität:** Technische Universität Berlin  
**Kurs/Modul:** Informationssysteme und Datenanalyse  
**Erstellungsdatum:** September 19, 2025



Zielorientierte Lerninhalte, kostenlos!  
Entdecke zugeschnittene Materialien für deine Kurse:

<https://study.AllWeCanLearn.com>

Informationssysteme und Datenanalyse

**Lernzettel: Überwachtes Lernen: Klassifikation, Regression, Modellbewertung**

(1) **Überblick.** Beim überwachten Lernen geht es darum, aus gegebenen Daten  $(\mathbf{x}_i, y_i)$  eine Vorhersagefunktion  $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$  abzuschätzen, sodass neue Eingaben  $\mathbf{x}$  möglichst korrekt beschrieben werden.

(2) **Grundidee des überwachten Lernens.** Gegeben Trainingsdaten  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , wähle eine Hypothesenklasse  $\mathcal{H}$  und minimiere

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(\mathbf{x}_i)) + \lambda R(f),$$

wobei  $\ell$  die Verlustfunktion ist und  $R(f)$  eine Regularisierung.

(3) **Klassifikation vs. Regression.**

- *Klassifikation:* Ziel  $y_i \in \{1, \dots, K\}$  (diskrete Klassen).
- *Regression:* Ziel  $y_i \in \mathbb{R}$  (stetige Werte).

(4) **Modellbewertung – Klassifikation.** Typische Metriken für ein Klassifikationsproblem:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1} = \frac{2 \text{Precision} \text{Recall}}{\text{Precision} + \text{Recall}}.$$

ROC-AUC = Fläche unter der ROC-Kurve.

Wichtige Begriffe: True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN).

(5) **Modellbewertung – Regression.** Gängige Metriken:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

mit  $\bar{y} = \frac{1}{n} \sum y_i$ .

(6) **Validierung und Bias-Varianz.** Um Überanpassung zu vermeiden, nutzt man Validierungstechniken:

- *Cross-Validation* (z.B. *k-fach*): Dataset wird in  $k$  Falten geteilt; jeweils eine Falte Tests, Resten Training.
- *Train/Validation/Test-Aufteilung*: Training zum Anpassen, Validation zur Auswahl der Modelle/Hyperparameter, Test zur abschließenden Beurteilung.

Darstellung des Fehlers als Trade-off:

$$\text{Fehler} \approx \text{Bias}^2 + \text{Varianz} + \text{Irreduzierter Fehler.}$$

### (7) Lernprozess – Training, Validierung, Test.

- Training: Modellparameter  $\theta$  werden geschätzt.
- Validierung: Hyperparameter  $\lambda$ , Lernrate etc. angepasst.
- Test: Endbewertung auf unabhängigen Daten.

Typischer Workflow: Splitte Daten, trainiere, bewerte, wähle das beste Modell, reportiere Metriken.

### (8) Hyperparameter-Tuning. Strategien zur Suche idealer Hyperparameter:

- Grid Search: systematisches Durchlaufen definierter Werte.
- Random Search: zufällige Auswahl von Werten.
- Bayesian Optimization: sequentiales Optimieren anhand von Evidenz.

Beachtung: geringe Rechenlast vs. gute Abdeckung des Suchraums.

### (9) Typische Algorithmen (Klassifikation).

- Logistische Regression (linear trennbar oder reguliert).
- K-Nearest Neighbors (KNN).
- Entscheidungsbaum, Random Forest.
- Gradient Boosting (LightGBM, XGBoost).
- Support Vector Machines (SVM) mit Kernel.

### (10) Typische Algorithmen (Regression).

- Lineare Regression, inklusive Ridge/Lasso.
- Entscheidungsbaum-Regressor, Random Forest Regressor.
- Gradient Boosting Regressor.
- Support Vector Regressor (SVR).

### (11) Datenvorverarbeitung.

- Skalierung: Standardisierung (zweiseitig zentriert), Robustskalierung.
- Kodierung kategorialer Merkmale: One-Hot Encoding.
- Umgang mit Ausreißern, Missing Values.

**(12) Hinweise zur Praxis.**

- Vermeide Overfitting durch Regularisierung, frühzeitiges Stoppen, Cross-Validation.
- Wähle passende Metriken zur Aufgabenstellung.
- Berücksichtige Klassenunbalancen bei Klassifikation.

**(13) Beispiel – einfache Klassifikation (Konzept).** Gegeben sei ein Datensatz mit zwei Merkmalen  $x_1, x_2$  und binärer Zielgröße  $y \in \{0, 1\}$ .

- Wähle eine Hypothesenklasse, z.B. lineare Trennung  $f(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$  mit  $\sigma(z) = \frac{1}{1+e^{-z}}$ .
- Trainiere mit der logistischen Verlustfunktion

$$\ell(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}),$$

- Beurteile das Modell mit Accuracy, ROC-AUC und ggf. F1.