Maschinelles Lernen 03

Prof. Dr. David Spieler – david.spieler@hm.edu

Hochschule München

4. November 2019

Logistische Regression

Logistische Regression

Logistische Regression

Klassifikation

Frage

Können wir die Prinzipien der linearen Regression auch für Klassifikation verwenden?

Beispiel

Wir wollen wissen, ob wir als Bank verschiedenen Kunden einen Kredit geben wollen. Hierfür wollen wir wissen, ob ein Kunde wahrscheinlich den Kredit zurückzahlen wird (Klasse 1) oder nicht (Klasse 0). Wir nehmen an, dass lediglich das durchschnittliche Bruttoeinkommen x in Euro dafür ausschlaggebend ist.

Logistische Regression

Klassifikation

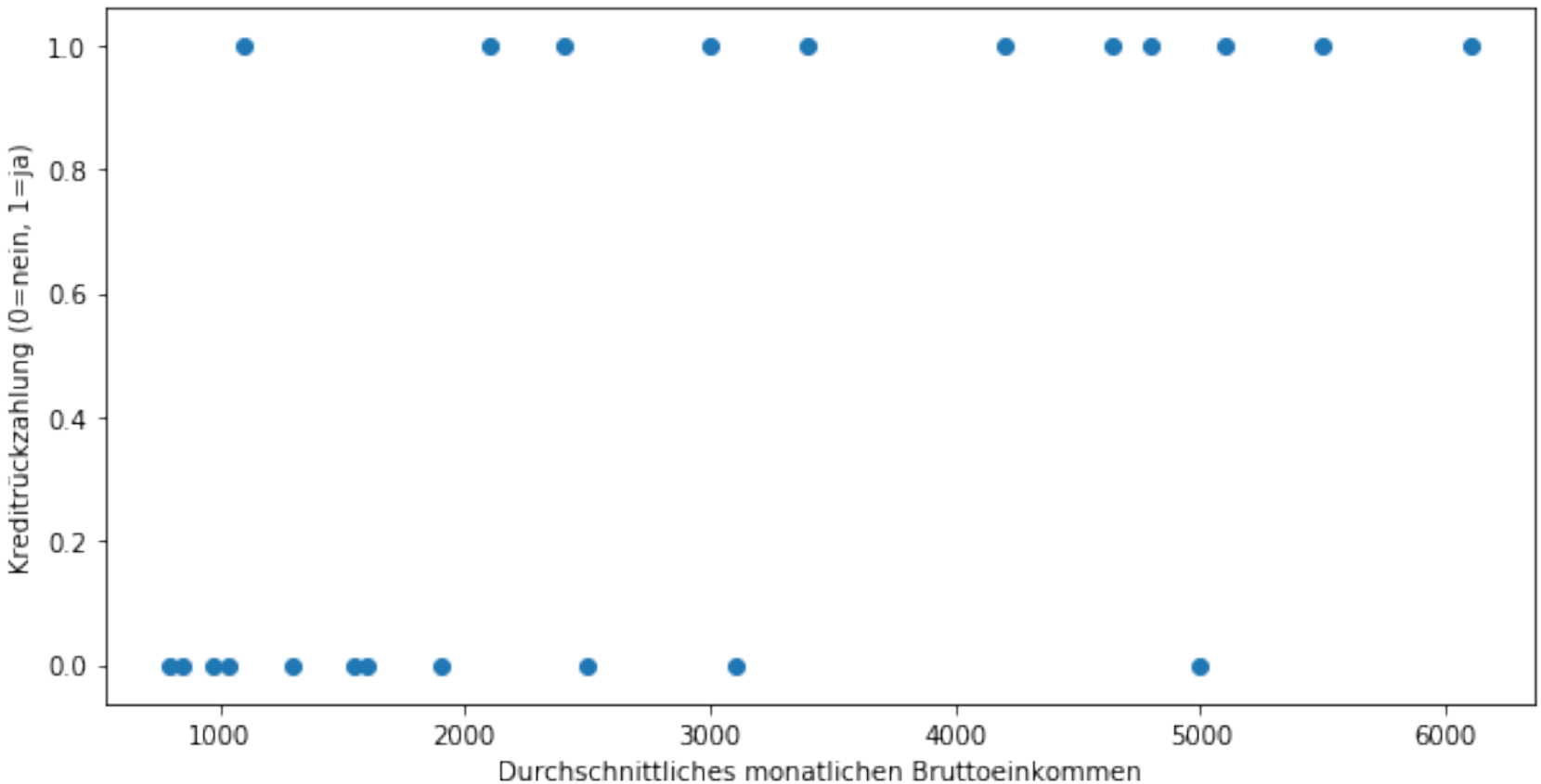

Abbildung 1: Kreditdaten als Scatterplot.

Logistische Regression

Klassifikation

Wir könnten auch hier versuchen eine Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) = \mathbf{w}_1 x + \mathbf{w}_0$$

durch lineare Regression bestimmen. Die Eingabe x wäre dann das durchschnittle Monatseinkommen und die Ausgabe $f(x)$ sollte dann angeben, ob der Kunde kreditwürdig ist oder nicht,

Logistische Regression

Klassifikation

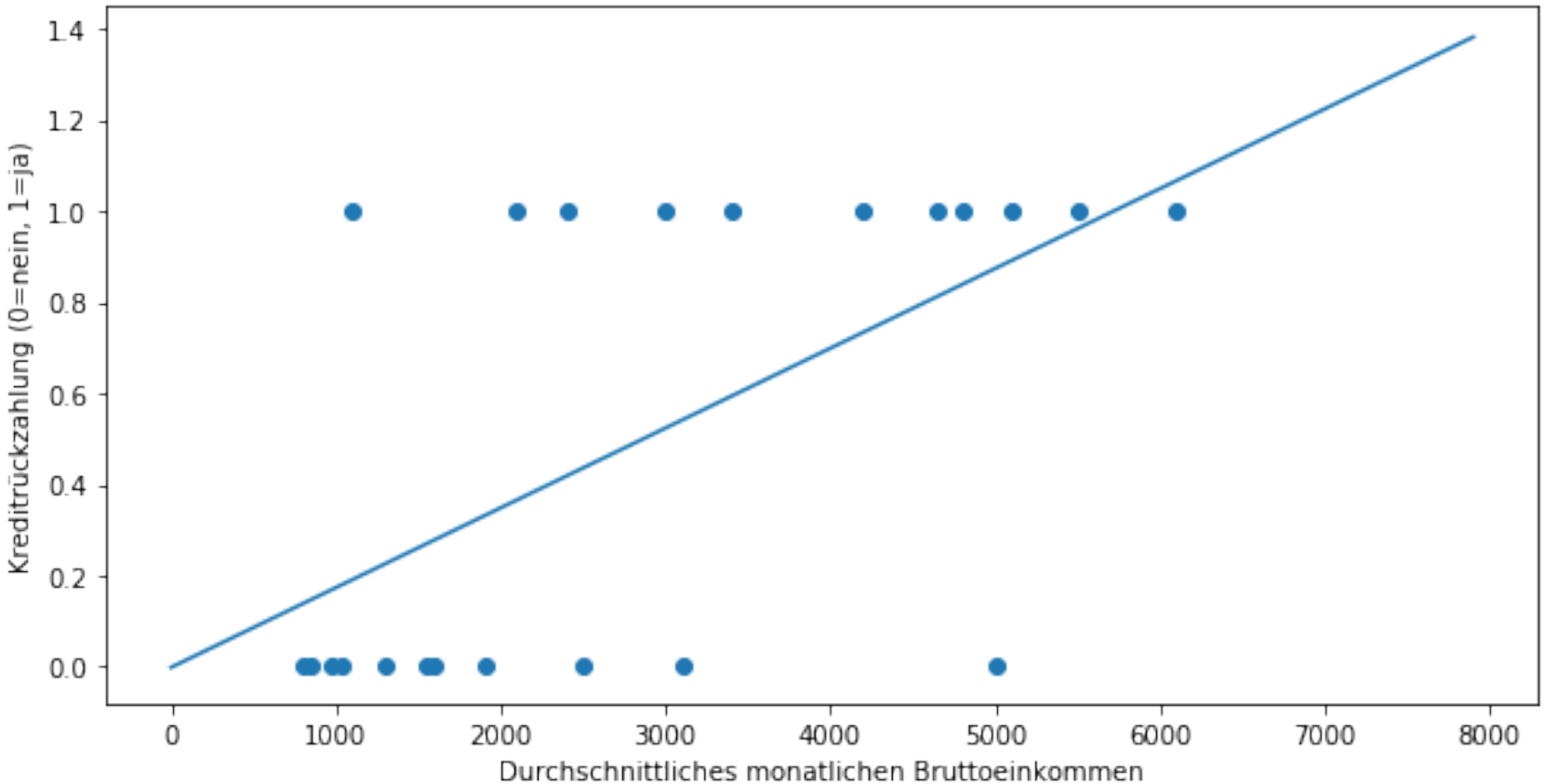

Abbildung 2: Lineares Regressionsmodell zur Bestimmung der Kreditwürdigkeit.

Logistische Regression

Klassifikation

Probleme:

- Die diskrete Ausgabemodalität (Klasse 0 oder 1) wird nicht eingehalten, da im Beispiel nicht nur die Ausgabewerte 0 und 1 sondern alle Werte im Intervall $[-0.00139646, 1.40044273]$ angenommen werden.
- Eventuell könnte man den Wert $f(x)$ als Wahrscheinlichkeit interpretieren, dass ein Kunde kreditwürdig ist. Aber auch dann machen die extremen Werte wie $f(0) = -0.00139646 < 0$ oder $f(8000) = 1.40044273 > 1$ keinen Sinn.

Logistische Regression

Klassifikation

Die Idee der Logistischen Regression ist es, die Idee der Schätzung der Wahrscheinlichkeit der Klassenzugehörigkeit aufzugreifen und den Wertebereich von f mit Hilfe der logistischen Funktion

$$logistic(x) = \frac{e^x}{1 + e^x}$$

unter Kontrolle zu bekommen.

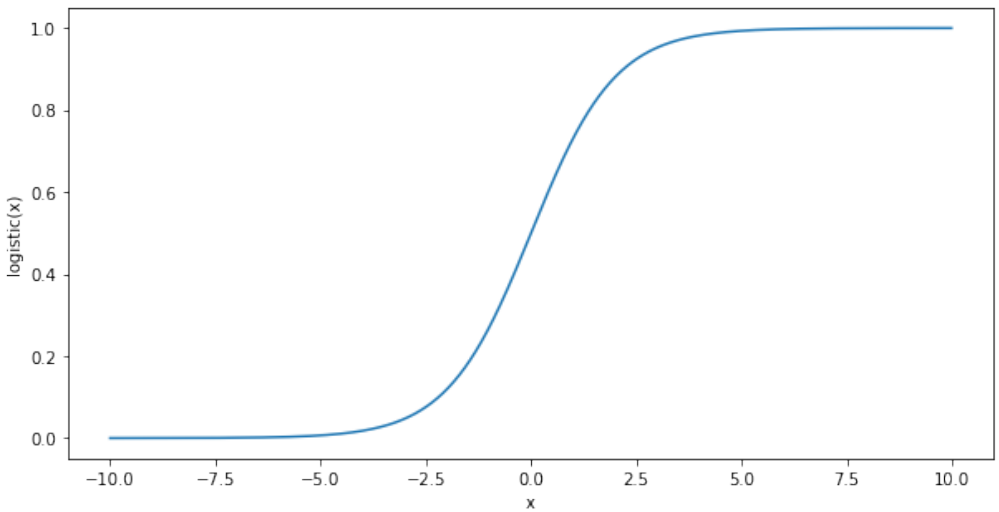

Abbildung 3: Logistische Funktion.

Logistische Regression

Klassifikation

Wir kombinieren also nun das Modell der linearen Regression

$$f(x) = \mathbf{w}_1 x + \mathbf{w}_0$$

mit der logistischen Funktion und erhalten

$$p(x) = \mathit{logistic}(f(x)) = \frac{e^{\mathbf{w}_1 x + \mathbf{w}_0}}{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}}.$$

Offensichtlich ist nun $p(x) \in (0, 1)$ für alle $x \in \mathbb{R}$.

Logistische Regression

Klassifikation

Wir interpretieren $p(x)$ als die Wahrscheinlichkeit, dass x zur Klasse 1 gehört, also

$$p(x) = \Pr(y = 1 \mid X = x).$$

Demnach gehört x zu Klasse 0 mit Wahrscheinlichkeit $1 - p(x)$, also

$$\Pr(y = 0 \mid X = x) = 1 - \Pr(y = 1 \mid X = x) = 1 - p(x).$$

Logistische Regression

Maximum Likelihood

Während wir bei der linearen Regression beim Training zur Bestimmung der Parameter den Fehler (RSS,MSE) minimiert haben, werden wir bei der logistischen Regression die Maximum Likelihood Methodik verwenden.

Beispiel

Wir betrachten ein eine (möglicherweise unfaire) Münze, welche mit unbekannter WK $w \in [0,1]$ Kopf zeigt und mit WK $1-w$ Zahl. Wir werfen die Münze n mal und erhalten k mal Kopf (und $n-k$ mal Zahl). Wie groß ist w?

Logistische Regression

Maximum Likelihood

Das Grundprinzip von Maximum Likelihood (ML) ist, dass wir die Parameter eines Modells so bestimmen, dass die Wahrscheinlichkeit, dass das Modell die beobachteten Daten generiert maximiert wird. Im Beispiel des n-maligen Münzwurfs ist die WK einer bestimmten Sequenz, die genau k mal Kopf enthält in Abhängigkeit des Parameters w:

$$L(w) = w^k(1-w)^{n-k}.$$

Der Term $L(w)$ wird auch Likelihood genannt. Formal suchen wir also bei ML

$$\arg \max_{w \in [0,1]} L(w).$$

Logistische Regression

Maximum Likelihood

Auch bei ML bilden wir meist die Ableitung, um das Optimum zu finden.

$$\begin{aligned}\frac{\partial L(w)}{\partial w} &= \frac{\partial}{\partial w} w^k (1-w)^{n-k} \\ &= k w^{k-1}(1-w)^{n-k} + w^k (n-k)(1-w)^{n-k-1}(-1) \\ &= w^{k-1}(1-w)^{n-k-1}\left[k(1-w) - (n-k)w\right]\end{aligned}$$

$$\frac{\partial L(w)}{\partial w} = 0 \Leftrightarrow w = 0 \vee w = 1 \vee w = \frac{k}{n}$$

da

$$k(w-1) - (n-k)w = 0 \Rightarrow k - kw - nw + kw = 0 \Rightarrow k - nw = 0.$$

Logistische Regression

Maximum Likelihood

Zur Bestimmung des Maximums überprüfen wir

- $w = 0$: $L(0) = 0^k(1-0)^{n-k} = 0$
- $w = 1$: $L(1) = 1^k(1-1)^{n-k} = 0$
- $w = \frac{k}{n}$: $L(\frac{k}{n}) = \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k} > 0$ (falls $k > 0$ und $k \neq n$)

und erhalten somit die Maximum Likelihood Schätzung

$$w = \frac{k}{n},$$

welche auch der Intuition entspricht.

Logistische Regression

Maximum Likelihood

Auch bei der logistischen Regression bestimmen wir nun die Likelihood durch

$$\begin{aligned} L(\mathbf{w}) &= \prod_{i=1}^{n} \begin{cases} p(x^{(i)}) & \text{falls } y^{(i)} = 1 \\ 1 - p(x^{(i)}) & \text{falls } y^{(i)} = 0 \end{cases} \\ &= \prod_{\{i \mid y^{(i)}=1\}} p(x^{(i)}) \prod_{\{i \mid y^{(i)}=0\}} (1 - p(x^{(i)})). \end{aligned}$$

Hiervon das Maximum mit Hilfe der Ableitung zu berechnen ist mühselig. Stattdessen können wir auch das Minimum des negativen Logarithmus suchen gegeben durch

$$-\log L(\mathbf{w}) = -\sum_{\{i \mid y^{(i)}=1\}} \log p(x^{(i)}) - \sum_{\{i \mid y^{(i)}=0\}} \log(1 - p(x^{(i)})).$$

Logistische Regression

Maximum Likelihood

$$\frac{\partial}{\partial \mathbf{w}_j}(-\log L(\mathbf{w})) = \frac{\partial}{\partial \mathbf{w}_j}\left(-\sum_{\{i \mid y^{(i)}=1\}} \log p(x^{(i)}) - \sum_{\{i \mid y^{(i)}=0\}} \log(1-p(x^{(i)}))\right)$$

$$= -\sum_{\{i \mid y^{(i)}=1\}} \frac{\frac{\partial}{\partial \mathbf{w}_j} p(x^{(i)})}{p(x^{(i)})} + \sum_{\{i \mid y^{(i)}=0\}} \frac{\frac{\partial}{\partial \mathbf{w}_j} p(x^{(i)})}{1-p(x^{(i)})}$$

Um hier weiter umformen zu können benötigen wir einige Zwischenschritte.

Logistische Regression

Maximum Likelihood

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{w}_j} p(x) &= \frac{\partial}{\partial \mathbf{w}_j} \frac{e^{\mathbf{w}_1 x + \mathbf{w}_0}}{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}} \\
&= \frac{\frac{\partial}{\partial \mathbf{w}_j}\left(e^{\mathbf{w}_1 x + \mathbf{w}_0}\right)\left(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}\right) - \left(e^{\mathbf{w}_1 x + \mathbf{w}_0}\right)\frac{\partial}{\partial \mathbf{w}_j}\left(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}\right)}{\left(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}\right)^2} \\
&= \frac{e^{\mathbf{w}_1 x + \mathbf{w}_0}\left(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}\right) - \left(e^{\mathbf{w}_1 x + \mathbf{w}_0}\right)^2}{\left(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}\right)^2} \frac{\partial}{\partial \mathbf{w}_j}\left(\mathbf{w}_1 x + \mathbf{w}_0\right) \\
&= \frac{e^{\mathbf{w}_1 x + \mathbf{w}_0}}{\left(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}\right)^2} \begin{cases} 1 & \text{falls } j = 0 \\ x & \text{falls } j = 1 \end{cases}
\end{aligned}
$$

Logistische Regression

Maximum Likelihood

$$
\begin{aligned}
1 - p(x) &= 1 - \frac{e^{\mathbf{w}_1 x + \mathbf{w}_o}}{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}} \\
&= \frac{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0} - e^{\mathbf{w}_1 x + \mathbf{w}_0}}{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}} \\
&= \frac{1}{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}}
\end{aligned}
$$

Logistische Regression

Maximum Likelihood

$$
\begin{aligned}
\frac{\frac{\partial}{\partial \mathbf{w}_j} p(x)}{p(x)} &= \frac{1}{p(x)} \frac{\partial}{\partial \mathbf{w}_j} p(x) \\
&= \frac{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}}{e^{\mathbf{w}_1 x + \mathbf{w}_0}} \frac{e^{\mathbf{w}_1 x + \mathbf{w}_0}}{(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0})^2} \begin{cases} 1 & \text{falls } j = 0 \\ x & \text{falls } j = 1 \end{cases} \\
&= \frac{1}{1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}} \begin{cases} 1 & \text{falls } j = 0 \\ x & \text{falls } j = 1 \end{cases} \\
&= (1 - p(x)) \begin{cases} 1 & \text{falls } j = 0 \\ x & \text{falls } j = 1 \end{cases}
\end{aligned}
$$

Logistische Regression

Maximum Likelihood

$$
\begin{aligned}
\frac{\frac{\partial}{\partial \mathbf{w}_j} p(x)}{1 - p(x)} &= \frac{1}{1 - p(x)} \frac{\partial}{\partial \mathbf{w}_j} p(x) \\
&= (1 + e^{\mathbf{w}_1 x + \mathbf{w}_0}) \frac{e^{\mathbf{w}_1 x + \mathbf{w}_0}}{(1 + e^{\mathbf{w}_1 x + \mathbf{w}_0})^2} \begin{cases} 1 & \text{falls } j = 0 \\ x & \text{falls } j = 1 \end{cases} \\
&= p(x) \begin{cases} 1 & \text{falls } j = 0 \\ x & \text{falls } j = 1 \end{cases}
\end{aligned}
$$

Logistische Regression

Maximum Likelihood

Zusammenfassung

$$\frac{\partial}{\partial \mathbf{w}_0}(-\log L(\mathbf{w})) = -\sum_{\{i \mid y^{(i)}=1\}}(1-p(x^{(i)})) + \sum_{\{i \mid y^{(i)}=0\}} p(x^{(i)})$$

$$\frac{\partial}{\partial \mathbf{w}_1}(-\log L(\mathbf{w})) = -\sum_{\{i \mid y^{(i)}=1\}}(1-p(x^{(i)}))x^{(i)} + \sum_{\{i \mid y^{(i)}=0\}} p(x^{(i)})x^{(i)}$$

Ab hier kann dann das gewohnte Gradientenabstiegsverfahren verwendet werden, um die optimale Parameterbelegung zu finden.

Logistische Regression

Beispiel

Wenn wir diesen Ansatz auf unser Beispiel anwenden, erhalten wir die Parameter

$$\mathbf{w}_0 = -1.25238942, \mathbf{w}_1 = 0.000542.$$

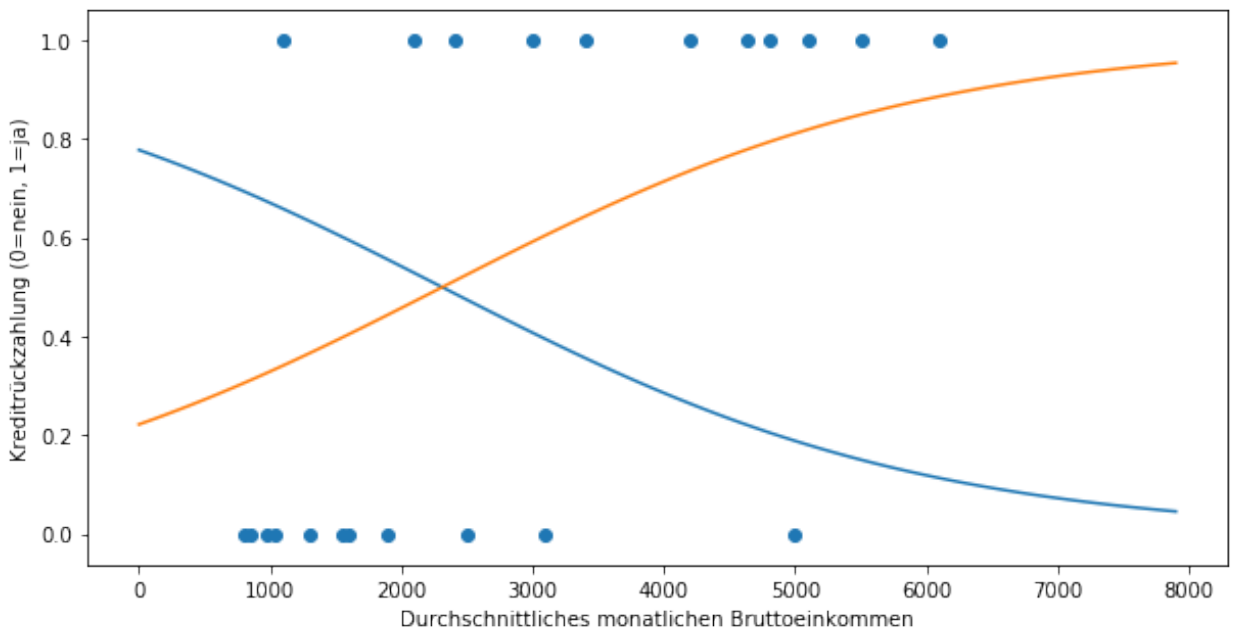

Abbildung 4: Kreditbeispiel: Wahrscheinlichkeit für die Rückzahlung (orange) $p(x)$ bzw. Nicht-Rückzahlung (blau) $1 - p(x)$.

Logistische Regression

Beispiel

Wir könnten also entscheiden, einer Person einen Kredit genau dann zu geben, wenn die Wahrscheinlichkeit, dass sie zurückzahlt größer ist als dass sie es nicht tut. Angenommen, unsere Schätzung der Wahrscheinlichkeit

$$\Pr(y \mid X = x)$$

würde der Wirklichkeit entsprechen, dann weist der Bayes Klassifikator jeder Beobachtung $x \in \mathcal{X}$ die wahrscheinlichste Klasse zu, also

$$f(x) = \arg\max_{y^* \in \mathcal{Y}} \Pr(y = y^* \mid X = x).$$

Logistische Regression

Beispiel

In unserem Fall würde der Bayes Klassifikator Personen mit einem Monatbrutto >2300 EUR als kreditwürdig einstufen (angenommen die beiden Klassen wären in der Gesamtgesellschaft gleich häufig vertreten).

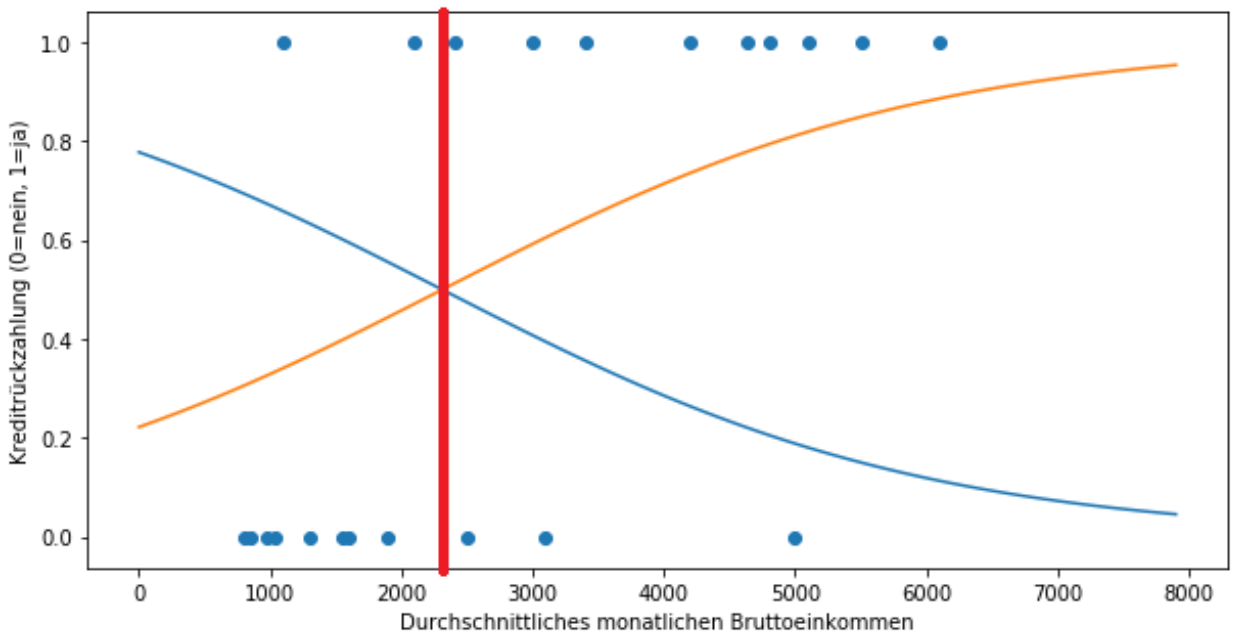

Abbildung 5: Kreditbeispiel: In Rot ist nun die Entscheidungsgrenze zwischen den beiden Klassen eingezeichnet.

Logistische Regression

Mehrdimensionale Logistische Regression

Natürlich lässt sich das Prinzip der logistischen Regression auch im Falle mehrdimensionaler Eingaben $\mathbf{x} \in \mathbb{R}^d$ verwenden. Hier wird das Modell (mit $\mathbf{x}_0 = 1$) beschrieben durch

$$p(\mathbf{x}) = \frac{e^{\mathbf{w}^T \mathbf{x}}}{1 + e^{\mathbf{w}^T \mathbf{x}}}$$

und der Gradient der negativen logarithmierten Likelihood ergibt sich aus

$$\frac{\partial}{\partial \mathbf{w}_j} \left(-\log L(\mathbf{w})\right) = -\sum_{\{i \mid y^{(i)}=1\}} (1-p(\mathbf{x}_j^{(i)}))\mathbf{x}_j^{(i)} + \sum_{\{i \mid y^{(i)}=0\}} p(\mathbf{x}_j^{(i)})\mathbf{x}_j^{(i)}$$

Logistische Regression

Nichtlineare Logistische Regression

Mit Hilfe einer Basiserweiterung

$$\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$$

und der mehrdimensionalen logistischen Regression können auch nichtlineare Klassifikatoren gelernt werden.

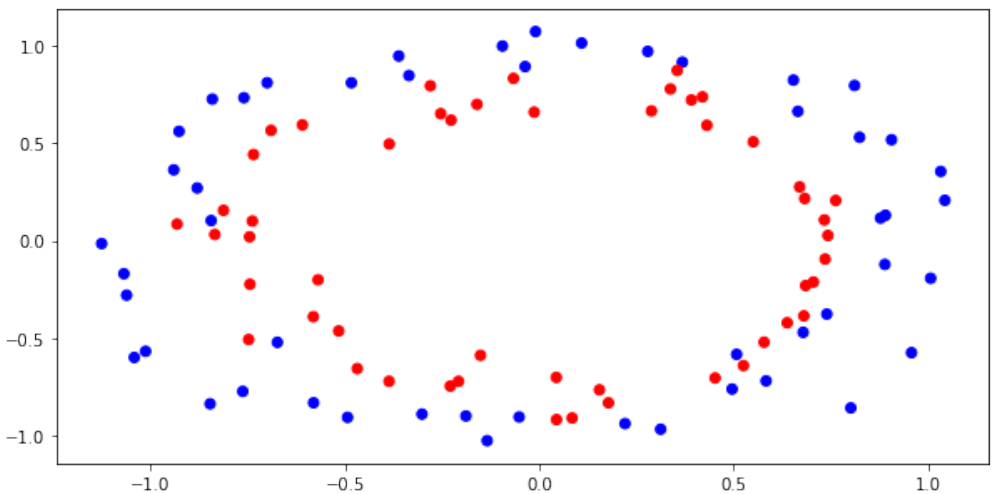

Abbildung 6: Zweidimensionaler Datensatz mit zwei Klassen, die nicht linear trennbar sind.

Logistische Regression

Nichtlineare Logistische Regression

Mit Hilfe der Basiserweiterung $\phi(\mathbf{x}) = (\mathbf{x}_1^2, \mathbf{x}_2^2)$ kann ein logistisches Regressionsmodell gelernt werden, welches die beiden Klassen trennt.

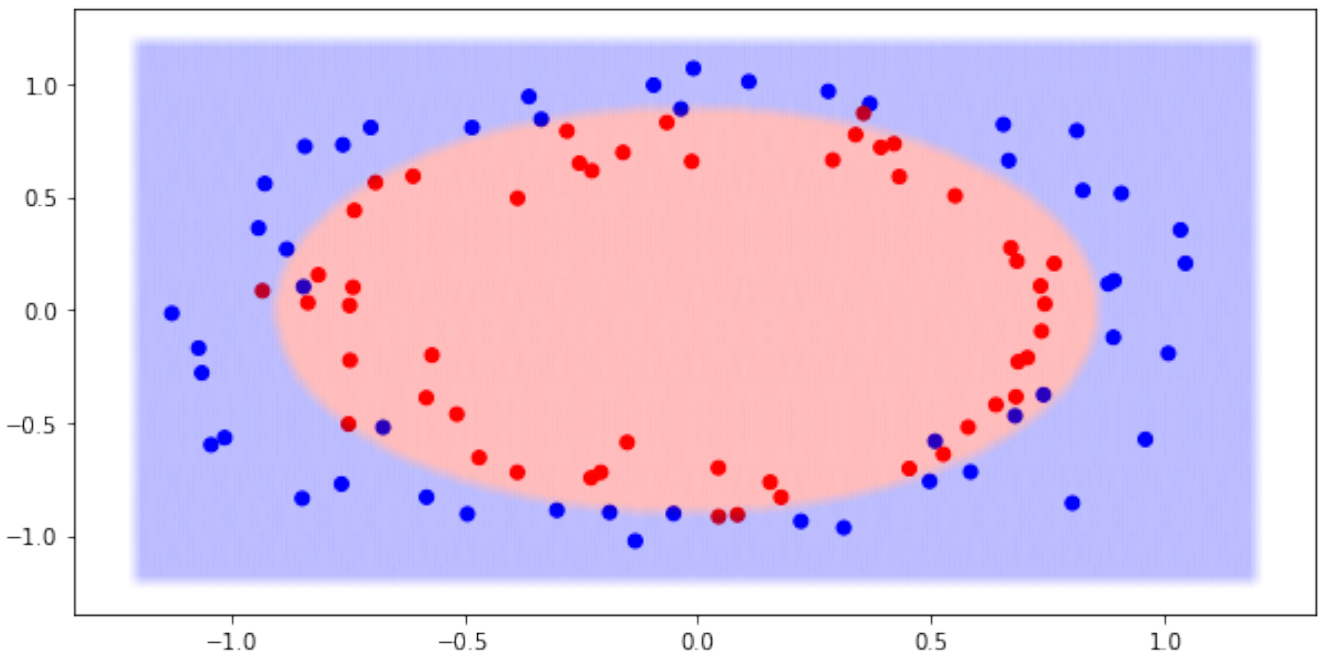

Abbildung 7: Zweidimensionaler Datensatz mit zwei Klassen, die nichtlinear (Kreis) trennbar sind.

Logistische Regression

Leistungsmetriken

Bei der binären Klassifikation können aus Kombination von Wahrheit und Vorhersage vier Fälle auftreten.

		prediction: +	prediction: -
truth	+	true ✓ positive tp	false ✗ negative fn
truth	-	false ✗ positive fp	true ✓ negative tn

Abbildung 8: Wahrheitsmatrix der binären Klassifikation.

Logistische Regression

Leistungsmetriken

Zählt man die Auftrittshäufigkeit der jeweiligen Fälle bei Anwendung des Klassifikators (üblicherweise) auf den Testdatensatz, so lassen sich Aussagen über die Güte des Modells treffen. Die häufigste Leistungsmetrik, die verwendet wird ist die Genauigkeit definiert als

$$\text{Genauigkeit} = \frac{tp + tn}{tp + tn + fp + fn}.$$

Die Genauigkeit (accuracy) ist der Anteil der korrekt klassifizierten Daten am Gesamtdatensatz. Meist wird daher versucht, die Genauigkeit zu maximieren.

Logistische Regression

Leistungsmetriken

Die Fehlerrate ist das intuitive Gegenteil der Genauigkeit und somit definiert als

$$\text{Fehlerrate} = \frac{fp + fn}{tp + tn + fp + fn} = 1 - \text{Genauigkeit}.$$

Meist ist man daran interessiert, die Fehlerrate zu minimieren.

Logistische Regression

Leistungsmetriken

Die Präzision (precision) ist der Anteil der korrekt positiv vorhergesagten Datensätze an der Gesamtheit der als positiv vorhergesagten Datensätze und definiert als

$$\text{Präzision} = \frac{tp}{tp + fp}.$$

Die Trefferquote (recall) ist der Anteil der korrekt positiv vorhergesagten Datensätze an der Gesamtheit der echt positiven Datensätze und definiert als

$$\text{Trefferquote} = \frac{tp}{tp + fn}.$$

Beide Metriken werden üblicherweise maximiert, wobei meist ein Kompromiss getroffen werden muss.

Logistische Regression

Leistungsmetriken

Es kommt auf die Anwendung an, welche Leistungsmetrik verwendet werden sollte:

- Medizinische Tests: Angenommen positiv ist gleich bedeutend mit krank, wie in "HIV positiv", dann sollte ein Test einen hohen Recall haben.
- Spam-Erkennung: Angenommen positiv bedeutet gewollte E-Mail, dann sollte ein Spam-Erkenner eine hohe Precision haben.