Masterarbeit

Daniel Siemmeister

2022 - 07 - 07

Titel der Arbeit

Erprobung unterschiedlicher Machine Learning Ansätze für die Vorhersage der Prüfungsaktivität von Studierenden

Wie viele prüfungsaktive Studierende gibt es in drei Jahren?

Wie viele prüfungsaktive Studierende gibt es in drei Jahren?

Ansätze des LQM

Wie viele prüfungsaktive Studierende gibt es in drei Jahren?

Ansätze des LQM

Prädiktion der Wahrscheinlichkeit, in drei Jahren prüfungsaktiv zu sein - ohne konkrete Klassifizierung

Machine Learning

$\mathcal{X}$ . . . Menge der Inputdaten

$\mathcal{Y}$ . . . Menge der Outputdaten

$\mathcal{D}_{\mathbf{X}}$ . . . Verteilung über $\mathcal{X}$

$f(\cdot)$ . . . mit $Y = f(\mathbf{X}) + \epsilon$

$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ mit n Datenpunkten

$\mathcal{D}$ . . . gemeinsame Verteilung von $(\mathbf{X}, Y)$

Machine Learning

$\mathcal{H} = \{h(\cdot, \mathbf{w}) | \mathbf{w} \in \mathbf{W}\}$

$\mathcal{A} \ldots$ Algorithmus mit $h_S = \mathcal{A}(S)$

$l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+ \ldots$ loss-Funktion, wobei $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

$L_D(\mathcal{A}) =$
$\mathbb{E}[l(\mathcal{A}(S), (\mathbf{X}, Y))] \ldots$ wahre Risikofunktion

$L_S(h_S) =$
$\frac{1}{n} \sum_{i=1}^{n} l(h_S, (\mathbf{x}_i, y_i)) \ldots$ empirische Risikofunktion

Machine Learning

mit ϵ wird Verteilung von $\mathcal{D}_{Y|\mathbf{X}}$ festgelgt

Parameter von $\mathcal{D}_{Y|\mathbf{X}}$ soll mittels h_S approximiert werden

Sinvolle Wahlen: Erwartungswert, Median

loss-Funktion entscheidet darüber, welcher Parameter approximiert wird

- $l(h_S, (\mathbf{X}, Y)) = (Y - h_S(\mathbf{X}))^2$ approximiert $\mathbb{E}[Y|\mathbf{X}]$
- $l(h_S, (\mathbf{X}, Y)) = |Y - h_S(\mathbf{X})|$ approximiert $m(Y|\mathbf{X})$ (Median)

Machine Learning

erwarteter Output: $\bar{y}(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$

erwartete Vorhersagefunktion:
$\bar{h} = \mathbb{E}[\mathcal{A}(S)] = \int_{\mathbb{R}^{(d+1)n}} h_s p_S(s)\, ds$

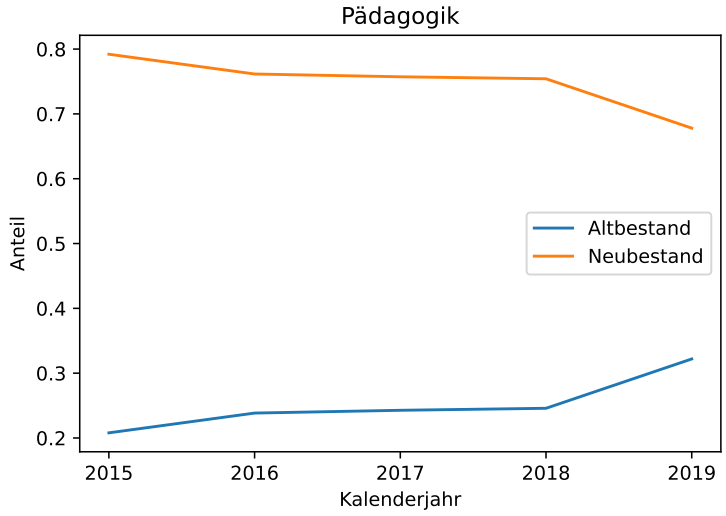
$\mathcal{L}_{\mathcal{D}}(\mathcal{A}) = \mathbb{E}[(h_S(\mathbf{X}) - Y)^2]$ wird zu

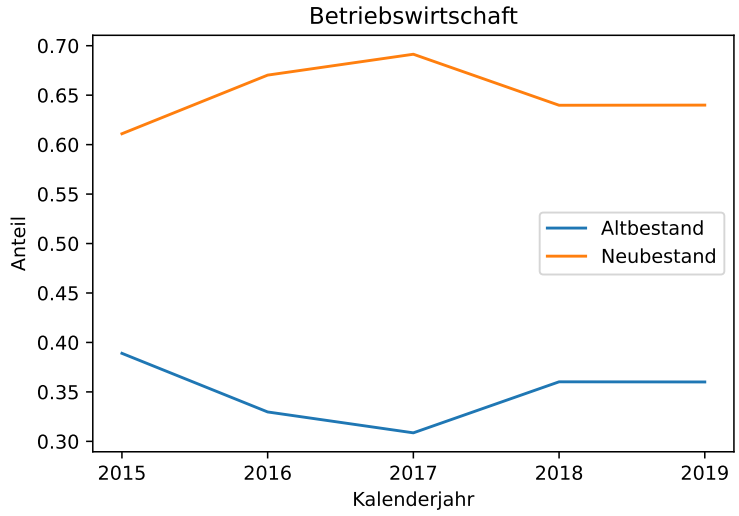
$$\underbrace{\mathbb{E}[(h_S(\mathbf{X}) - \bar{h}(\mathbf{X}))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}[(\bar{h}(\mathbf{X}) - \bar{y}(\mathbf{X}))^2]}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(\bar{y}(\mathbf{X}) - Y)^2]}_{\text{Noise}}$$

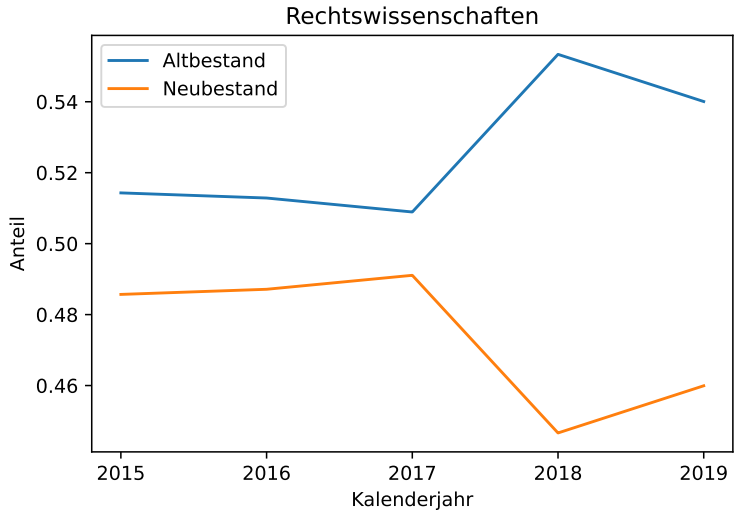
Problemstellung

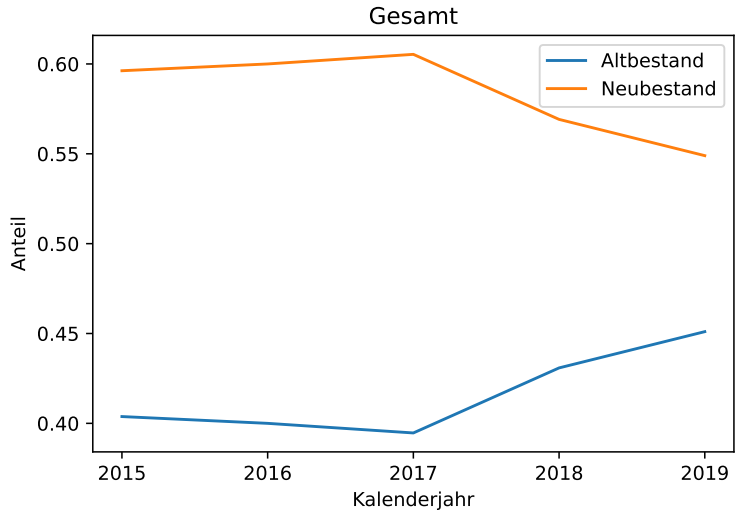
Daten von 2022 und davor	Daten von 2023, 2024, 2025
☐ Anzahl vorhanden	☐ Anzahl nicht vorhanden
☐ Merkmalskombinationen vorhanden	☐ Merkmalskombinationen nicht vorhanden

Relevanz der Problemstellung









Problem 1:
Schätzung der prüfungsaktiven Studierenden, von den bereits Anzahl und Merkmalskombinationen vorhanden sind

Problem 1:
Schätzung der prüfungsaktiven Studierenden, von den bereits Anzahl und Merkmalskombinationen vorhanden sind

Problem 2:
Schätzung der prüfungsaktiven Studierenden, von denen weder Anzahl noch Merkmalskombinationen vorhanden sind

Herangehensweise Problem 1

- Regression der ECTS

Herangehensweise Problem 1

- Regression der ECTS
- Markov Ketten Modell

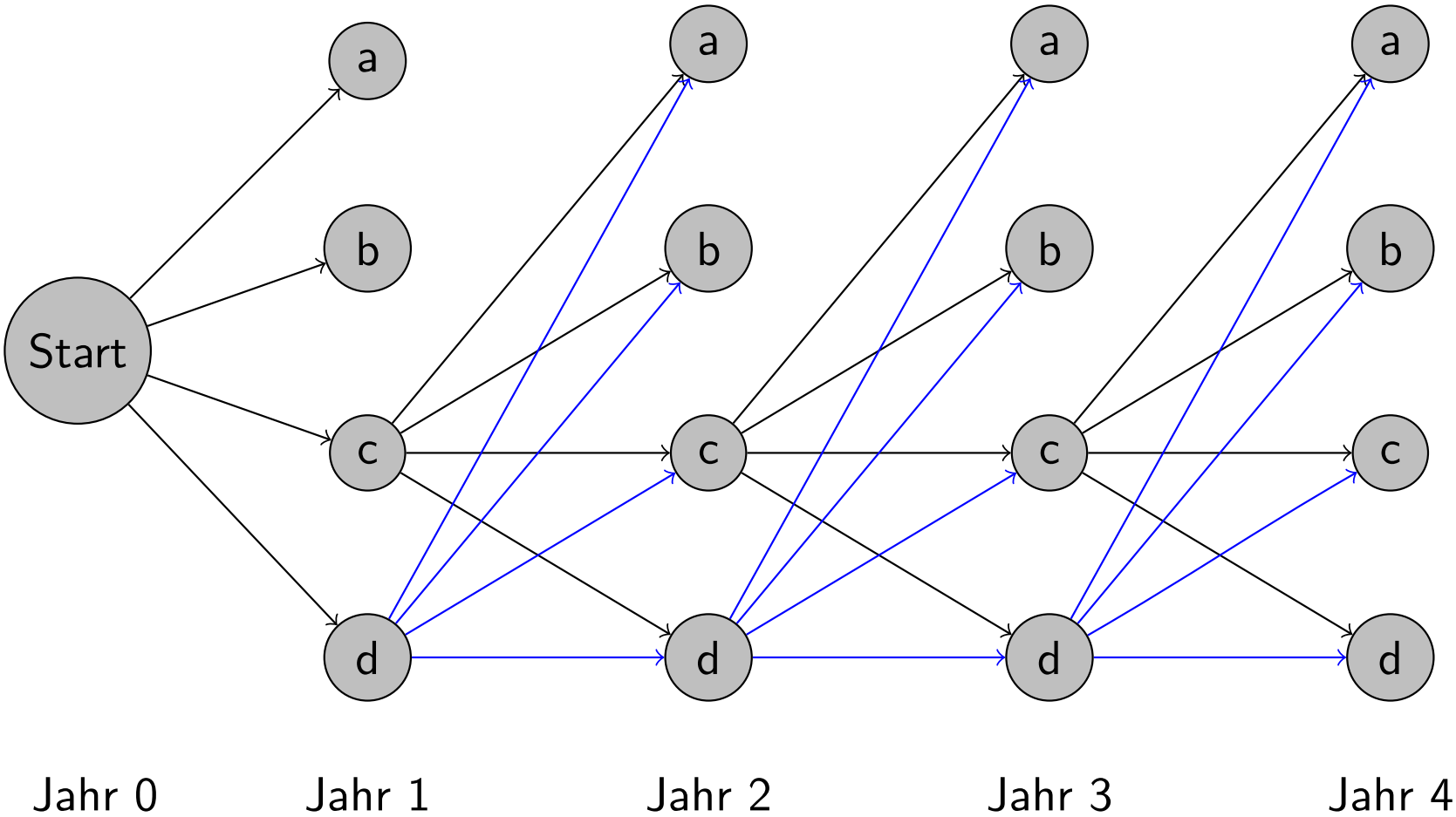
Herangehensweise Problem 1

- Regression der ECTS
- Markov Ketten Modell
- Schätzung der Wahrscheinlichkeit aktiv zu sein, ohne zu klassifizieren

Ergebnisse für Problem 1 Ansatz 1

Metrik		lineare Re-gres-sion	Random Forest	SVM	KNN (ohne CV)
RMSE (Crossvalidation)	1 Jahr	18.7 ± 0.2	19.2 ± 0.3	19.7 ± 0.4	18.72
	$\geq$ 2 Jahre	16.8 ± 0.2	15.4 ± 0.2	19.2 ± 0.3	14.8
MAE (Trainingsdaten)	1 Jahr	15.6	15.9	15.9	14.5
	$\geq$ 2 Jahre	13.3	11.7	16.2	10.4
R2-Score	1 Jahr	0.06	$-.01$	-0.05	0.06
	$\geq$ 2 Jahre	0.38	0.48	0.17	0.52
% Accurancy	1 Jahr	61	61	60	63
	$\geq$ 2 Jahre	80	78	66	80

Ergebnisse für Problem 1 Ansatz 2



Ergebnisse für Problem 1 Ansatz 2

$$\begin{bmatrix} 0.05 & 0.19 & 0.53 & 0.23 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.02 & 0.13 & 0.79 & 0.07 \\ 0.02 & 0.49 & 0.19 & 0.29 \end{bmatrix},$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.01 & 0.05 & 0.82 & 0.13 \\ 0.01 & 0.46 & 0.24 & 0.29 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.05 & 0.01 & 0.83 & 0.11 \\ 0.00 & 0.18 & 0.25 & 0.50 \end{bmatrix},$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.15 & 0.02 & 0.71 & 0.12 \\ 0.03 & 0.27 & 0.28 & 0.42 \end{bmatrix}.$$

Ergebnisse für Problem 1 Ansatz 3

		log. Reg.	RF	SVM	KNN
1 Jahr	Predicted	129.39	128.17	128.84	129.29
	Real	129	129	129	129
$\geq$ 2 Jahre	Predicted	121.25	117.46	120.59	120.9
	Real	121	121	121	121
1 Jahr	CV Scores	0.63 ± 0.00	0.62 ± 0.01	0.63 ± 0.02	0.62 ± 0.01
$\geq$ 2 Jahre	CV Scores	0.70 ± 0.01	0.73 ± 0.01	0.68 ± 0.01	0.72 ± 0.01

Ergebnisse für Problem 1

- X Ansatz 1 funktioniert nicht - zu große Fehler bei Schätzung der ECTS

Ergebnisse für Problem 1

- X Ansatz 1 funktioniert nicht - zu große Fehler bei Schätzung der ECTS
- X Ansatz 2 benötigt mehr Daten, um ihn seriös zu erproben

Ergebnisse für Problem 1

- ✗ Ansatz 1 funktioniert nicht - zu große Fehler bei Schätzung der ECTS
- ✗ Ansatz 2 benötigt mehr Daten, um ihn seriös zu erproben
- ✓ Ansatz 3 funktioniert auf kleinem Datensatz (sehr!) gut - man benötigt mehr Daten um ihn noch besser zu erproben

Herangehensweise Problem 2

- Schätzung der Anzahl der Studierenden mit gleicher Merkmalskombination wie im Jahr zuvor

Herangehensweise Problem 2

- Schätzung der Anzahl der Studierenden mit gleicher Merkmalskombination wie im Jahr zuvor
- Clustering der Studierenden und anschließende Schätzung der Anzahl nach Cluster

Ergebnisse für Problem 2

Zeitspanne der Schätzung		Prediction reale Daten	Prediction dummy Daten (Anzahl gegeben)	tatsächliche Anzahl
1 Jahr	2016	1118	1105	1092
	2017	1000	984	973
2 Jahre	2016	867	878	819
	2017	769	769	721

Ergebnisse für Problem 2

Zeitspanne der Schätzung		Prediction reale Daten	Prediction dummy Daten (Anzahl gegeben)	tatsächliche Anzahl
1 Jahr	2016	1118	1105	1092
	2017	1000	984	973
2 Jahre	2016	867	878	819
	2017	769	769	721

$\sim$ Legitimation von Ansatz 1 für Problem 2

Ergebnisse für Problem 2

Zeitspanne der Schätzung		Prediction reale Daten	Prediction dummy Daten (Anzahl gegeben)	tatsächliche Anzahl
1 Jahr	2016	1118	1105	1092
	2017	1000	984	973
2 Jahre	2016	867	878	819
	2017	769	769	721

- ∼ Legitimation von Ansatz 1 für Problem 2
- ∼ Zu wenige Daten vorhanden, um Ansatz 2 für Problem 2 seriös zu erproben

Beiträge

Klare Darstellung der Problemstellung

Beiträge

Klare Darstellung der Problemstellung

Erprobung unterschiedlicher Ansätze

Beiträge

Klare Darstellung der Problemstellung

Erprobung unterschiedlicher Ansätze

Machine Learning Ansatz für Problem 1, der gute Ergebnisse liefert

Beiträge

Klare Darstellung der Problemstellung

Erprobung unterschiedlicher Ansätze

Machine Learning Ansatz für Problem 1, der gute Ergebnisse liefert

Notwendigkeit von mehr Daten, um Ansätze weiter zu Erproben