

13.06.24

Einführung in Data Science und maschinelles Lernen

BEHANDLUNG VON FEHLENDEN WERTEN

- **Quiz**
- **Nutzung von Dropout-Layern**
- **Behandlung fehlender Werte**
- **(Support-Vektor-Maschinen)**

QUIZ



NEURONALES NETZ MIT DROPOUT LAYER

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import InputLayer, Dense, BatchNormalization, Dropout
from tensorflow.keras.optimizers import Adam

model = Sequential([
    InputLayer(input_shape=(training_features.shape[1], )),
    BatchNormalization(),
    Dense(10, activation='relu'),
    Dropout(.3),
    Dense(4, activation='relu'),
    Dense(1)
])
```

Model: "sequential"

Layer (type)	Output Shape	Param #
batch_normalization (Batch Normalization)	(None, 34)	136
dense (Dense)	(None, 10)	350
dropout (Dropout)	(None, 10)	0
dense_1 (Dense)	(None, 4)	44
dense_2 (Dense)	(None, 1)	5

Total params: 535 (2.09 KB)

Trainable params: 467 (1.82 KB)

Non-trainable params: 68 (272.00 Byte)

DROPOUT LAYER ALS FORM DER REGULARISIERUNG

- **Setzt bei jedem Iterationsschritt die einzelnen Aktivierungen im vorherigen Layer mit der definierten Dropout-Wahrscheinlichkeit auf 0.**
- **Integriert Redundanz in das Netz.**
- **Hilft Overfitting zu vermeiden.**
- **Wird nur angewendet im Training, für die Inferenz werden immer alle Neuronen genutzt.**

FEHLENDE WERTE

GRÜNDE FÜR FEHLENDE WERTE

- **Nicht gegebene Antworten in Umfragen**
- **Zusammenführung von Daten aus verschiedenen Quellen mit unterschiedlichen Variablen**
- **Technische Probleme in der Datenerhebung oder Aufzeichnung**
- ...

TYPEN VON FEHLENDEN WERTEN

- **Missing Completely at Random (MCAR)**
- **Missing at Random (MAR)**
- **Missing not at Random (MNAR)**

BREAKOUT

Diskutiert Lösungen für folgende mögliche Fälle im Wetterdatensatz:

- **Temperaturdaten für einen Monat mit fehlenden Daten für zwei Tage:**
[20, 19, 23, 19, 17, 17, NA, 24, 16, 20, 22, 21, 20, 19, 17, 22, 24, 21, 23, 15, 18, 18, 21, 19, 19, 21, 21, 19, 23, NA]
- **Temperaturdaten für einen Monat mit fehlenden Daten für ein Woche:**
[18, 15, 21, 15, 24, 16, 21, 16, 22, 18, 17, 25, 22, 21, 16, 19, 17, 23, NA, NA, NA, NA, NA, NA, 21, 20, 20, 16, 15]
- **Wettercode-Daten für 20 Tage mit fehlenden Daten für einen Tag:**
[10, 60, NA, 95, 61, 1, 29, 81, 21, 25, 25, 80, 80, 63, 81, 80]


BEHANDLUNG VON FEHLENDEN WERTE

- **Zugehörige Fälle löschen (listwise deletion)**
- **Einfache „Spender“-basierte Imputation (donor-based)**
 - **Mittelwertschätzung (bzw. Median oder Mode)**
 - **nach „Ähnlichkeit“ (Hot-Deck Imputation)**
 - **durch minimalen Abstand (k Nearest Neighbors)**
- **Einfache modellbasierte Imputation**
 - **Iterative Regression**
- **Multiple Imputation**


HOT-DECK IMPUTATION

Nach Domänen

PhysActive	Weight
TRUE	51
TRUE	63
FALSE	98
TRUE	NA
FALSE	81
FALSE	88



PhysActive	Weight
TRUE	51
TRUE	63
FALSE	98
TRUE	NA
FALSE	81
FALSE	88



Nach Korrelation

Height	Weight
150	51
161	63
189	98
155	NA
182	81
184	88



Height	Weight
150	51
155	NA
161	63
182	81
184	88
189	98



K-NEAREST NEIGHBORS (KNN)

Suche nach den k Fällen mit minimalem Abstand

- je nach Variablentyp unterschiedliche Abstandsmessung
- Zusammenführung der Abstände über eine Summenfunktion

Verschiedene Vorgehensweisen zur Berechnung des Imputationswerts:

- Der Wert mit minimalem Abstand wird genommen (1NN)
- Zufällige Ziehung aus den k Fällen
- Berechnung aus den k-Fällen über den (gewichteten Mittelwert)

ITERATIVE REGRESSION

1) Vorhersage fehlender Werte in A

A	B	C	D
5	34	NA	1
1	22	NA	4
NA	65	55	2
4	87	27	2
NA	23	10	1

ITERATIVE REGRESSION

1) Vorhersage fehlender Werte in A

A	B	C	D
5	34	NA	1
1	22	NA	4
5	65	55	2
4	87	27	2
2	23	10	1

ITERATIVE REGRESSION

2) Vorhersage Fehlender Werte in C mit den imputierten werten von A

A	B	C	D
5	34	NA	1
1	22	NA	4
5	65	55	2
4	87	27	2
2	23	10	1

ITERATIVE REGRESSION

2) Vorhersage Fehlender Werte in C mit den imputierten werten von A

A	B	C	D
5	34	32	1
1	22	16	4
5	65	55	2
4	87	27	2
2	23	10	1

ITERATIVE REGRESSION

3) Vorhersage fehlender Werte in A mit den imputierten Werten von C

A	B	C	D
5	34	32	1
1	22	16	4
NA	65	55	2
4	87	27	2
NA	23	10	1

→ Wiederholung bis keine Änderung mehr eintritt

ITERATIVE REGRESSION

- 1) Gehe schrittweise durch alle Variablen des Datensatz**
 - 2) Stelle dabei für jede Variable ein Regressionsmodell basierend auf allen anderen Variablen auf**
 - 3) Berechne für alle fehlenden Werte eine Vorhersage**
- **Jetzt wiederhole Schritt 1) bis 3) erneut und schätze die fehlenden Werte erneut - dieses Mal mit den bereits imputierten fehlenden Werten.**
 - **Wiederhole dies, bis sich die imputierten Werte nicht mehr ändern.**

IMPUTATION EXAMPLES

missing_value_imputation.ipynb U x

Fehlende Werte > missing_value_imputation.ipynb > ...

+ Code + Markdown | ▶ Run All ↺ Restart ⌵ Clear All Outputs | 📄 Variables 📄 Outline ...

Python 3.10.12

```
# Import libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import missingno as msno
from fancyimpute import IterativeImputer, KNN
```

[2] ✓ 0.0s Python

```
# Load the dataset
url = 'https://raw.githubusercontent.com/opencampus-sh/einfuehrung-in-data-science-und-ml/main/Fehlende%20Werte/airquality.csv'
airquality = pd.read_csv(url)

airquality.head()
```

[3] ✓ 0.1s Python

	Ozone	Solar.R	Wind	Temp	Month	Day
0	41.0	190.0	7.4	67	5	1
1	36.0	118.0	8.0	72	5	2
2	12.0	149.0	12.6	74	5	3
3	18.0	313.0	11.5	62	5	4
4	NaN	NaN	14.3	56	5	5

...

Visualization of Missing Data

markdown

```
# Matrix plot of missing data
msno.matrix(airquality, figsize=(12, 6))
plt.title('Missing Data Matrix Plot')
plt.show()
```

[4] ✓ 0.4s Python

...

Missing Data Matrix Plot

1

Ozone Solar.R Wind Temp Month Day

BERECHNUNG DES IMPUTATIONSFEHLERS

- 1. Erstellen eines vollständigen Datensatz („Referenz-Datensatz“)**
- 2. Zufälliges Entfernen von Daten**
- 3. Imputieren der fehlenden Daten mit der gewählten Methode (ggf. auch mehreren zum Vergleich)**
- 4. Vergleich der imputierten Daten mit den Originaldaten:, z.B. durch Berechnung**
 - **des mittleren quadratischen Fehlers (MSE) oder**
 - **des absoluten Fehlers.**
- 5. Bewertung des Fehlers (ggf. Anpassung des Imputationsverfahrens)**

BREAKOUT

- **Nutzt die `missingno` library, um Ihre fehlenden Werte darzustellen.**
- **Führt eine erste Imputation durch.**

LERNMATERIAL

- Schaut [dieses Video](#) (5 Minuten) zu Zeitreihenanalysen.
- Als zusätzlichen Input zu Missing Values könnt Ihr das erste Kapitel [dieses Kurses](#) bei datacamp absolvieren.

AUFGABEN

- **Wählt ein (bzw. verschiedene) Verfahren, um die fehlenden Werte in Eurem Datensatz zu ersetzen.**
- **Teilt Euch die Aufgaben im Team gut auf:
Wer arbeitet an der Datenoptimierung, wer an der Modelloptimierung?**