Woche 06 Theorie: Regression

Skript

Erarbeitet von
Katja Theune

Lernziele 1
Inhalt 2
Einstieg 2
Lineare Regression – Beispiel 2
Lineare Regression – Idee 3
Lineare Regression – Finden der besten Gerade 4
Logistische Regression – Beispiel 5
Logistische Regression – Idee 6
Abschluss 8
Weiterführendes Material 9
Disclaimer 10

Lernziele

- Definieren, was eine Regression ist
- Erläutern der Idee und Vorgehensweise der linearen und logistischen Regression
- Anwenden der Vorgehensweise der Verfahren auf ein neues Beispiel
- Beispiele nennen, wozu man eine lineare und logistische Regression verwendet

Projekt „KI für alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Habt ihr euch auch schon mal gefragt, ob und wie sich euer Einkommen verändern würde, wenn ihr anstatt eines Bachelorstudiums auch noch ein Masterstudium oder sogar eine Promotion absolvieren würdet? Solche Überlegungen begegnen uns im Alltag sehr häufig und hier kommt z. B. die lineare oder logistische Regression zum Einsatz. Beides sind beliebte Verfahren aus dem supervised learning. Wir können mit ihnen Zusammenhänge zwischen einer Zielgröße, wie z. B. dem Einkommen und verschiedenen features, wie z. B. die Art des Studiums, bestimmen und damit Vorhersagen für die Zielgröße treffen. Wichtig ist hier, dass die Regression üblicherweise bei metrischen Zielgrößen zum Einsatz kommt.

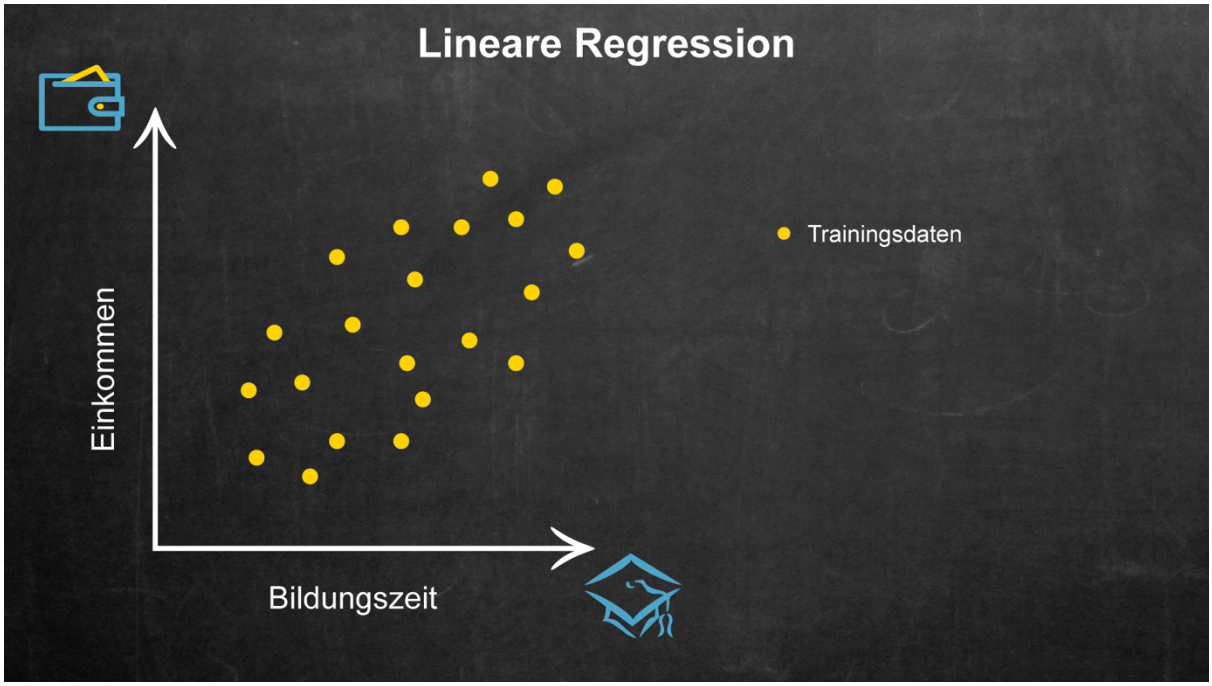
Lineare Regression – Beispiel

Wenn wir zwischen den features und der Zielgröße einen linearen Zusammenhang vermuten, dann verwenden wir eine sogenannte lineare Regression. Linearer Zusammenhang bedeutet, dass wir bei einer gleichbleibenden Veränderung eines features von einer gleichbleibenden Veränderung der Zielgröße ausgehen. Wir werden sehen, dass man dies mit einer geraden Linie optisch abbilden kann.

Einblendung kleines Koordinatensystem mit Gerade

Betrachten wir zum besseren Verständnis mal ein konkretes Beispiel. Wir könnten uns fragen, wie wir das Einkommen einer Person – das wäre dann unsere Zielgröße – vorhersagen können. Wir vermuten, dass das Einkommen z. B. von der investierten Zeit in Bildung – also der Zeit in der Schule, Ausbildung, Universität usw. – abhängt. Zudem vermuten wir, dass hier ein linearer Zusammenhang besteht. D. h. bei einer Veränderung der investierten Zeit in Bildung, sagen wir um ein Jahr, verändert sich auch das Einkommen immer um einen gleichbleibenden Betrag.

Wollen wir nun das Einkommen vorhersagen, können wir aus dem Bereich des Maschinellen Lernens eine sogenannte einfache lineare Regression verwenden. Einfach, da wir nur ein feature betrachten. Wir könnten uns aber auch überlegen, dass das Einkommen von mehreren features abhängt, z. B. der Bildungszeit, dem Alter und der Berufsgruppe einer Person. Das wäre dann eine sogenannte multiple lineare Regression. Jetzt gehen wir aber zur besseren Veranschaulichung der Vorgehensweise von einer einfachen Regression aus und betrachten nur das feature „Bildungszeit“. Unsere Trainingsdaten beinhalten nun für verschiedene Personen sowohl die jeweilige absolvierte Bildungszeit als auch die Höhe ihrer Einkommen. Wir tragen jetzt diese Daten in einem Koordinatensystem als Datenpunkte ein. Da wir nur eine Zielvariable und ein feature verwenden, haben wir hier ein zweidimensionales Koordinatensystem.

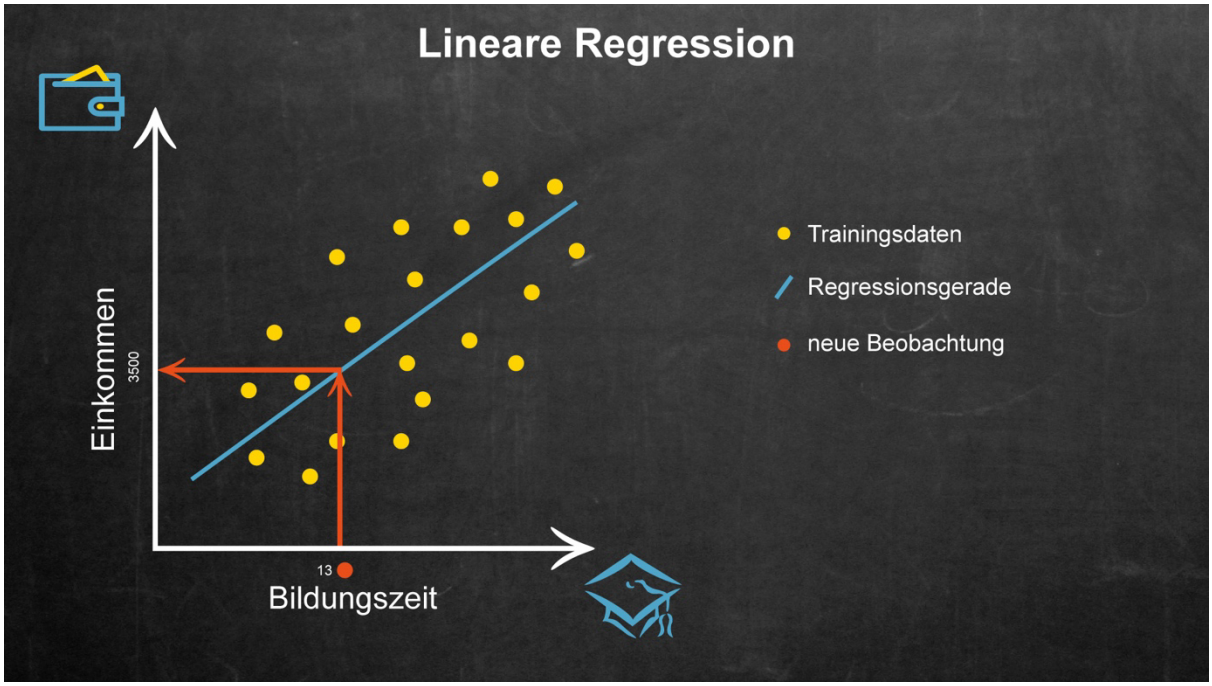


Wir tragen die Bildungszeit in Jahren auf der horizontalen Achse und das monatliche Brutto-Einkommen in Euro auf der vertikalen Achse ab. Die gelben Datenpunkte repräsentieren unsere Beobachtungen, also einzelne Personen und ihre Eigenschaften. Hier sind das bestimmte Werte von Bildungszeit und Einkommen. Je weiter rechts der Punkt liegt, desto mehr investierte Zeit in Bildung kann die Beobachtung vorweisen. Je weiter oben der Punkt liegt, desto höher ist ihr Einkommen.

Lineare Regression – Idee

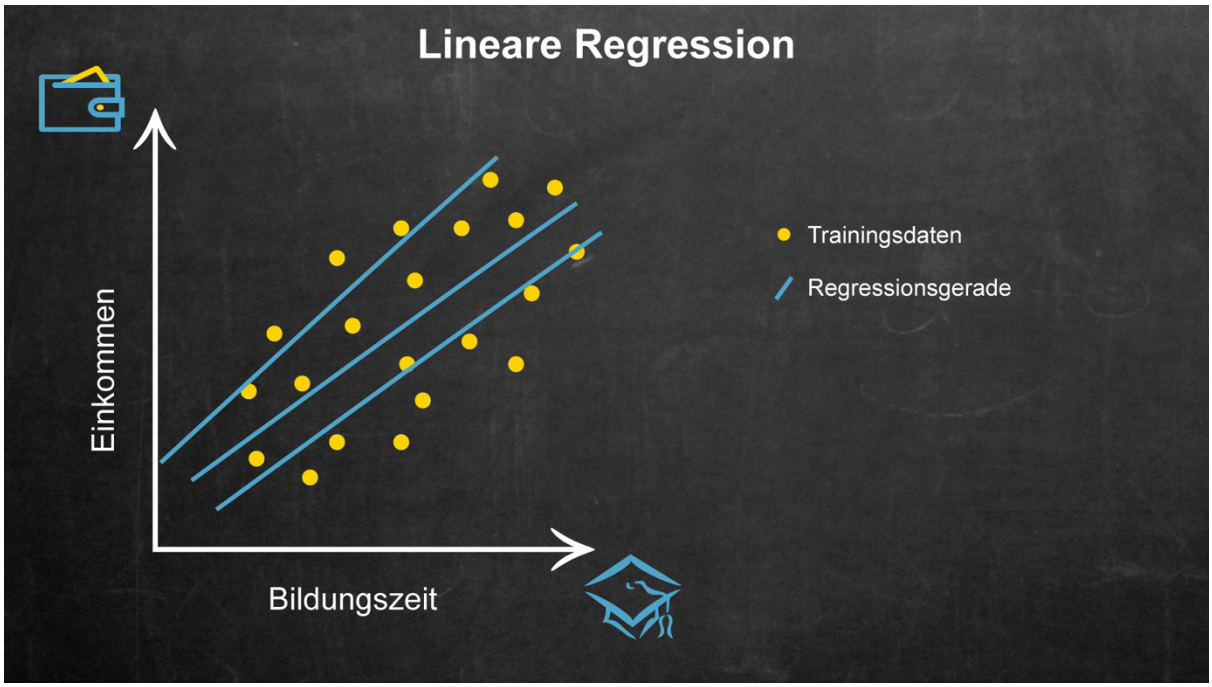
Aber was ist jetzt genau unser Ziel? Im Prinzip wollen wir jetzt in diese Punktwolke eine Kurve anpassen, die den Zusammenhang zwischen feature und der Zielgröße möglichst gut widerspiegelt. In unserem Beispiel kann man schon einen Trend erkennen, nämlich dass Personen mit höherer investierter Bildungszeit auch ein höheres Einkommen erzielen. Zudem kann man schon vermuten, dass sich hier eine stetig ansteigende gerade Linie – auch Gerade genannt – am besten in die Wolke einfügen würde. Eine lineare Regression ist hier also passend. Die Gerade hilft uns dann, für neue Beobachtungen eine Prognose ihres Einkommens zu machen.

Eine neue Beobachtung ist hier jetzt durch den orangenen Punkt gekennzeichnet. Von dieser kennen wir ihre investierte Bildungszeit, hier z. B. genau 13 Jahre, aber nicht ihr Einkommen. Wir können jetzt von diesem Punkt aus vertikal nach oben gehen, bis wir unsere Gerade erreichen und von hier aus horizontal, bis wir die Achse für das Einkommen erreichen. Den Wert, den wir hier ablesen können, entspricht dann unserer Einkommens-Prognose für diese neue Beobachtung. Hier wären das z. B. 3500 €.

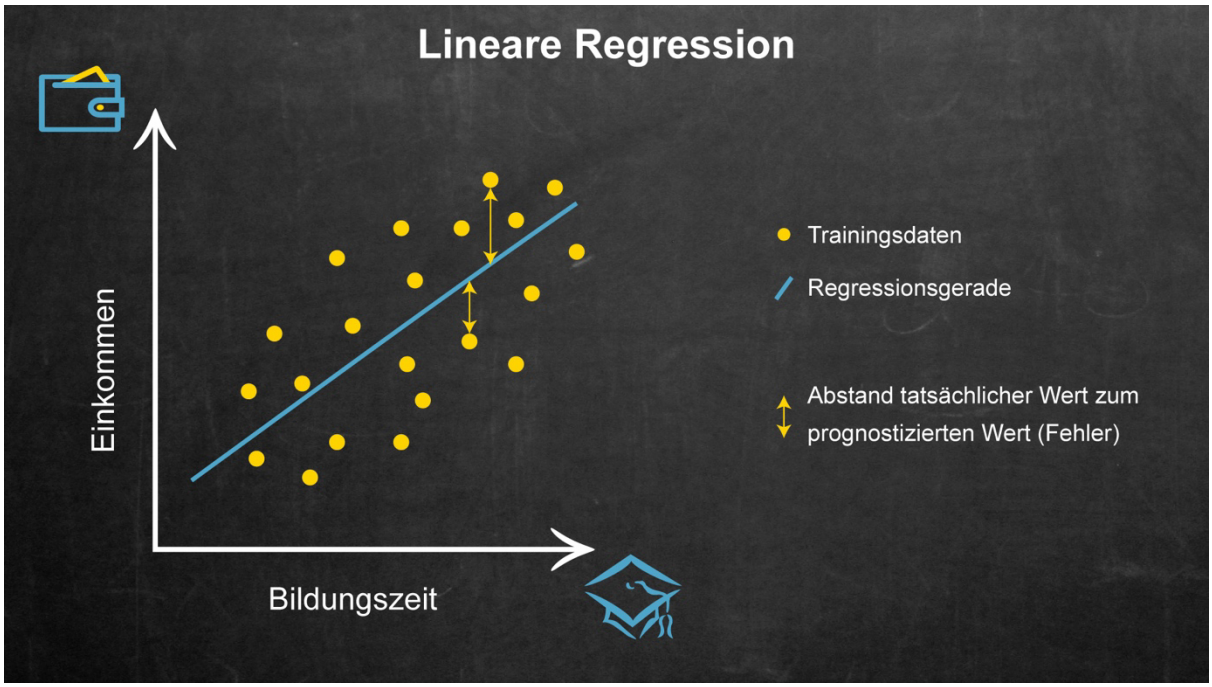


Lineare Regression – Finden der besten Gerade

Wie findet man nun die beste Gerade? Die Gerade könnte ja z. B. so oder so oder auch so aussehen.



Und was bedeutet überhaupt beste Gerade? Hier gibt es verschiedene Herangehensweisen. Die gemeinsame Idee ist, dass sie versuchen, die Abstände zwischen den Datenpunkten und der Geraden zu minimieren. Wir wollen also die Gerade finden, von der möglichst viele Datenpunkte möglichst wenig Abstand haben.

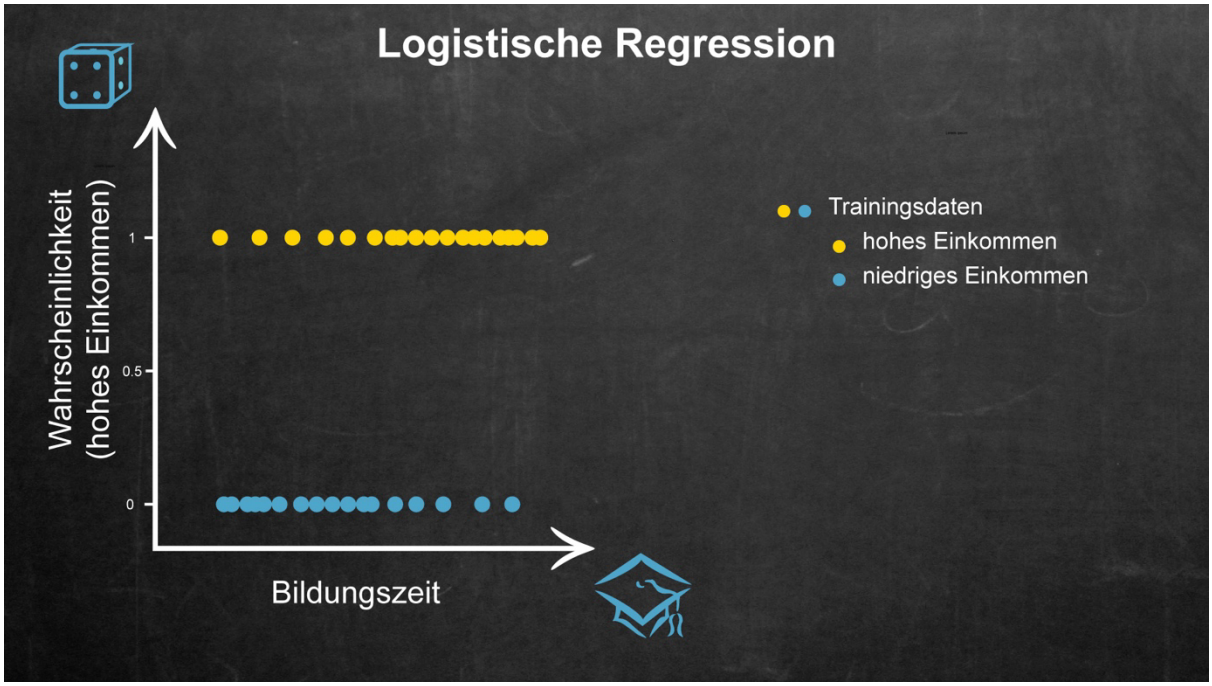


Warum? Diese Abstände kann man auch als Fehler interpretieren, den wir beim Prognostizieren des Einkommens machen. Dieser Fehler sollte natürlich möglichst gering sein. Der Fehler beschreibt hier im Beispiel die Differenz zwischen dem tatsächlichen Einkommen einer Person aus den Trainingsdaten und dem durch unsere Geraden prognostizierten Einkommen dieser Person. Wenn wir den gesamten Fehler der durch die Geraden prognostizierten Werte messen wollen, können wir dazu die Abstände jedes einzelnen Datenpunkts zu der Geraden aufsummieren. Da sowohl Abstände nach oben als auch nach unten Fehler sind und sich diese beim Summieren nicht aufheben sollen, wird häufig die quadrierte Abweichung verwendet.

Um nun die beste Gerade, also die mit dem geringsten Fehler, zu finden, könnten wir verschiedene Geraden ausprobieren, d.h. wir verändern ihre Lage in der Punktwolke. Dann können wir für die Geraden die jeweiligen Abstände zu den Beobachtungen messen, aufsummieren und diese Summe zwischen verschiedenen Geraden vergleichen. Am Ende wählen wir dann die Gerade mit dem geringsten Fehler.

Logistische Regression – Beispiel

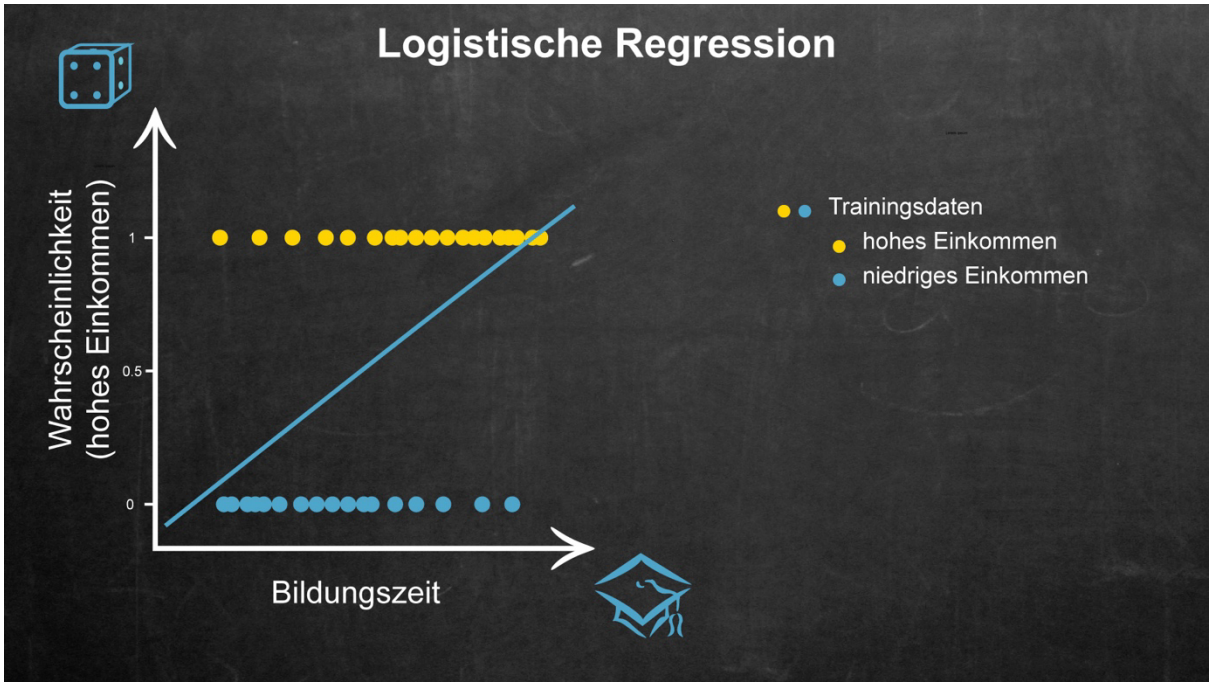
Bisher haben wir uns eine metrische Zielgröße angeschaut. Wir können uns aber auch Situationen vorstellen, bei denen wir eigentlich eine kategoriale Zielgröße haben. Um bei unserem vorherigen Beispiel zu bleiben, betrachten wir jetzt z. B. nicht das genaue Einkommen, sondern nur die beiden Einkommenskategorien „hoch“ und „niedrig“. Wir versuchen jetzt aber nicht einen direkten Zusammenhang zwischen feature und der eigentlichen Zielgröße herzustellen, sondern einen Zusammenhang zwischen feature und der Wahrscheinlichkeit, dass eine der beiden Kategorien zutrifft. Diese Wahrscheinlichkeit ist dann wieder eine metrische Größe. Schauen wir uns mal unser Beispiel näher an.



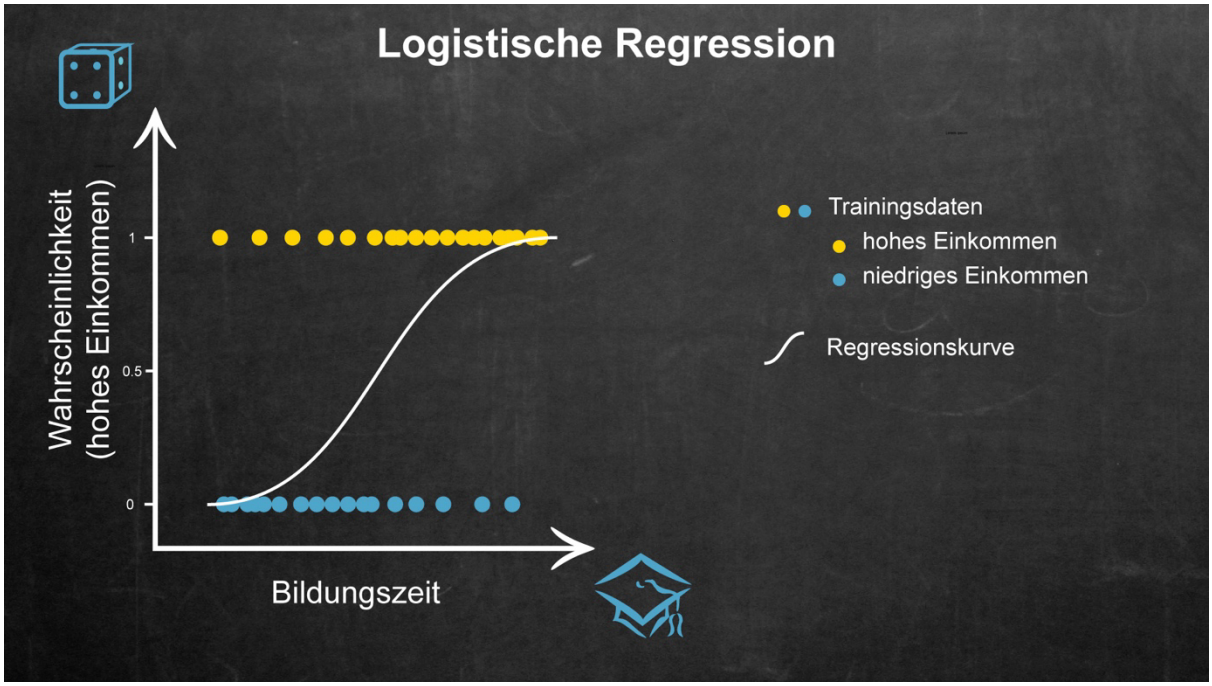
Wir tragen wieder unsere Trainingsdaten in ein Koordinatensystem ein. Die investierte Zeit in Bildung ist weiterhin auf der horizontalen Achse abgetragen. Beobachtungen mit hohem Einkommen sind hier durch die gelben Punkte und Beobachtungen mit niedrigem Einkommen durch die blauen Punkte repräsentiert. Auf der vertikalen Achse sind jetzt aber, wie bereits gesagt, nicht die direkten Einkommen abgetragen, sondern die Wahrscheinlichkeiten, z. B. in die Kategorie „hoch“ zu fallen. Wir sehen schon, dass es zu einer Ballung der gelben Punkte bei einer höheren Bildungszeit kommt und eine Ballung der blauen Punkte bei geringerer Bildungszeit.

Logistische Regression – Idee

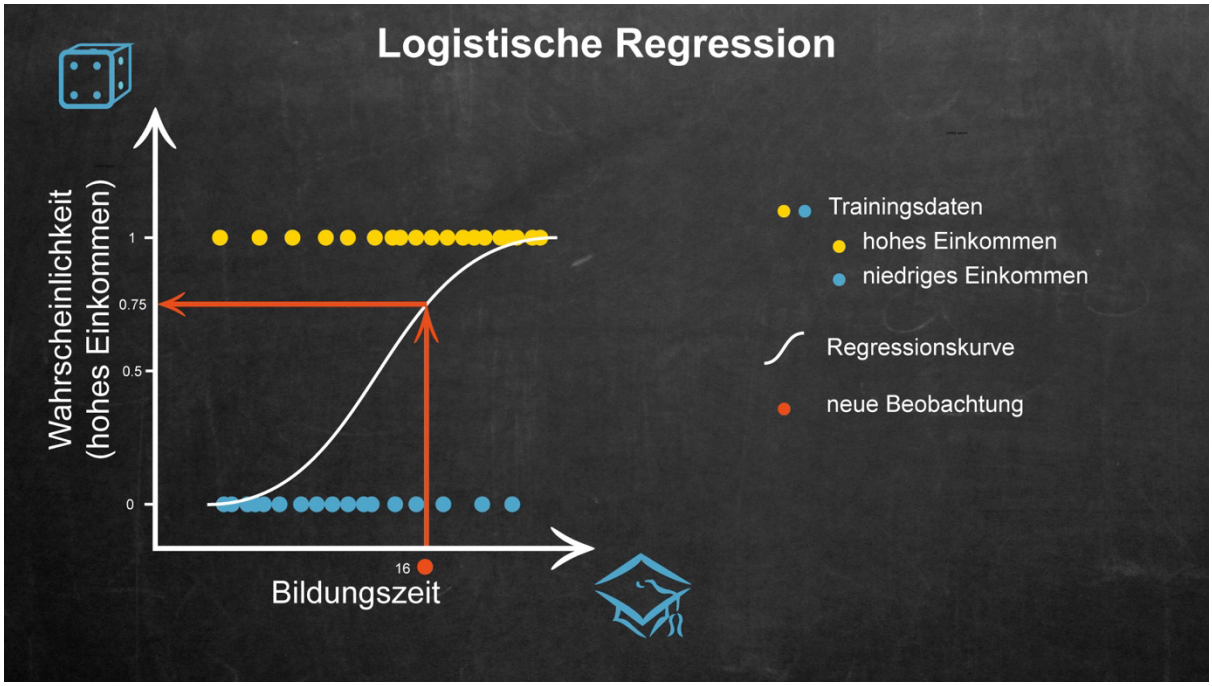
Ähnlich zur linearen Regression wollen wir auch hier eine Kurve in die Datenwolke einpassen. Hier ist eine Gerade allerdings nicht sinnvoll, da dann auch prognostizierte Wahrscheinlichkeiten kleiner oder größer als 0 bzw. 1 auftreten könnten. Wahrscheinlichkeiten liegen aber naturgemäß zwischen 0 und 1.



Hier kommt die logistische Regression zum Einsatz. Mit ihrer Hilfe passen wir anstatt einer Geraden eine S-Kurve in die Datenpunkte ein. Die genaue Ermittlung unterscheidet sich von der linearen Regression. Das wollen wir hier aber erstmal nicht weiter thematisieren.



Ganz ähnlich zur linearen Regression können wir jetzt aber mit dieser S-Kurve für eine neue Beobachtung die Wahrscheinlichkeiten vorhersagen, zu einer bestimmten Einkommenskategorie zu gehören.



Eine neue Beobachtung ist hier durch den orangenen Punkt gekennzeichnet. Von dieser kennen wir wieder nur die investierte Bildungszeit, hier 16 Jahre. Folgen wir wieder dem vertikalen orangenen Pfeil bis zur S-Kurve und gehen von dort aus entlang des horizontalen Pfeils bis zur Achse, können wir die Wahrscheinlichkeit ablesen, dass diese neue Beobachtung ein hohes Einkommen hat. Hier wären es 75 %. Die Wahrscheinlichkeit, ein niedriges Einkommen zu haben, läge dann bei 25%.

Ein Vorteil dieser Regressionsverfahren ist ihre intuitive Herangehensweise und Interpretierbarkeit. Ein Nachteil ist, dass wir im Vorhinein starke Annahmen treffen müssen über den Zusammenhang zwischen Zielgröße und feature, hier z. B. einen linearen Zusammenhang. Das kann bei komplexen Zusammenhängen in der Praxis schwierig sein und bei falschen Vermutungen zu verzerrten Ergebnissen führen.

Abschluss

Einblendungen kleine Grafiken Lineare und Logistische Regression

Wir kennen nun die lineare und die logistische Regression und ihre gemeinsame Idee, eine Gerade oder S-Kurve möglichst gut an unsere Punktwolke anzupassen. Mit ihrer Hilfe können wir metrische Zielgrößen oder Wahrscheinlichkeiten für kategoriale Zielgrößen vorhersagen. Anwendungsbeispiele finden sich überall. Neben der bereits besprochenen Prognose von Einkommen werden Regressionen z. B. für Prognosen von Mieten oder Krankheitsrisiken oder auch für Wettervorhersagen verwendet.

Weiterführendes Material

Fachbücher:

Guter Einstieg ins Thema, anschaulich erläutert, keine Formeln oder tiefergehende methodische Erläuterungen:

Kossen, J., & Müller, M. E. (2019). Regression – Voll im Trend. In K. Kersting, C. Lampert, & C. Rothkopf (Hrsg.), *Wie Maschinen lernen* (S. 39-43). Springer, Wiesbaden.

Kossen, J., & Müller, M. E. (2019). Lineare Regression – Einfach nur ein Strich? In K. Kersting, C. Lampert, & C. Rothkopf (Hrsg.), *Wie Maschinen lernen* (S. 61-67). Springer, Wiesbaden.

Schüler, T. (2019). Logistische Regression – Schubladendenken mit Wahrscheinlichkeiten. In K. Kersting, C. Lampert, & C. Rothkopf (Hrsg.), *Wie Maschinen lernen* (S. 105-110). Springer, Wiesbaden.

Auch wenn dieses Buch mit R anstatt Python arbeitet, anschauliche Erklärung der Methoden, tiefergehende methodische Erläuterungen:

Lantz, B. (2015). *Machine learning with R* (2. Auflage). Packt Publishing Ltd, Birmingham.
- Chap.6: Forecasting Numeric Data – Regression Methods

Klassisches Werk für Statistisches/Maschinelles Lernen, tiefergehende methodische Erläuterungen:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2. Auflage). Springer.
- Chap. 3: Linear Regression
- Chap. 4.3: Logistic Regression

Videos/Kurse:

Kurzer Einstieg ins Thema, anschaulich erläutert:

So lernen Maschinen: #3 Überwachtes Lernen - Regression.
https://ki-campus.org/videos/solernenmaschinen

Etwas weitergehender Einstieg ins Thema, anschaulich erläutert:

Elements of AI, Chap. 4: Machine learning, III. Regression.
https://course.elementsofai.com/4/3

Disclaimer

Transkript zu dem Video „Woche 06 Theorie: Regression“, Katja Theune.