

2. Unüberwachte Regression

Die unüberwachte Regression ist ein Dimensionsreduktionsverfahren, das durch die Umkehr eines Regressionsmodells f einen latenten Raum $\mathbf{X}$ ermittelt [13]. Das heißt, die Muster des Datenraums sind in diesem Fall $\mathbf{Y}$ und es sind latente Variablen im Raum $\mathbf{X}$ gesucht. Die Idee dabei ist, dass der latente Raum $\mathbf{X}$ eine optimale Form besitzt, wenn sich über das gewählte Regressionsmodell f folgender Zusammenhang ergibt:

$$f(\mathbf{X}) = \mathbf{Y} \quad (2.1)$$

Dies würde bedeuten, dass es über ein allgemeines Regressionsverfahren möglich wäre, den Datenraum $\mathbf{Y}$ aus dem latenten Raum $\mathbf{X}$ vollständig zu rekonstruieren. Das würde wiederum heißen, dass alle Informationen des Datenraums in umgewandelter Form auch im latenten Raum enthalten sein würden und somit alle Fragen, die mit Hilfe des Datenraums $\mathbf{Y}$ beantwortet werden können, auch durch den latenten Raum $\mathbf{X}$ beantwortet werden könnten.

In der Realität wird es nicht möglich sein, den Zusammenhang $f(\mathbf{X}) = \mathbf{Y}$ herzustellen. Daher wird versucht, die Differenz, häufig in quadratischer und auf die Anzahl der Muster N normierter Form, zwischen $f(\mathbf{X})$ und $\mathbf{Y}$ zu minimieren. Das Quadrat der Differenz zwischen $f(\mathbf{X})$ und $\mathbf{Y}$ nennt man Datenraumrekonstruktions-Fehler [10]:

$$E(\mathbf{Y}, \mathbf{X}) = \frac{1}{N} \|\mathbf{Y} - f(\mathbf{X})\|_F^2 \quad (2.2)$$

Bei der unüberwachten Regression gibt es also einen hochdimensionalen Eingangsdatensatz $\mathbf{Y}$ und einen niedrigdimensionalen Ausgangsdatensatz $\mathbf{X}$, der mit Hilfe eines Regressionsmodells f erzeugt wurde. Man könnte das Regressionsmodell f dabei als eine Art *Bauanleitung*

ansehen, die maßgeblich den erzeugten latenten Raum $\mathbf{X}$ beeinflusst, jedoch nach der Erzeugung nicht mehr von Interesse ist.

2.1. Loss-Funktion

Es gibt verschiedene Fehler-Funktionen / Loss-Funktionen, mit denen ein Loss-Fehler berechnet werden kann. In dieser Arbeit werden die Einzelfehler wie beim Datenraumrekonstruktions-Fehler in quadratischer Form addiert, sodass für ein einzelnes Muster bei Verwendung des Nadaraya-Watson-Schätzers als Regressionsmodell gilt:

$$L(\mathbf{y}, f(\mathbf{x}, \mathbf{X})) = \|\mathbf{y} - f(\mathbf{x}, \mathbf{X})\|_2^2 \quad (2.3)$$

Und für alle Muster:

$$L(\mathbf{Y}, f(\mathbf{X})) = \sum_{i=1}^N \|\mathbf{y}_i - f(\mathbf{x}_i, \mathbf{X})\|_2^2 \quad (2.4)$$

Wenn ein Muster $\mathbf{x}_i \in \mathbf{X}$ mit der Loss-Funktion $L(\mathbf{y}_i, f(\mathbf{x}_i, \mathbf{X}))$ optimiert werden soll, liegt es Sonderfall vor, da das Muster $\mathbf{x}_i$ das Ergebnis von $f(\mathbf{x}_i, \mathbf{X})$ beeinflusst. Dies kann dazu führen, dass das Muster bei der Optimierung von den anderen Mustern entfernt wird, also in einen Bereich verschoben wird, in dem nur noch das Muster selber Einfluss auf das Ergebnis von $f(\mathbf{x}_i, \mathbf{X})$ hat. Damit gilt praktisch $f(\mathbf{x}_i, \mathbf{X}) \approx f(\mathbf{x}_i, \mathbf{x}_i) = \mathbf{y}_i$, was einen Loss-Fehler im Bereich von 0 verursacht.

Abbildung 2.1 zeigt einen unter Betrachtung des Loss-Fehlers perfekten zweidimensionalen latenten Raum mit Verwendung des Nadaraya-Watson-Schätzers und einer Kernel-Funktion, die ab einer bestimmten Entfernung den Wert 0 zurückgibt, wenn keine Leave-One-Out Kreuzvalidierung verwendet wird. Zu sehen sind dabei neun Muster und ihre Einflussbereiche, außerhalb der Bereiche ist der Nadaraya-Watson-Schätzer undefiniert. Das Ergebnis aus der Abbildung ist nicht gewünscht, da hierbei weder Entfernungen noch Nachbarschaften erhalten bleiben. Soll das Muster $\mathbf{x}_i \in \mathbf{X}$ optimiert

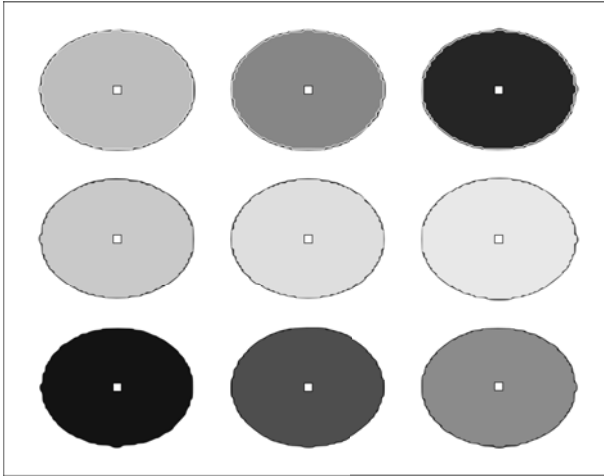


Abbildung 2.1.: Ein perfekter, latenter Raum ohne Leave-One-Out

werden, sollte daher die Loss-Funktion mit der Methode Leave-One-Out Kreuzvalidierung verwendet werden [15], hierbei gilt:

$$L(\mathbf{y}_i, f(\mathbf{x}_i, \mathbf{X} \setminus \mathbf{x}_i)) \quad (2.5)$$

Durch die Loss-Funktion kann eine Dimensionsreduktion bewertet werden. Auch ist es möglich den Loss-Fehler durch eine differenzierbare Loss-Funktion mit einem Optimierungsverfahren wie dem Gradientenabstieg zu minimieren. Die Loss-Funktion selber ist differenzierbar, jedoch müssen auch das verwendete Regressionsmodell und die verwendete Kernel-Funktion differenzierbar sein, um einen Gradientenabstieg durchführen zu können.

2.2. Co-Ranking-Matrix

Ein weiteres Fehlermaß, mit welchem die Qualität einer Dimensionsreduktion geprüft werden kann, ist die Co-Ranking-Matrix. Sie vergleicht Nachbarschaften, repräsentiert durch Ränge, im Datenraum und im latenten Raum [12]. Zunächst wird die Distanz im Datenraum

zwischen den Mustern $\mathbf{y}_i$ und $\mathbf{y}_j$ als δ_{ij} und die Distanz im latenten Raum zwischen den Mustern $\mathbf{x}_i$ und $\mathbf{x}_j$ als d_{ij} eingeführt. Der Rang im Datenraum ist nun wie folgt definiert:

$$p_{ij} = |\{k \mid \delta_{ik} \leq \delta_{ij} \text{ und } 1 \leq k < j < N\}| \quad (2.6)$$

Und der Rang für den latenten Raum:

$$r_{ij} = |\{k \mid d_{ik} \leq d_{ij} \text{ und } 1 \leq k < j < N\}| \quad (2.7)$$

Die Co-Ranking-Matrix $\mathbf{Q}$ definiert die Abweichungen zwischen den Rängen im Datenraum und im latenten Raum. Die einzelnen Matrixeinträge sind dabei wie folgt definiert:

$$q_{kl} = |\{(i, j) \mid p_{ij} = k \text{ und } r_{ij} = l\}| \quad (2.8)$$

Die Diagonaleinträge der Matrix stehen für fehlerfrei in den latenten Raum übertragene Ränge. Bei $p_{ij} > r_{ij}$ liegt eine Intrusion vor, bei $p_{ij} < r_{ij}$ eine Extrusion. Unter Angabe einer Nachbarschaftsgröße K gibt es das Qualitätsmaß $Q_{NX}(K)$. Dieses Qualitätsmaß gibt an, in wie weit die Nachbarschaften unter Berücksichtigung von K erhalten bleiben, das heißt, in wie weit die Menge der k -nächsten Nachbarn mit $k = K$ erhalten bleibt. Die Reihenfolge innerhalb der Menge spielt dabei keine Rolle. Da sich das Ergebnis je nach K ändert, wird $Q_{NX}(K)$ mitunter auch als Funktion von K dargestellt. Definiert ist $Q_{NX}(K)$ wie folgt:

$$Q_{NX}(K) = \frac{1}{KN} \cdot \sum_{k=1}^K \sum_{l=1}^K q_{kl} \quad (2.9)$$

Die Co-Ranking-Matrix ist ein Maß, das für die Überprüfung des Erhalts von Nachbarschaften geeignet ist. Der Erhalt von Entfernungen wird von der Co-Ranking-Matrix nicht ausgewertet, auch wenn in vielen Fällen davon auszugehen ist, dass sich die Beibehaltung von Nachbarschaften auch positiv auf den Erhalt von Distanzen auswirkt. Durch die Auswertung von Nachbarschaften eignet sich dieses Qualitätsmaß vor allem für latente Räume, die mit einem nachbarschaftsauswertenden Verfahren weiterverwendet werden.

2.3. Unüberwachte Nächste Nachbarn

Das Verfahren der *Unüberwachten Nächsten Nachbarn* (UNN) ist eine von Kramer vorgestellte Methode zur Dimensionsreduktion [9]. Es handelt sich dabei um eine unüberwachte Regression, die als Regressionsmodell k -nächste Nachbarn nutzt und den latenten Raum iterativ aufbaut. UNN ist somit die Basis für das in dieser Arbeit entwickelte Verfahren und unterscheidet sich hauptsächlich durch die Verwendung des Regressionsmodells der k -nächsten Nachbarn.

Bei k -nächste Nachbarn handelt es sich um ein Klassifikationsverfahren, das ein zu klassifizierendes Muster anhand seiner k -nächsten Nachbarn klassifiziert [1]. Soll beispielsweise ein Muster $\mathbf{x}$ mit $k = 3$ klassifiziert werden und sind in seiner Menge der drei nächsten Nachbarn zwei Muster mit dem Label Wert A und eins mit dem Label Wert B , so wird klassischerweise eine Mehrheitsentscheidung getroffen und das Muster $\mathbf{x}$ wird mit dem Label A klassifiziert. Um nicht eine Klassifikation sondern eine Regression durchzuführen, kann das arithmetische Mittel über die Label $\mathbf{y}_i$ der k -nächsten Nachbarn gebildet werden [5]:

$$f_{\text{KNN}}(\mathbf{x}) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(\mathbf{x})} \mathbf{y}_i \quad (2.10)$$

In Formel 2.10 ist $\mathcal{N}_k$ die Indexmenge der k -nächsten Nachbarn. Zur Bestimmung der nächsten Nachbarn kann beispielsweise die euklidische Distanz genutzt werden. Es gibt auch Varianten, in denen die nächsten Nachbarn anti-proportional zu ihrem Abstand zum Muster $\mathbf{x}$ gewichtet werden [5].

2.4. Regressionsmodell

Das Grundprinzip, nach dem das in dieser Arbeit entwickelte Verfahren den latenten Raum iterativ aufbaut, gleicht dem Verfahren der unüberwachten nächsten Nachbarn. Die Muster aus dem Datenraum werden nacheinander im latenten Raum platziert und das auf der

Position, die zu einem minimalen Datenraumrekonstruktions-Fehler führt. Im Gegensatz zu dem Verfahren der unüberwachten nächsten Nachbarn kommt hier als Regressionsmodell der Nadaraya-Watson-Schätzer zum Einsatz. Da der Nadaraya-Watson-Schätzer somit ein zentraler Bestandteil dieser Arbeit ist, wird er in diesem Kapitel Schritt für Schritt entwickelt. Begonnen wird dabei mit einem klassischen Histogramm, welches in einem zweiten Schritt um Label-Informationen erweitert wird, wodurch sich eine Art diskreter Nadaraya-Watson-Schätzer ergibt. Mit Hilfe der Kernel-Funktion und des Parzen-Window-Schätzers, welcher eine Art kontinuierliches Histogramm ist, wird aus dem diskreten Nadaraya-Watson-Schätzer der Nadaraya-Watson-Schätzer. Dabei wird auch auf die Bandbreite eingegangen.

2.4.1. Histogramm

Ein Histogramm bietet die Möglichkeit, die Häufigkeitsverteilung vorhandener Daten zu visualisieren. Dabei werden die Daten in Klassen eingeteilt und die Klassen in Form von Balken mit der Größe entsprechend ihrer Häufigkeit dargestellt.

Tabelle 2.1.: Anzahl der Internetnutzer in 1000

	10 bis 15	16 bis 24	25 bis 44	45 bis 64	65 und älter	Insgesamt
Männlich	2277	4207	10478	9275	3102	29338
Weiblich	2169	4025	10263	8581	2056	27093
Insgesamt	4446	8232	20741	17856	5158	56431

Basierend auf Daten des Statistischen Bundesamts² über die Anzahl der Internetnutzer in Deutschland – siehe Tabelle 2.1 – wurde ein

²Statistisches Jahrbuch 2013 – 7 Kultur, Medien, Freizeit

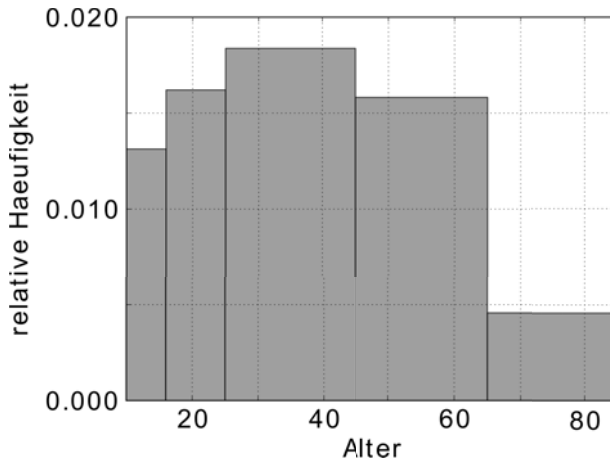


Abbildung 2.2.: Beispiel eines Histogramms

Histogramm als Beispiel erstellt. Abbildung 2.2 zeigt dabei auf der x -Achse das Alter der Nutzer und in Richtung der y -Achse die relative Häufigkeit an Menschen, die aktiv das Internet nutzen. Die Einteilung der Klassen wurde dabei vom statistischen Bundesamt übernommen. Man sieht an Abbildung 2.2, dass es möglich ist, die Klassen unterschiedlich breit einzuteilen. Die Höhe des Balkens, der die Häufigkeit seiner Klasse symbolisiert, ermittelt sich wie folgt:

$$h(\text{Klasse}) = \frac{1}{\text{Breite}_{\text{Klasse}}} \cdot \frac{N_{\text{Klasse}}}{N} \quad (2.11)$$

Wobei $\text{Breite}_{\text{Klasse}}$ die Breite der Klasse ist – zum Beispiel für die Klasse *16 bis 24* gilt $\text{Breite}_{\text{Klasse}} = 9$. Die Anzahl der Daten einer Klasse ist N_{Klasse} und N ist die Gesamtanzahl der Daten.

Um aus einem Histogramm ein Regressionsmodell zu entwickeln, werden den Daten X Label-Informationen Y hinzugefügt. Es wurde sich dafür entschieden, über das Label anzugeben, wie hoch die Wahrscheinlichkeit ist, dass ein Nutzer täglich im Internet aktiv ist, siehe Tabelle 2.2.

Ist der Nutzer täglich aktiv, so bekommt er das Label 1 (100%), da die Chance bei 100% liegt, dass er täglich im Internet aktiv ist.

Tabelle 2.2.: Anteil der Internetnutzer mit täglicher Aktivität

	10 bis 15	16 bis 24	25 bis 44	45 bis 64	65 und älter	Insgesamt
Männlich	61%	90%	84%	76%	69%	79%
Weiblich	61%	87%	80%	66%	47%	72%
Insgesamt	61%	89%	82%	71%	60%	76%

Ist er nicht täglich aktiv, so erhält er das Label 0 (0%). Über das arithmetische Mittel kann für die einzelnen Klassen die Gesamtwahrscheinlichkeit für die entsprechende Klasse ermittelt werden. Sie ergibt sich zu:

$$f(\text{Klasse}) = \sum_{i \in \mathcal{N}_{\text{Klasse}}} y_i \cdot \frac{1}{N_{\text{Klasse}}} \quad (2.12)$$

Die Anzahl der Daten einer Klasse ist N_{Klasse} , die Indexmenge der Elemente der Klasse ist $\mathcal{N}_{\text{Klasse}}$ und es gilt $y_i \in Y$.

Mit den Daten aus dem Datensatz des Statistischen Bundesamts ergibt sich Abbildung 2.3. Diese Abbildung ist etwas anders zu interpretieren als ein klassisches Histogramm. Die Höhe der Balken sagt nun etwas über die Klasse der Balken aus und ist nicht mehr eine Eigenschaft der Klasse relativ zum Gesamtdatensatz. So kann beispielsweise gesagt werden, dass 60% aller über 65-Jährigen täglich das Internet nutzen. Es handelt sich hierbei um eine Art diskreten Nadaraya-Watson-Schätzer.

2.4.2. Kernel-Funktion

Die Anwendung der Kernel-Funktion ist der entscheidende Schritt zum kontinuierlichen Nadaraya-Watson-Schätzer. Sie ersetzt das Rechteck, welches im klassischen Histogramm genutzt wird, durch

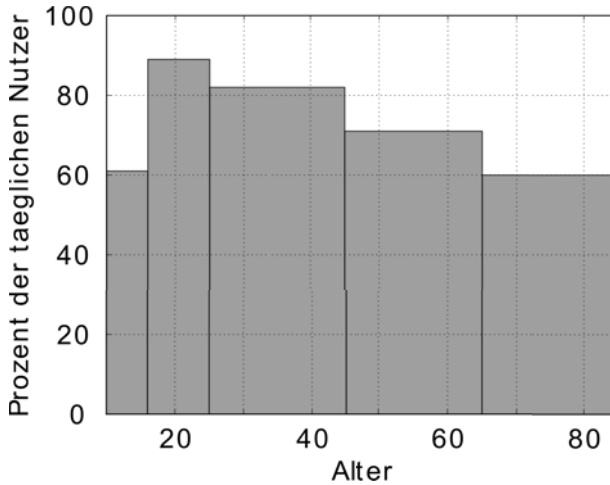


Abbildung 2.3.: Histogramm mit Label-Informationen

einen kontinuierlichen Verlauf. Hierdurch wird eine kontinuierliche Funktion mit Ableitung ohne Unstetigkeitsstellen möglich.

Es gibt eine große Anzahl an Kernel-Funktionen. Verbreitet sind zum Beispiel der Gauss-Kernel und der Epanechnikov-Kernel. Für Kernel-Funktionen gelten zwei Anforderungen [2]: Zum einen muss die Fläche unter der Funktion den Wert 1 ergeben, wie in Formel 2.13 beschrieben. Die Funktion muss zum anderen symmetrisch zur y -Achse sein, wie Formel 2.14 zeigt:

$$\int K(x)dx = 1 \quad (2.13)$$

$$\int x \cdot K(x)dx = 0 \quad (2.14)$$

Kernel-Funktionen können über den durchschnittlichen, integrierten, quadratischen Fehler (MISE: mean integrated squared error) bewertet werden [4]. Für diesen Fehler ist der Epanechnikov-Kernel optimal [22]. Daher wurde er für den ersten Ansatz ausgewählt und wird an dieser Stelle als Beispiel dargestellt. Abbildung 2.4 zeigt die Kernel-Funktion.

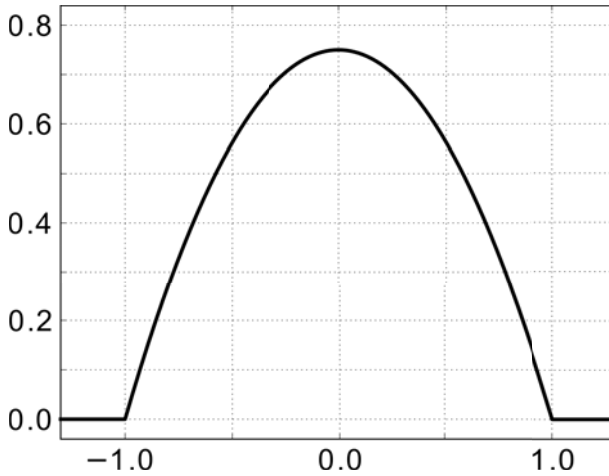


Abbildung 2.4.: Epanechnikov-Kernel-Funktion

$$K_{\text{Epa}}(x) = \begin{cases} \frac{3}{4} \cdot (1 - x^2) & \text{falls } |x| \leq 1 \\ 0 & \text{falls } |x| > 1 \end{cases} \quad (2.15)$$

$$K'_{\text{Epa}}(x) = \begin{cases} \frac{3}{4} \cdot (-2x) & \text{falls } |x| \leq 1 \\ 0 & \text{falls } |x| > 1 \end{cases} \quad (2.16)$$

Definiert ist der Epanechnikov-Kernel wie in Formel 2.15 zu sehen ist. Formel 2.16 macht deutlich, dass die Ableitung des Epanechnikov-Kernels nicht stetig ist. Die Unstetigkeitsstellen befinden sich an den Positionen -1 und 1 .

2.4.3. Parzen-Window-Schätzer

Wie einleitend beschrieben, stellt der Parzen-Window-Schätzer eine Art kontinuierliches Histogramm dar. Anstatt die Daten *in Rechtecken zu sammeln*, ist die Idee beim Parzen-Window-Schätzer, für

Hybride Optimierung für Dimensionsreduktion
Unüberwachte Regression mit Gradientenabstieg und
evolutionären Algorithmen

Lückehe, D.

2015, VIII, 99 S. 51 Abb., Softcover

ISBN: 978-3-658-10737-6