1. Sitzung

18.06.2021

Lineare Regression

Regressions-Residuen

1. Verstehen, was der Fehler einer Linearen Regression ist

Analytische Lösungen in Linearer Regression

2. Einsicht, dass die Parameter einer Linearen Regression analytisch, d.h. ohne iterative Verfahren eineindeutig bestimmt werden können.
3. Matrizen-Schreibweise für die Multivariate Lineare Regression: Visuelles Wiedererkennen der Formel; Wissen, dass die analytische Lösung die Inversion der Variablen-Kreuzprodukt-Matrix benötigt; Diese Inversion war in der Vergangenheit mathematisch eine Herausforderung
4. Polynomiale Regression als Beispiel für Multivariate Lineare Regression; Einsicht, dass die Formel unverändert bleibt, die Regressions-Linie aber keine Gerade mehr ist.

Overfitting

5. Einsicht, wie overfitting entsteht; Kenntnis um den Zusammenhang von Anzahl Datenpunkte und Anzahl anzupassender Parameter

Lösungen für Overfitting

6. Regularization; L1 und L2 penalties für Parameter in der Fehlerfunktion (loss-function)
7. Für Ridge-Regression existiert eine analytische Lösung (dieses Verfahren benötigt auch keine schrittweise Annäherung).
8. Unterschied zwischen Ridge- und Lasso-Parameters
9. Elastic-Net; was ist es; welche Hyper-Parameter müssen optimiert werden
10. Regularization; L1 und L2

Interaktionen

11. Was sind Interaktionen von Variablen; Wie bildet man in python Interaktions-Terme -> durch Multiplikation der beiden Variablen
12. Problematik, die besten Interaktions-Terme für die lineare Regression zu finden

Konfidenz-Intervalle

13. Es gibt analytische Konfidenzintervalle, die nur gültig sind, wenn die Voraussetzungen der linearen Regression erfüllt sind

14. Es gibt numerische Verfahren (Bootstrapping) um sich auch für nicht-parametrische Verfahren oder bei Verletzung der Voraussetzungen, Konfindenzintervalle abzuleiten

Generalized Linear Model (GLM)

15. Unterschied Generalized Lineare Model and General Lineare Model kennen
16. Logistic-Regression ist auch eine Lineare Regression

Neuronale Netzwerke und Regression

17. Ähnlichkeit von Linearer Regression und einem Perceptron verstehen; Übereinstimmung in der Vektorschreibweise erkennen
18. weight decay als regularization in Neuronalen Netzwerken ist ein L2-penalty wie in Ridge-Regression

Data Leakage

19. wie entsteht Data Leakage
20. warum müssen wir es verhindern, wenn wir unseren Datensatz zusammenstellen
21. Beispiele für Data Leakage
22. Wie erkennen wir Data Leakage
23. Was sind Abhängige Daten; Oversampling erzeugt abhängige Datensätze
24. Richtige Strategie für Oversampling: Nach split in Train- und Test-Data
25. Python-Pipeline

Validierunsgs-Schemata

Kreuzvalidierung

26. Warum macht man das
27. Nested-Cross-Validation

Stacking

28. Warum macht man das

Mean- oder Target-Encoding

29. Wie funktioniert es?
30. Catboost als einen Algorithmus für kategorielle Variablen ist bekannt

House-Prices Example

Preprocessing

31. Wann dürfen die Test-Daten beim Preprocessing im gesamten Datensatz verbleiben
32. Preprocessing-Steps in Regression: Box-Cox-Transformation, Missing-Value Imputation, Dummy- aka One-Hot-Encoding, Interaktions-Terme

Hyper-Parameter-Search

33. Warum erzeugt man für manche Parameter die Werte für die Grid-Search auf einer logarithmischen Skala?
34. GridSearchCV sklearn-Klasse ist bekannt

Abschätzen der Güte des trainierten Algorithmus

35. Wie kann ich Abschätzen, wie gut mein trainierter Algorithmus bei unbekannten Daten sein wird?
36. Zusammenhang Modell-Güte und Overfitting