Data Science Zusammenfassung

By Yannis Schmutz

todo

Zusammenfassung

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2 Statistik

2.1 Begriffe

Dieses Kapitel bietet eine grobe Übersicht über einige hilfreiche Begriffe der Statistik.

2.1.1 Lageparameter

Lageparameter beschreiben die Lage der Stichprobenelemente im Bezug auf die Messskala.

Mittelwert
Auch Durchschnitt (oder mean im Englischen) gennant. In der Wahrscheinlichkeitsrechnung spricht man oft vom Erwartungswert.

$$\bar{x}_{arithm} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Median
Der Median oder auch Zentralwert genannt, beschreibt den Wert aus der auf-/ absteigend geordneten Stichprobe, der genau in der Mitte liegt.

$$\widetilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{1}+1}\right), & n \text{ gerade.} \end{cases}$$

Quantile
Schwellenwert (engl. percentile) der angibt, dass ein bestimmter prozentualer Wert einer Menge an Werten kleiner ist als das Quantil. Das Quantil bei 50% ist der Median. Weitere spezielle Quantile sind die Quartile, die Quintile, die Dezile und die Perzentile.

Modus
Definiert den häufigsten Wert, der in der Stichprobe vorkommt.

2.1.2 Streuungsparameter

Streuungsparameter beschreiben die Streuung von Werten einer Stichprobe um einen bestimmten Lageparameter. So ergeben sich je nach gewählten Lageparameter unterschiedliche Berechungsformen. Diese unterscheiden sich in ihrer Beeinflussung durch Ausreisser. So wird beispielsweise der Median tendenziell weniger von einem einzelnen, sehr hohen Ausreisser beeinflusst als der arithmetische Mittelwert.

Spannweite
Die Spannweite (eng. range) gibt den Abstand des grössten gegenüber dem kleinsten vorkommenden Wert der Stichprobe an. $R = x_{max} - x_{min}$.
Die Spannweite wird stark durch Ausreisser beeinflusst. Dem kann jedoch durch das alternative Verwenden des **Interquartilsabstands** (engl. interquartile range) entgegengewirkt werden. Dieser berechnet nämlich die Spannweite zwischen zwei Quantilen. Somit können Ausreisser ignoriert werden.

Varianz
Korrelation
Kovarianz
Kausalität

2.1.3 Daten

NOIR

- **Nominal**
  - Meist diskret
  - Keine Rangordnung
  - Bspw: Farben, Geschlecht, Ortschaft etc.
- **Ordinal**
  - Meist diskret
  - Rangordnung
  - Keine interpretierbare Abstände
  - Bspw: Schlecht, okay, gut, sehr gut
- **Interval**
  - Meist stetig
  - Rangordnung
  - Kein interpretierbarer Nullpunkt
  - Bspw: Grad Celsius (geht unter Null)
- **Rational**
  - Meist stetig
  - Rangordnung
  - Definierter Nullpunkt
  - Bspw: Grad Kelvin (geht nicht unter 0K)

3 Probabilistik

3.1 Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit ist die Wahrscheinlichkeit des Eintreten eines Ereignisses A unter der Bedingung, dass die Wahrscheinlichkeit für das Eintreten eines Ereignisses B bereits bekannt ist. Man spricht von "A unter der Bedingung B". Oder auch $P(A|B)$.

Sind zwei Ereignisse E, F voneinander **unabhängig**, so gilt:

$$P(E \cap F) = P(E)P(F)$$

$$P(E|F) = P(E)$$

Sind jedoch zwei Ereignisse A, B **nicht unabhängig** so lautet die Formel für A unter der Bedingung B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Daraus erschliesst sich:

$$P(A \cap B) = P(A|B)P(B)$$

Das Aufzeichnen eines Wahrscheinlichkeitsbaumes hilft zur Veranschaulichung:

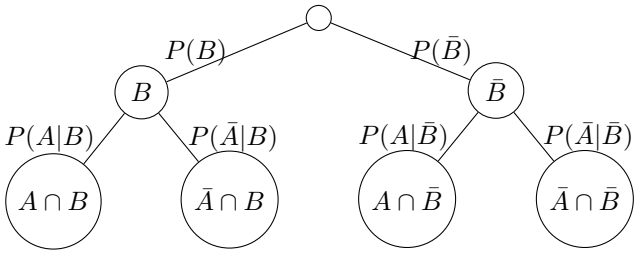


Abbildung 1: Wahrscheinlichkeitsbaum

3.1.1 Satz von Bayes

Der Satz von Bayes zeigt den Zusammenhang zwischen $P(A|B)$ und $P(B|A)$ auf:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Diese Gleichung 1 entsteht, wenn man den Ausdruck $P(A \cap B)$ anhand den umgekehrten Wahrscheinlichkeitsbaums 3.1.1 ausdrückt.

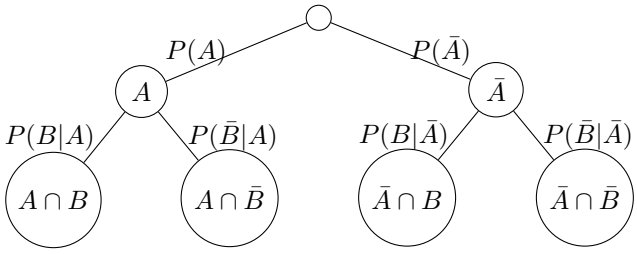


Abbildung 2: Umgekehrter Wahrscheinlichkeitsbaum

4 Machine Learning

4.1 Übersicht

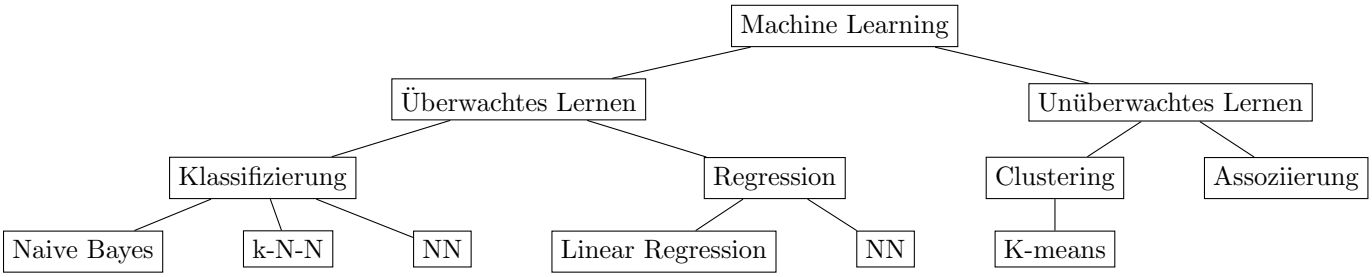


Abbildung 3: Machine Learning Kategorien

4.1.1 Überwachtes Lernen

Überwachte Lern-Algorithmen probieren Beziehungen und Abhängigkeiten zwischen den Input-Features und des zu erzielenden Outputs zu erschliessen. Dies unter der Verwendung von **beschrifteten Daten**. Diese können zu Trainingszwecken verwendet werden. Jeder Satz an Daten besteht aus Input-Werten sowie einem dazugehörigen bekannten Output-Wert. Nach dem Trainieren des Algorithmus versucht dieser anhand von **neuen** Input-Features den dazugehörigen **unbekannten** Output vorherzusagen.

Das überwachte Lernen kann in zwei Kategorien unterteilt werden:

- **Klassifizierung:** Ziel der Klassifizierung ist es, diskrete Werte vorherzusagen (bspw. Wahr/ Falsch, Spam-Mail/ normales Mail).
- **Regression:** Das Ziel der Regression ist die Vorhersage kontinuierlicher Werte (bspw. Hauspreise in Abhängigkeit von Fläche und Anzahl Zimmer).

4.1.2 Unüberwachtes Lernen

Im unüberwachten Lernen stehen den Algorithmen **keine** beschrifteten Daten zur Verfügung. Die Algorithmen versuchen eigenständig Pattern in den zu behandelnden Daten zu erkennen und sie dadurch beispielsweise gruppieren zu können.

4.2 Algorithmen

4.2.1 Naive Bayes

5 Tryout

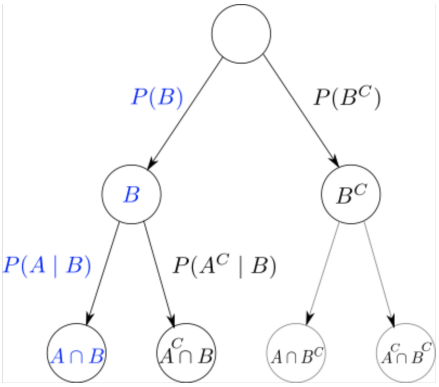


Abbildung 4: Entscheidungsbaum

Wie auf der Abbildung 4 zu sehen ist.....

Tabelle 1: Local caption, with reference [1–3]

Area	Number of rooms	Price
80	4	1680
100	5	2300
50	2.5	1500

- This is an item
- This is another item
  - This is a further item

blub This is an item with a custom bullet point

1. This is a numbered item
2. And so on

5.1 Math examples

Here's an example within a sentence $E = mc^2$.

And here one example

$$a = v/t$$

which is centred.

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi}{dx^2} = E\Psi$$

Fractions

$$d = v_i t + \frac{1}{2} \cdot at^2$$

$$d = v_i t + {}^1\!/_2 \cdot at^2$$

Brackets:

$$\left(\frac{1}{2}\right) \cdot 2 = 1$$

$$|-7| = 7$$

$$\sqrt{4} = 2$$

$$\sqrt{4} \neq 1$$

$$\sqrt{4} < 5$$

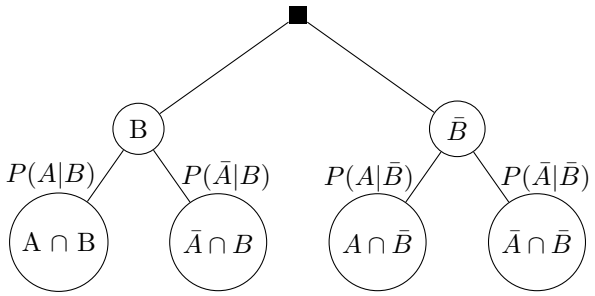
$$\pi \approx 3$$

$$\pi \times \sqrt{4} < 15$$

$$\begin{aligned} 3x + 14 &= 20 && (2)\\ 3x &= 6 && (3)\\ x &= 2 && (4) \end{aligned}$$

$$x^2 + 3x - 7 = 0 \quad (5)$$

5.2 Graphs


B
$\bar{B}$
$P(A|B)$
$P(\bar{A}|B)$
$P(A|\bar{B})$
$P(\bar{A}|\bar{B})$
A ∩ B
$\bar{A} \cap B$
$A \cap \bar{B}$
$\bar{A} \cap \bar{B}$

Literatur

[1] J. Grus, *Einführung in Data Science*. O'REILLY, 2016.

[2] T. Rashid, *Neuronale Netze selbst programmieren*. O'REILLY, 2017.

[3] U. Lämmel and J. Cleve, *Künstliche Intelligenz*. Carl Hanser Verlag München, 2012.

A Cheatsheet