

Advanced Machine Learning

Zusammenfassung

Stephan Stofer

27. Februar 2021

Inhaltsverzeichnis

1	Data Classification	4
1.1	Data Quality Assessment	4
1.1.1	Data Cleaning	4
1.1.2	Approaches to DQA	4
1.1.3	Statistische Kennzahlen	5
1.2	Replacement Strategies für NULL Values	6
1.2.1	Feature Engineering	6
1.2.2	Vector Space Model	6
1.3	Pandas Profiling	7
1.4	Fazit	7
2	this is a test	8

Abbildungsverzeichnis

1 Data Classification

Daten werden in zwei Klassen unterteilt. *Numerische* und *Kategorische* Daten. Bei numerische Daten gibt es *stetige* oder *diskrete* Zahlen. Bei Kategorischen sind entweder *ordinal* oder *nominal*. Ordinale haben eine Hierarchie.

1.1 Data Quality Assessment

Daten sind sehr wichtig, der beste ml-Algo nützt nicht, wenn Daten rubbish sind. Mögliche Fehlerquellen:

- Technische Fehler
- Qualität
- schlecht Design
- menschliche Fehler
- Input in Web-Apps (ungeprüfte Eingabefelder)
- Exporte der Daten, falsche Formate - oder Pre-Processing
- Falschangaben durch Benutzer
- Daten haben immer ein Ablaufdaten! (z.B. emailadressen, Adressen)

Ein DQA kommt immer zuerst! Schützt auch die Reputation gegenüber Kunden.

1.1.1 Data Cleaning

Prozess um Fehler in Daten zu beheben (automatisch)/bereinigen. Duplikate entfernen, null-values entfernen, Datenformate ml-friendly aufbereiten (data wrangling). Die Änderungen müssen dokumentiert und versioniert werden, den data provider darüber informieren und die Ursache für die data quality issues untersuchen.

1.1.2 Approaches to DQA

Dies ist detektiv-Arbeit. Wenn etwas verdächtig erscheint, weitergraben! Die Daten werden überprüft, ob sie vertrauenswürdig sind (plausibilieren).

- Datenquellen und vertrauenswürdigkeit prüfen
- statistische Kennzahlen interpretieren
- daten visualisieren
- Datenranges prüfen (Alter sollte unter 200 sein, Salär > 0, usw.)
- Korrelation zwischen Attributen prüfen (Tachostand und Preis eines Autos)
- Redundanz -> je weniger umso bessere Daten
- Anomalieprüfung in Syntax und Semantik
- NULL Werte und Duplikate erforschen

1.1.3 Statistische Kennzahlen

Geben uns einen Fingerabdruck und erste Plausibilisierung der Daten. Die wichtigsten Kennzahlen sind:

- Mittelwert - *mean* - $O(n)$
- Modus - *mode* die Zahl die am meisten vorkommt
- Median - *median* - $O(n * \log n)$, ist aussagekräftiger

1.1.3.1 Schiefheit

Der Mean, Modus und Median geben Auskunft über die Schiefheit der Daten. Wir haben eine negative, Links-Schiefe *skewness* wenn $mean - mode < 0$, wenn positiv, Rechts-Schiefe $mean - mode > 0$

1.1.3.2 Median

Sortiere Datenreihe. Der Median enthält 50% der Daten. Die Quantile entsprechen je 25%. Die Interquartils Differenz (IQR) entspricht $Q3 - Q1$.

1.1.3.3 Boxplots

Sehr nützlich zur grafischen Darstellung. *Outliers* sind die Werte die grösser sind als $Q3 + 1.5 * IQR$ respektive $Q1 - 1.5 * IQR$. Minimum bzw. Maximum sind die Werte, die gerade noch in diese Grenze $1.5 * IQR$ reinpassen.

Wenn viele Outliers müssen Daten genau angeschaut werden, ob sie trotzdem plausibel sind.

1.1.3.4 Five Number Summary of a Data Distribution

In mit Python kann sehr einfach die $Q1$, $Q2$, $Q3$, min und max einer Datenreihe ausgegeben werden:

```
import numpy as np
import pandas as pd

s = pd.Series(np.random.rand(100))
s.describe()
```

Auch Boxplots sind sehr einfach:

```
import matplotlib.pyplot as plt
plt.boxplot(x = [data.Mileage, data.Price], labels=['Mileage', 'Price'])
```

1.1.3.5 Datenverteilung

Die Verteilung wird mit der Varianz betrachtet, wobei diese *sample variance* die Besselkorrektur $(n - 1)$ nutzt. Die Standardabweichung entspricht aus der $\sqrt{Var(x)}$

1.1.3.6 Covarianz

Die Kovarianz zeigt die Variabilität von zwei Datensätzen auf. Ist der Wert positiv, verhalten sich die beiden Daten ähnlich. Ist sie negativ, entsprechend nicht. Ist aber schwierig zu interpretieren, weil sie nicht normiert ist.

1.1.3.7 Covarianzmatrix

Die covarianzmatrix ist sehr wichtig in ML. Sie enthält alle Covarianzen aller Varianzpaare. Die Diagonale kann durch die Varianz von \mathbf{X} ersetzt werden.

1.1.3.8 Pearson Korrelation

Covarianz wird durch die Standardabweichung dividiert. Deshalb ergeben sich Werte zwischen 1 (perfekte Korrelation) und -1 (perfect anti-correlation). Damit kann die Datenreihe verglichen werden. Die Korrelationsmatrix kann als Heatmap gut dargestellt werden.

1.2 Replacement Strategies für NULL Values

Kommen immer wieder vor. ML-Algos können selten damit umgehen und müssen bereinigt werden. Je nach Datenumfang sind versch. Verfahren denkbar:

- Zeilen mit NULL Werten löschen
- Fehlende Daten manuell einsetzen
- Globale Konstanten einsetzen (UNKNOWN, *inf*)
- Tendenzen verwenden (Mittelwert für symmetrische Daten, Medien für Schiefedaten)
- Tendenzen auch pro "Klasse" (Eigenschaften) berechnen (z.B Krebskranke und gesunde Patienten)
- Regressionsmodell (sehr aufwändig und ungewohnt in Praxis)

1.2.1 Feature Engineering

Features entsprechen Spalten. Null-Values können also mit ML erzeugen. Information verfügbar für ML-Algo machen.

1.2.2 Vector Space Model

Entspricht einem Datenset welches ausser dem Key nur numerische Werte enthält. Kategorische Daten können sehr einfach in nummersiche Daten transformiert werden. Zum Beispiel werden die Farben alle zu Spalten und entsprechende Zugehörigkeit mit 1 bzw. 0 gekennzeichnet. Diese werden als Dummy-Variable bezeichnet.

```
import pandas as pd
data = pd.read_csv('cars.csv')
data = pd.get_dummies(data)
```

Python code um Daten entsprechend aufzubereiten.

1.2.2.1 Dummy Variable Trap

Mit dem einfügen von Dummy-Variablen muss die *Multikolloniarität* im Auge behalten werden. Wenn n -Dummy Variablen erzeugt werden und $n - 1$ Spalten alle 0 sind, wissen wir zu 100, dass die n te Spalte 1 sein muss. Dies führt zu unterterminierten Matrizen. Die Matrix kann nicht invertiert werden. Um das zu verhindern, muss eine Spalte gelöscht werden! Es gibt aber Verfahren, die immun dagegen sind (z.B. Entscheidungsbäume).

1.3 Pandas Profiling

Effizient in drei Zeilen Code!

```
import pandas_profiling
data = pd.read_csv('cars.csv')
data.profile_report()
```

1.4 Fazit

Bei jedem ML-Projekt ist in Data Quality Assessment Pflicht

2 this is a test

I wrote this text.

Alice -> Bob: Authentication Request

Bob --> Alice: Authentication Response

Alice -> Bob: Another authentication Request

Alice <-- Bob: another authentication Response