Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

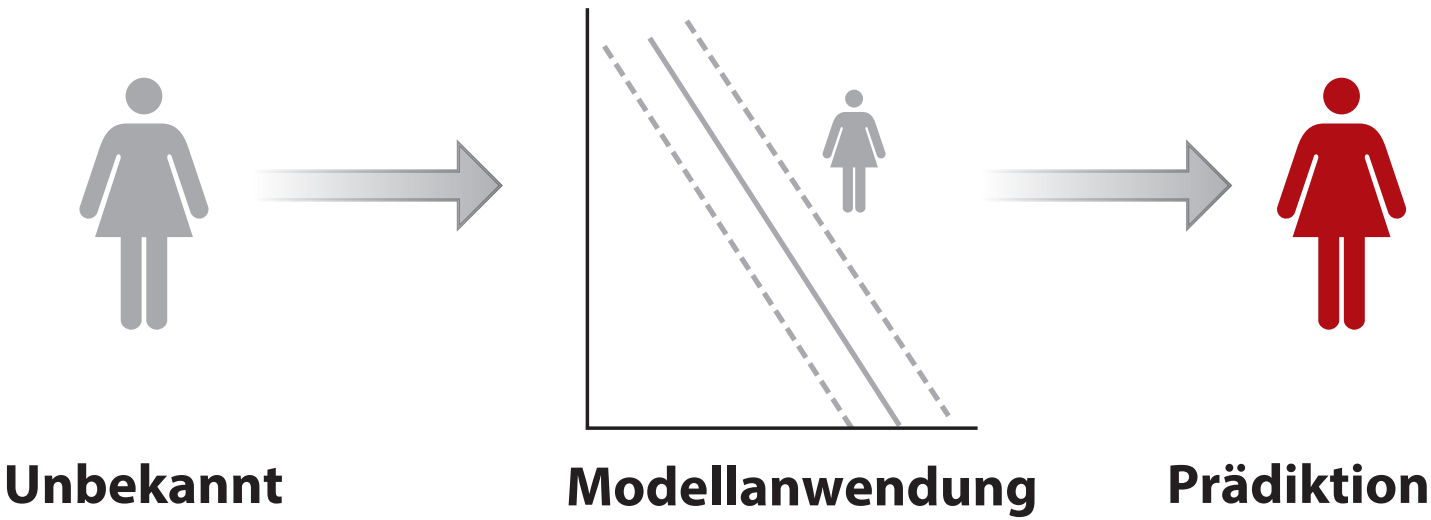
(8) Prädiktion und Kreuzvalidierung

Prädiktive Modellierung und Maschinelles Lernen

Kreuzvalidierung

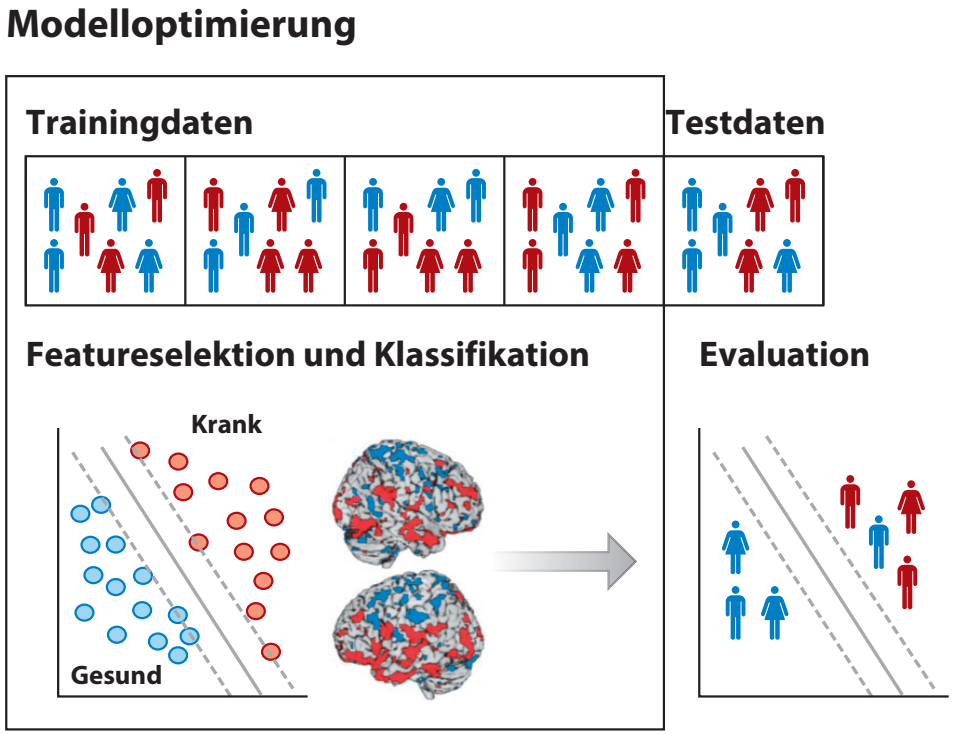
Selbstkontrollfragen

Rhetorik der Prädiktiven Modellierung und des Maschinellen Lernens



Dwyer, Falkai, and Koutsouleris (2018)

Rhetorik der Prädiktiven Modellierung und des Maschinellen Lernens



Dwyer, Falkai, and Koutsouleris (2018)

Daten	Trainingsdaten und Testdaten
Statistisches Modell	Modell, Machine Learning Algorithmus
Schätzen von Parametern	Trainieren des Modells, Parameterlernen, Supervised Learning

Definition (Binärer Klassifikationstrainingdatensatz)

Ein *binärer Klassifikationsdatensatz*

$$\mathcal{D} := \{(x_1, y_1), ..., (x_n, y_n)\} = \{(x_i, y_i)\}_{i=1}^{n} \tag{1}$$

ist eine Menge von n *Trainingsdatenpunkten*

$$(x_i, y_i) \text{ mit } x_i \in \mathbb{R}^m \text{ und } y_i \in \{0, 1\} \text{ for } i = 1, ..., n, \tag{2}$$

wobei x_i m-*dimensionaler Featurevektor* und y_i *Label* genannt wird. Üblicherweise werden die Trainingsdatenpunkte dabei als unabhängige und identische Realisierungen eines Zufallsvektors $m + 1$-dimensionalen Zufallsvektors $\zeta := (\xi, \upsilon)$ verstanden.

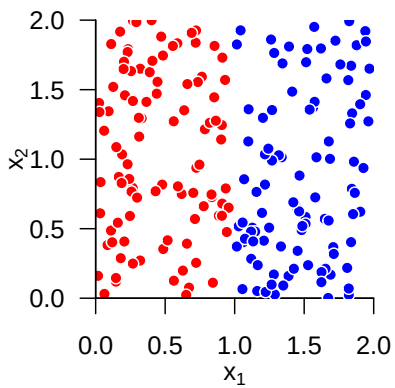
Bemerkungen

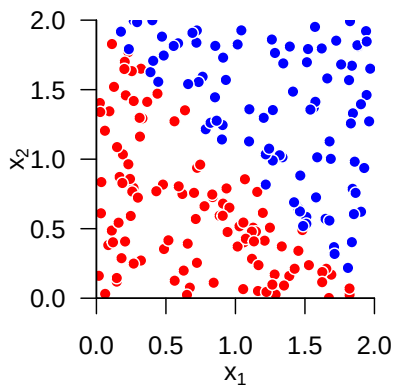
- $y_i \in \{0, 1\}$ bezeichnet die Klassenzugehörigkeit des Featurevektors $x_i \in \mathbb{R}^m$.
- Ein Beispiel für y_i ist “Kein Therapieerfolg” (0)” vs. “Therapieerfolg” (1).
- Beispiele für die m Komponenten der x_i sind Testscores, Biomarker, Soziodemographische Daten.

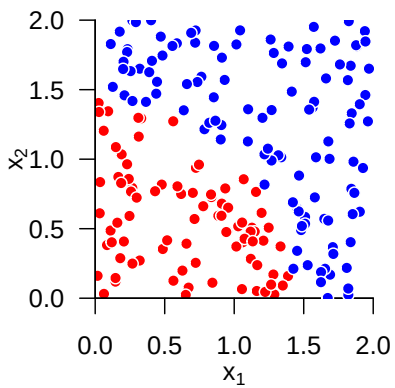
Beispiele bivariater Featureplotszenarien

$$x_i \in \mathbb{R}^2, y_i \in \{0, 1\}, i = 1, ..., n, \bullet\ y_i = 0, \quad \bullet\ y_i = 1$$

$$x_{i_1} > 1 \Leftrightarrow y_i = 1 \qquad x_{i_1} + x_{i_2} > 2 \Leftrightarrow y_i = 1 \qquad x_{i_1}^2 + x_{i_2}^2 > 2 \Leftrightarrow y_i = 1$$







Anwendung der prädiktiven Modellierung

Explanatorische Modellierung ⇔ Grundlagenforschung

Bestimmung von $\hat{\phi} \coloneqq \operatorname{argmin} \left\| \hat{\phi} - \phi \right\|$

Bestimmung von $\hat{f} \coloneqq \operatorname{argmin}_{f \in F} \| \upsilon - f(\xi) \|, F$ beliebig

Prädiktive Modellierung ⇔ Anwendungsorientierte Forschung

Shmueli (2010), Sainani (2014)

Prädiktive Modellierung und Maschinelles Lernen

Kreuzvalidierung

Selbstkontrollfragen

Workflow der prädiktiven Modellierung

Featureselektion

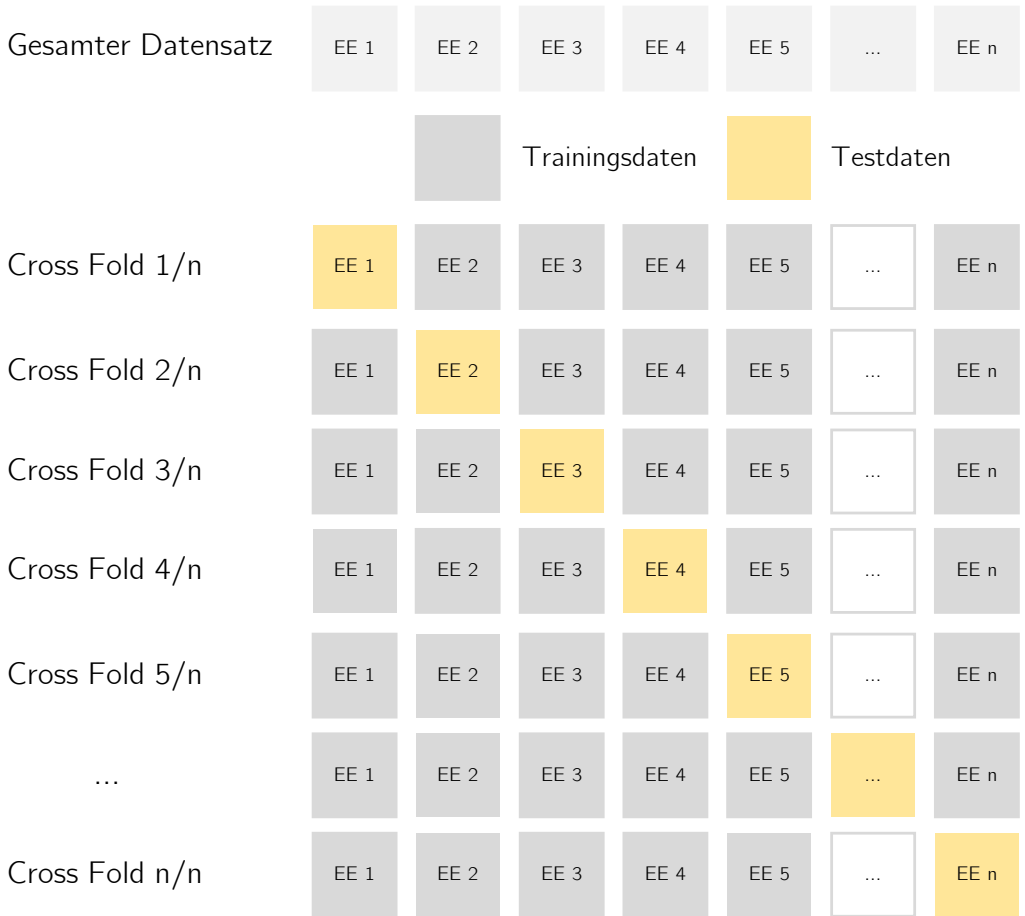
- Auswahl von möglichen prädiktiven Variablen
- Dimensionreduktion zur Verringerung des Curse of Dimensionality

Kreuzvalidierung

- Wiederholtes Trainieren und Testen eines Modells an einem Datensatz
- Einsatz zur Modelloptimierung
- Einsatz zur Messung der probabilistischen Assoziation von Featuren und Labeln

Kreuzvalidierung

Leave-One-Out-Crossvalidation (LOOCV) bei n experimentellen Einheiten (EE)



Kreuzvalidierung

Konfusionsmatrix bei LOOCV mit binärem Label für Testendatenpunkt (x_i, y_i)

Fall	Prädiktion $f(x_i) = 0$	Prädiktion $f(x_i) = 1$	
$y_i = 0$	Richtig Negativ r_n	Falsch Positiv f_p	Gesamt Negativ $r_n + f_p$
$y_i = 1$	Falsch Negativ f_n	Richtig Positiv r_p	Gesamt Positiv $f_n + r_p$
	Negative Prädiktion $r_n + f_n$	Positive Prädiktion $f_p + r_p$	

Exemplarische Performanzmaße bei LOOCV mit binärem Label

- Akkuratheit (Accuracy)

$$\text{ACC} = \frac{\text{Anzahl richtiger Prädiktionen}}{\text{Anzahl aller Prädiktionen}} = \frac{r_n + r_p}{r_n + r_p + f_n + f_p} \tag{3}$$

- Sensitivität (Richtig-positiv-Rate, True Positive Rate, Recall, Hit Rate)

$$\text{SEN} = \frac{\text{Anzahl richtiger Positivprädiktionen}}{\text{Anzahl positiver Fälle}} = \frac{r_p}{r_p + f_n} \tag{4}$$

- Spezifität (Richtig-negativ-Rate, True Negative Rate, Correct Rejection Rate)

$$\text{SPE} = \frac{\text{Anzahl richtiger Negativprädiktionen}}{\text{Anzahl negativer Fälle}} = \frac{r_n}{r_n + f_p} \tag{5}$$

Prädiktive Modellierung und Maschinelles Lernen

Kreuzvalidierung

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie die Rhetorik der Prädiktiven Modellierung.
2. Geben Sie die Definition eines binären Klassifikationstrainingdatensatzes wieder.
3. Erläutern Sie Unterschiede und Gemeinsamkeiten der explanatorischen und prädiktiven Modellierung.
4. Erläutern Sie die Idee der Leave-One-Out-Crossvalidation (LOOCV).
5. Erläutern Sie die Konfusionsmatrix bei LOOCV mit binärem Label.
6. Geben Sie die Definitionen von Akkuratheit, Sensitivität und Spezifität bei LOOCV wieder.

Referenzen

Dwyer, Dominic B., Peter Falkai, and Nikolaos Koutsouleris. 2018. "Machine Learning Approaches for Clinical Psychology and Psychiatry." *Annual Review of Clinical Psychology* 14 (1): 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Sainani, Kristin L. 2014. "Explanatory Versus Predictive Modeling." *PM&R* 6 (9): 841–44. https://doi.org/10.1016/j.pmrj.2014.08.941.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3). https://doi.org/10.1214/10-STS330.