KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Prognosemodelle (Klassifikation und Regression):
02_03Verfahren_Regression_02

Die logistische Regression

Erarbeitet von
Dr. Katja Theune

Lernziele

- Du kannst anhand eines einfachen Beispiels den prognostizierten Output einer neuen Beobachtung bestimmen
- Du kannst erklären, warum hier die Verwendung der logistischen anstatt der linearen Regression sinnvoll ist

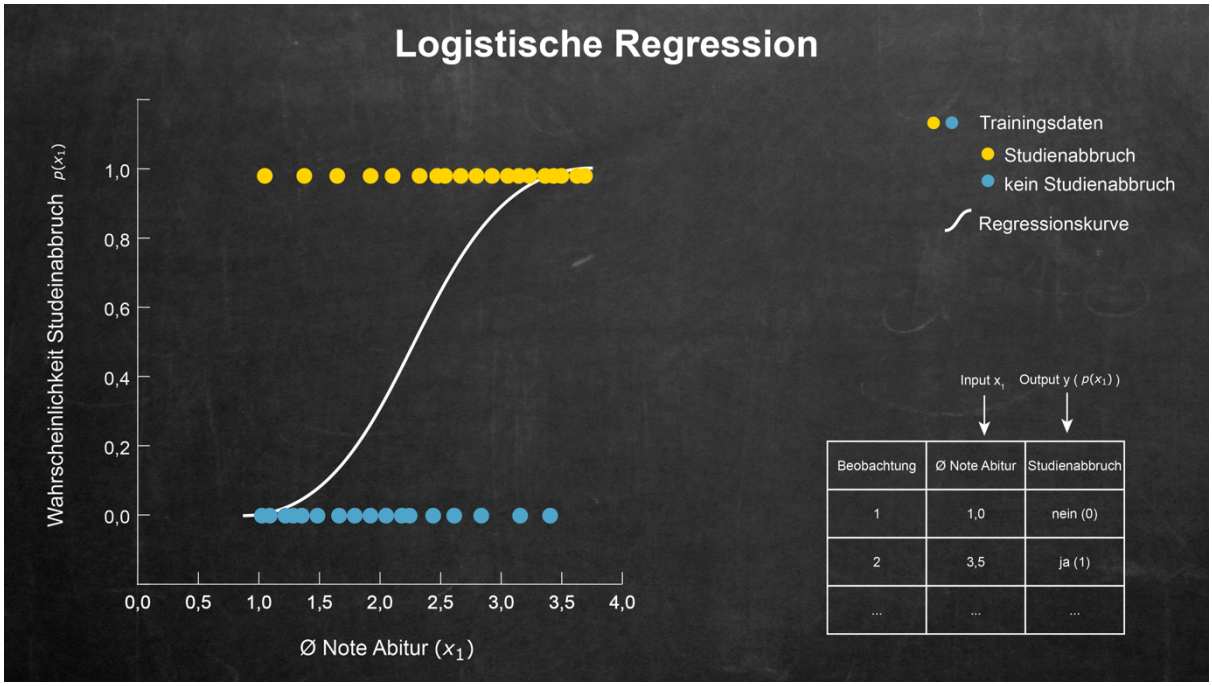
Inhalt

Einstieg

Befassen wir uns nun mit einer weiteren Art der Regression, der sogenannten logistischen Regression. Sie eignet sich, im Gegensatz zur linearen Regression, für kategoriale Outputs und damit zum Klassifizieren.

Logistische Regression: Idee

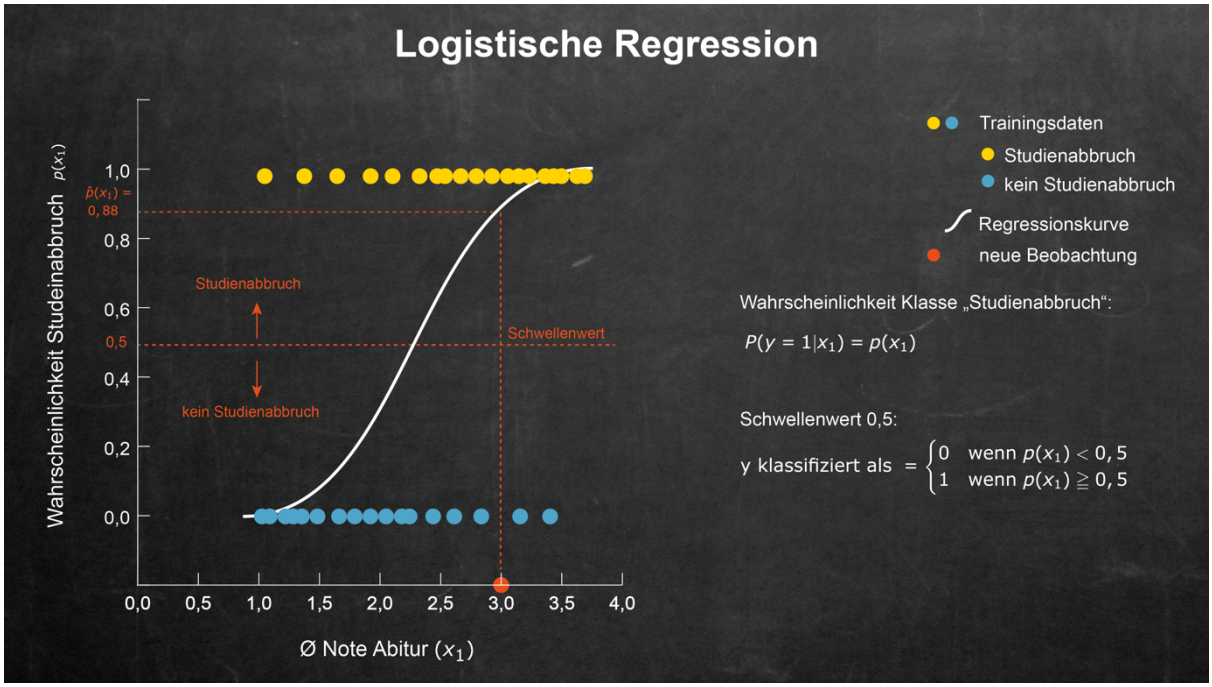
Wir beschäftigen uns hier ausschließlich mit der binomialen logistischen Regression, welche nur für Probleme mit zwei Klassen geeignet ist.
Kommen wir nochmal auf unser ausgedachtes Anwendungsbeispiel zurück und wollen diesmal vorhersagen, ob ein Studienabbruch stattfinden wird oder nicht. Wir haben also ein Klassifikationsproblem mit zwei Klassen. Zudem betrachten wir als einzigen Input die Abiturnote.



Die logistische Regression stellt jetzt aber nicht einen direkten Zusammenhang zwischen Inputs und Output bzw. den Klassen her, sondern einen Zusammenhang zwischen Inputs und der Wahrscheinlichkeit, dass eine der beiden Klassen zutrifft. Dieser wird anstatt mit einer Geraden durch eine S-Kurve dargestellt. Wahrscheinlichkeiten kann man wiederum als metrischen Output interpretieren. Daher verhält sich die logistische Regression, obwohl sie zu den Klassifikationsverfahren gehört, wie eine Regression.

Logistische Regression: Beispiel und Prognose

In unserem Beispiel betrachten wir z. B. die Wahrscheinlichkeit zur Klasse „Studienabbruch" zu gehören. Wir kodieren diese Klasse dann mit 1 und die andere Klasse, hier „kein Studienabbruch", mit 0. Die Wahrscheinlichkeit, zur Klasse „Studienabbruch" zu gehören, ist dann (siehe Grafik):

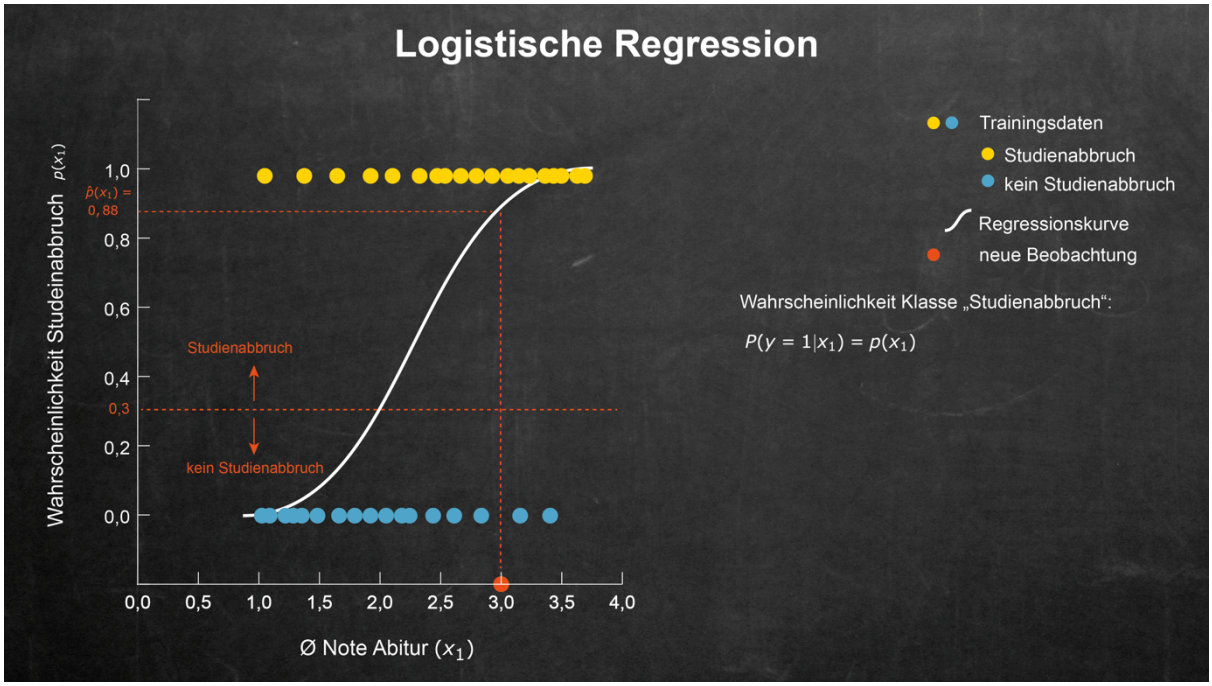


Das P ist eine Abkürzung für die Wahrscheinlichkeit und „unter der Bedingung", dargestellt durch den vertikalen Strich, bedeutet, dass wir die Wahrscheinlichkeit unter der Bedingung prognostizieren, dass ein bestimmter Wert für x_1, hier die Abiturnote, eingetroffen ist. Wir kürzen diese Wahrscheinlichkeit mit $p(x_1)$ ab und sie soll zwischen 0 und 1 liegen.

Wir können diese Wahrscheinlichkeiten jetzt verwenden, um z. B. neue Beobachtungen konkreten Klassen zuzuordnen. Üblich ist es, dass wir ab einer Wahrscheinlichkeit bzw. einem Schwellenwert von 0,5 sagen, dass eine Beobachtung der Klasse 1 angehört. Wir sehen das auch in der Grafik und an dieser Fallunterscheidung.

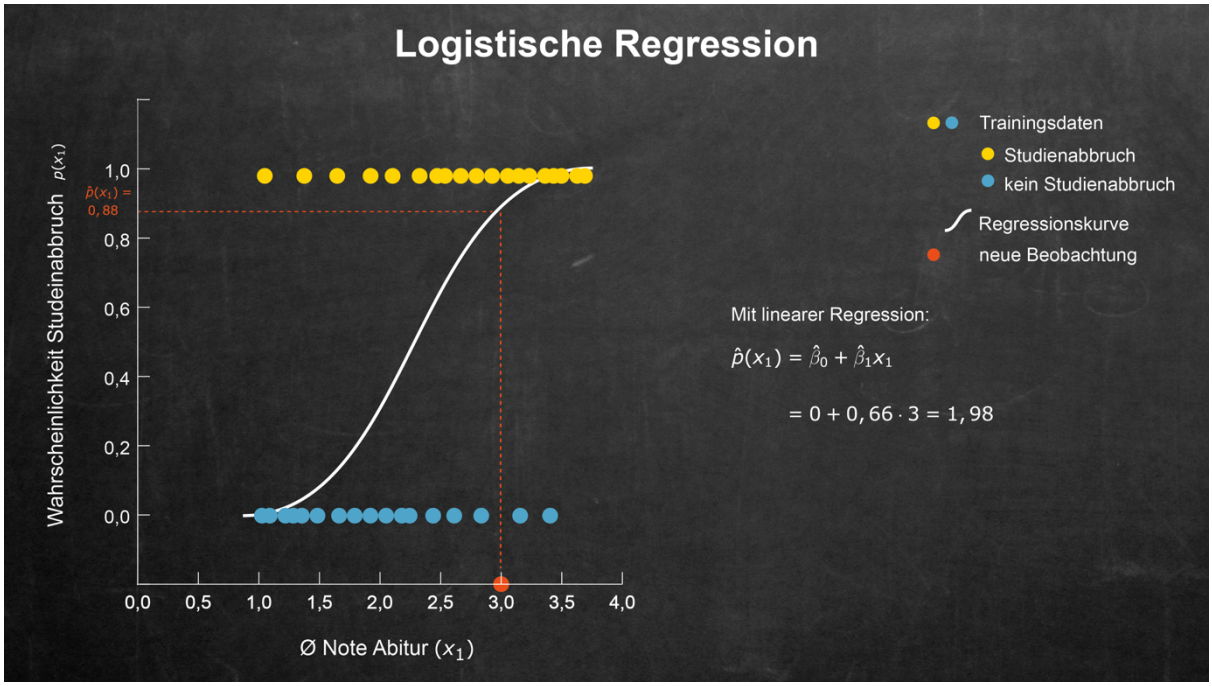
Zum Beispiel würden wir für eine neue Beobachtung mit einer Abiturnote von 3,0 eine Wahrscheinlichkeit $\hat{p}$ von 0,88 prognostizieren und sie damit der Klasse Studienabbruch zuordnen, da 0,88 größer als 0,5 ist.

Diesen Schwellenwert kann man aber individuell verändern, was bei manchen Anwendungsfällen sinnvoll sein kann. Eine alternative Zuteilung zu der Klasse „Studienabbruch" wäre z. B. ein Schwellenwert von 0,3 (siehe Grafik). Hier würde man dann vielleicht lieber in Kauf nehmen, dass fälschlicherweise erfolgreiche Studierende in diese Gruppe gelangen anstatt abbruchgefährdete Studierende zu übersehen.



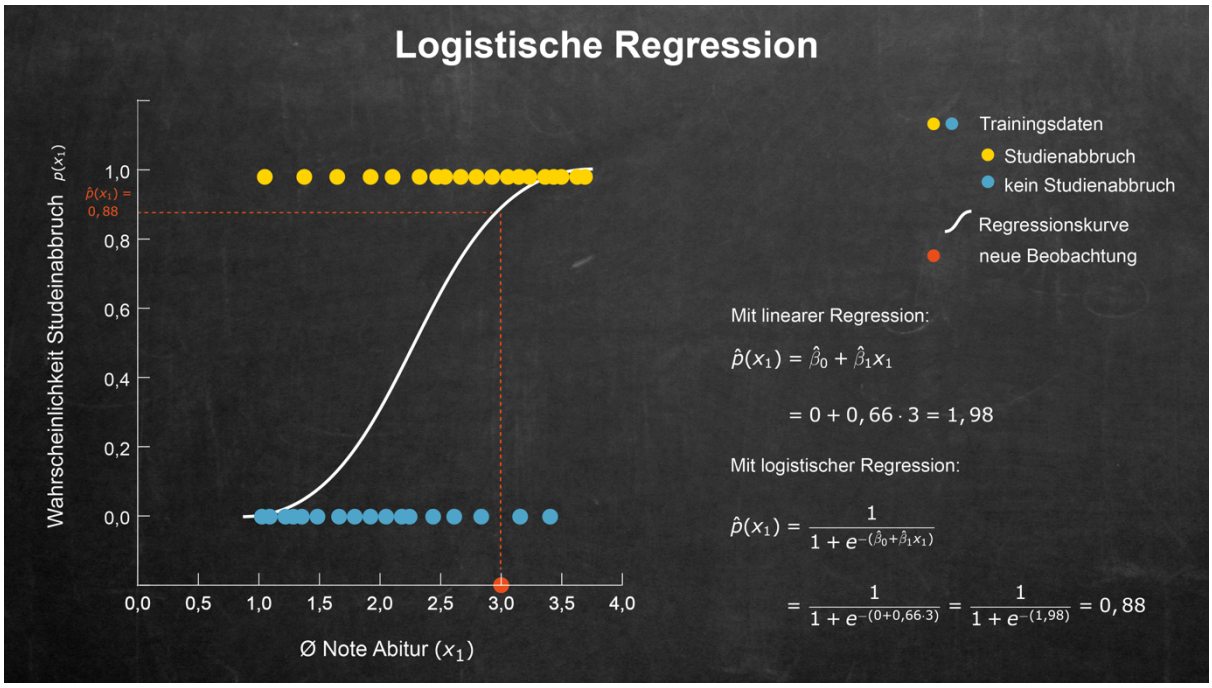
Logistische vs. lineare Regression: Beispiel

Würden wir nun wieder ein lineares Regressions-Modell verwenden, um die Wahrscheinlichkeiten $\hat{p}$ zu prognostizieren, dann würden wir für manche Werte von x_1 eine Wahrscheinlichkeit kleiner als 0 oder größer als 1 erhalten. Das ist natürlich nicht sinnvoll.



Um das zu veranschaulichen, sagen wir mal, wir hätten für $\hat{\beta}_0$ nun 0 und für $\hat{\beta}_1$ einen Wert von 0,66 geschätzt. Mit einer linearen Regression kämen wir nun für eine Beobachtung mit einer Abiturnote von $x_1 = 3$ zu einer geschätzten Wahrscheinlichkeit $\hat{p}$ von 1,98, was größer als 1 und nicht erwünscht ist (siehe Grafik).

Es gibt aber Möglichkeiten, Werte zwischen 0 und 1 zu gewährleisten. Eine davon ist die logistische Funktion. Sie ist folgendermaßen definiert (siehe Grafik):



Ich zeige hier die Formel direkt mit unserem Beispiel, um zu veranschaulichen, wie lineare und logistische Regression zusammenhängen. Nur der Vollständigkeit halber: e ist hier die Eulersche Zahl. Vielmehr müssen wir dazu aber erstmal nicht wissen. Im hochgestellten Teil, dem Exponenten von e, erkennen wir die lineare Regressionsgleichung, welche ja für unsere betrachtete Beobachtung einen Wert von 1,98 ergab.

Mit einer logistischen Regression erhalten wir jetzt für diese Beobachtung eine prognostizierte Wahrscheinlichkeit $\hat{p}$ von 0,88 (siehe Grafik).

Die Werte einer linearen Regression werden also durch die logistische Funktion in das Intervall 0 bis 1 überführt. Durch diese Überführung erhalten wir bei der logistischen Regression aber eine S-Kurve anstatt einer Geraden.

Wie man genau zu den Schätzungen für die Parameter $\hat{\beta}$ kommt, lassen wir hier außen vor. Die hier gewählten Werte dienen nur der Veranschaulichung und ergeben ansonsten nicht die dargestellte Kurve. Und ganz ähnlich zur multiplen linearen Regression, kann man natürlich auch hier bei mehreren Inputs eine multiple logistische Regression verwenden.

Diskussion, Vor- und Nachteile

Positiv ist hier ähnlich zur linearen Regression, dass wir einen Eindruck von Stärke und Richtung der Zusammenhänge zwischen Inputs und Outputs erhalten. Allerdings lassen sich diese nicht mehr ganz so intuitiv interpretieren.

Ein Nachteil ist auch hier, dass wir im Vorhinein Annahmen über die Art der Zusammenhänge zwischen Inputs und Output treffen müssen und oft eine aufwendigere Datenaufbereitung notwendig ist.

Abschluss

Wir haben jetzt ein häufig verwendetes Verfahren zur Klassifikation, die logistische Regression, kennengelernt. Mit diesem Verfahren können wir Wahrscheinlichkeiten für die Zugehörigkeit zu den Klassen prognostizieren. Auf Basis dieser Wahrscheinlichkeiten können wir dann Beobachtungen einer konkreten Klasse zuordnen.

Weiterführendes Material

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3. Auflage). Morgan Kaufmann.

James, G., Witten, D., Hastie, T., & Tibshirani, R., & Tylor, J. (2023). *An Introduction to Statistical Learning - with Applications in Python*. Springer.

Lantz, B. (2015). *Machine learning with R* (2. Auflage). Packt Publishing Ltd, Birmingham.

Disclaimer