Datenanalyse auf Basis von KI-Methoden

Einfache Lineare Regression

Repräsentation der Punktwolke durch eine Gerade der allgemeinen Form:

$$Y = b0 + b1 * X$$

Dabei stehen:

- *y* für die abhängige Variable,
- *x* für die unabhängige Variable,
- *b*0 für den Schnittpunkt der Geraden mit der *y*-Achse des Koordinatensystems
- *b*1 für die Steigung der Geraden, auch Regressionskoeffizient genannt

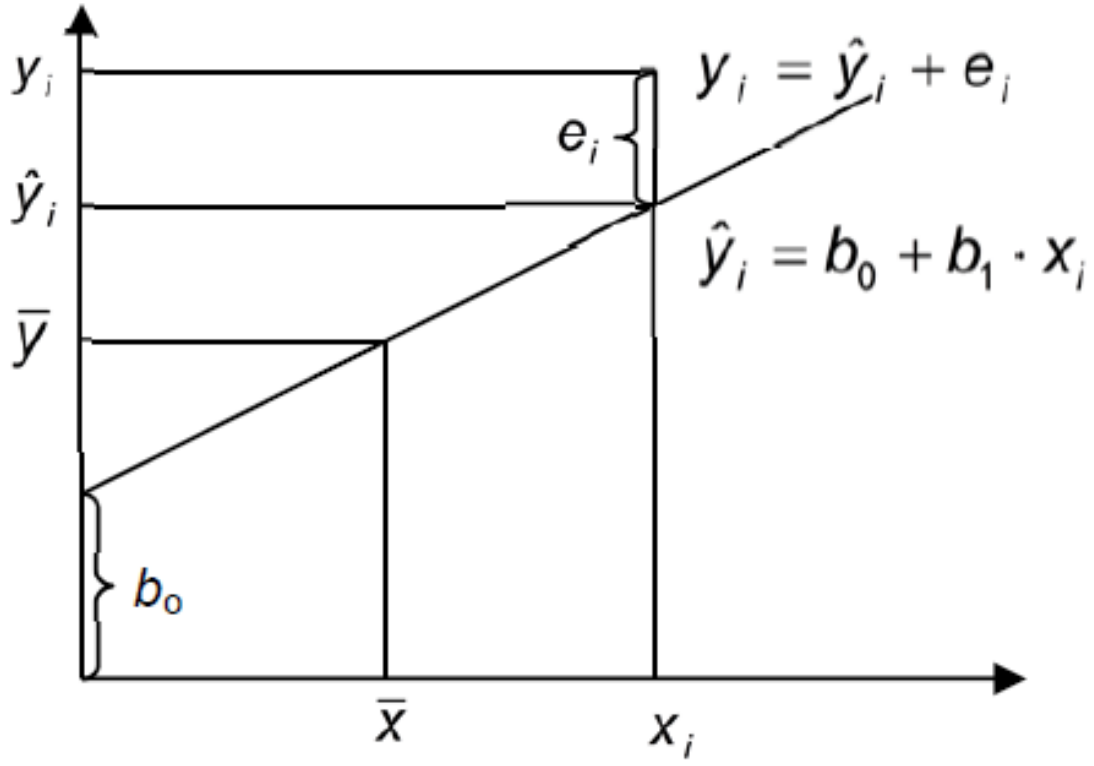
Regressionsgerade

- Zur Berechnung der Geraden werden in ein Koordinatensystem die Wertepaare übertragen und eine Punktwolke zeigen.

- Legt man nun rein graphisch irgendeine Gerade hinein, so sind stets Abweichungen der Einzelwerte *yi* von der Geraden festzustellen.

- Diese Abweichungen werden als Residuen *ei* bezeichnet.

Darstellung- Regressionsgerade



Regressionsgerade

- Damit das Datenmaterial durch die Regressionsgerade möglichst gut repräsentiert wird, muss die Abweichung der Einzelwerte yi von der Geraden minimiert werden.

- Ein Kriterium für die beste Anpassung der Regressionsgerade an die Beobachtungen muss gefunden werden.

- Methode der kleinsten Quadrate vorgestellt werden, die die Quadratsumme der Residuen minimiert.

Methode der kleinsten Quadrate

- Die Regressionsgerade ist diejenige Gerade, die die Summe der quadrierten Residuen (Abweichungen, Vorhersagefehler) minimiert.

Es gilt:

$$e_i = y_i - \hat{y}_i$$

$$e_i = y_i - (b_0 + b_1 \cdot x_i)$$

$$e_i^2 = [y - (b_0 + b_1 \cdot x_i)]^2$$

Gefordert ist:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y - (b_0 + b_1 \cdot x_i)]^2 \rightarrow Min$$

Methode der kleinsten Quadrate

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_j - \bar{x})^2} = \frac{\sum_{j=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i \right) \Big/ n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \Big/ n}$$

$$b_1 = \frac{Summe\ der\ Abweichungsprodukte_{xy}}{Summe\ der\ Abweichungsquadrate_{xy}} = \frac{SP_{xy}}{SQ_{xy}}$$

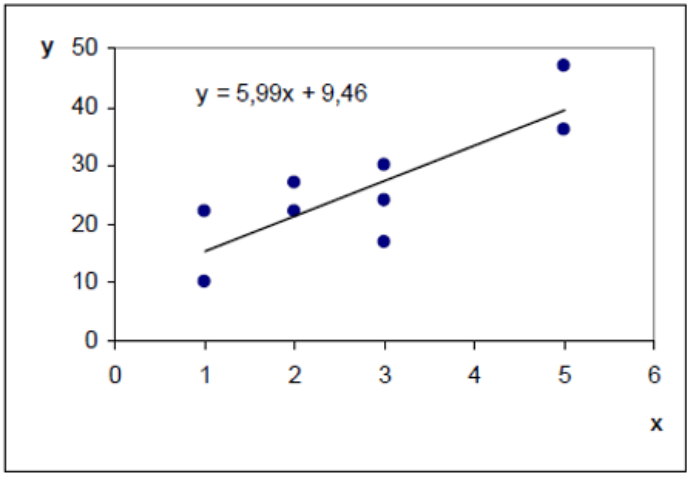
Beispiel

	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*
Schnittpunkt	9,4618	4,8596	1,9471	0,0926
X Variable 1	5,9937	1,5630	3,8347	0,0064

Die Regressionsgerade lautet damit:

$\hat{y} = b_0 + b_1 \cdot x = 9{,}4618 + 5{,}9937 \cdot x$

In Worten: Ändert sich die Einflussgröße x um eine Einheit, so ändert sich die Zielgröße y um 5,9937 Einheiten. Ist die Einflussgröße = 0, so beträgt der Wert der Zielgröße = 9,4618.



Anpassungsgüte

Den Anteil der durch die Regression erklärten Streuung an der Gesamtstreuung bezeichnet als **Bestimmtheitsmaß r^2**:

$$r^2 = \frac{SQ_{Reg}}{SQ_{Ges}} = \frac{b_1 \cdot \left[\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})\right]}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = b_1 \cdot \frac{SP_{xy}}{SQ_{xy}} = b_1^2 \cdot \frac{SQ_x}{SQ_y} \qquad ($$

Für das obige Beispiel folgt:

$$r^2 = \frac{5{,}9937 \cdot 105{,}2222}{930{,}8889} = 0{,}6775$$

Hochdimensionale Daten

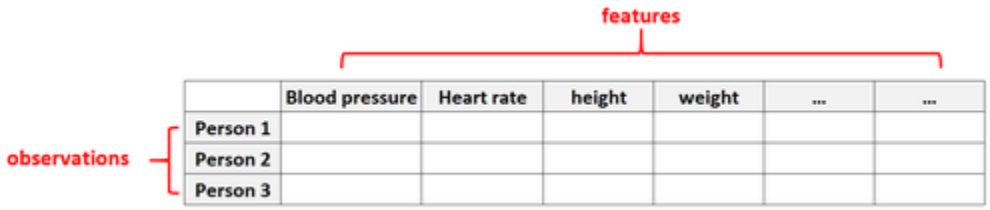
Definition

Hochdimensionale Daten zeichnen sich meist dadurch aus, dass die Anzahl der **Beobachtungen** (**n**) wesentlich kleiner ist als die Anzahl der **Variablen** (**p**) ist.

Kurz schreibt man dafür auch **n<<p**.

Beispielfelder sind:

- Gesundheitswesen
- Finanzwesen
- Bioinformatik
- ….



Regressionsanalyse: Hochdimensionale Daten

Probleme

Eine hohe Dimensionalität der Daten:

- erhöht die Komplexität der Modellen
- erhöht das Risiko einer Überanpassung (Overfitting).

Lösung

- Regularisierung

Regularisierung

Definition

Die Regularisierung ist eine Methode zur Verbesserung der Vorhersagegenauigkeit und der Modellinterpretierbarkeit.

Bei diesem Ansatz wird ein Modell mit allen unabhängigen Variablen geschätzt.
Die geschätzten Regressionskoeffizienten werden gegen Null geschrumpft.

Diese Schrumpfung hat den Effekt, dass die Varianz reduziert wird.

Je nachdem, welche Art von Schrumpfung durchgeführt wird, können einige der Koeffizienten als genau Null geschätzt werden.

Methoden zur Regularisierung

- Ridge Regression
- Lasso Regression
- Elastic Net Regression

Ridge Regression

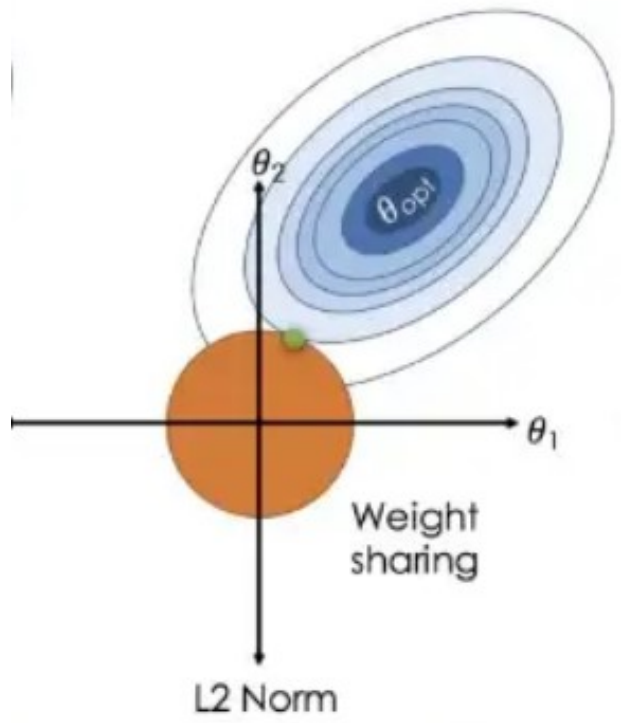
Bei der linearen Regression werden die Koeffizienten nach **der Methode der kleinsten Quadrate ausgewählt**, die die Summe der quadratischen Residuen (RSS) minimiert:

$$\sum \left(w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n - y\right)^2$$

Wenn die Prädiktorvariablen jedoch stark korreliert sind, kann **Multikollinearität** zu einem Problem werden.

Eine Möglichkeit, dieses Problem zu umgehen, ohne einige Prädiktorvariablen vollständig aus dem Modell zu entfernen, besteht darin, **Ridge-Regression** (L2-Penalty) zu verwenden, mit der stattdessen Folgendes minimiert werden soll:

$$\sum \left(w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n - y\right)^2 + \lambda \sum w_i^2$$



Quelle: Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net.

Ridge Regression

Schritte

1. Berechnung der Korrelationsmatrix und die VIF-Werte für die Prädiktorvariablen.
2. Standartisierung jeder Prädiktorvariable.
3. Anpassung des Ridge-Regressionsmodells und Auswahl des Werts für λ.

Vorteile

Fähigkeit, einen niedrigeren mittleren quadratischen Testfehler (MSE) im Vergleich zur Regression der kleinsten Quadrate zu erzeugen, wenn **Multikollinearität** vorliegt.

Nachteile

Unfähigkeit, eine **Variablenauswahl** durchzuführen, da alle Prädiktorvariablen im endgültigen Modell enthalten sind. Da einige Prädiktoren sehr nahe an Null geschrumpft werden, kann es schwierig sein, die Ergebnisse des Modells zu interpretieren.

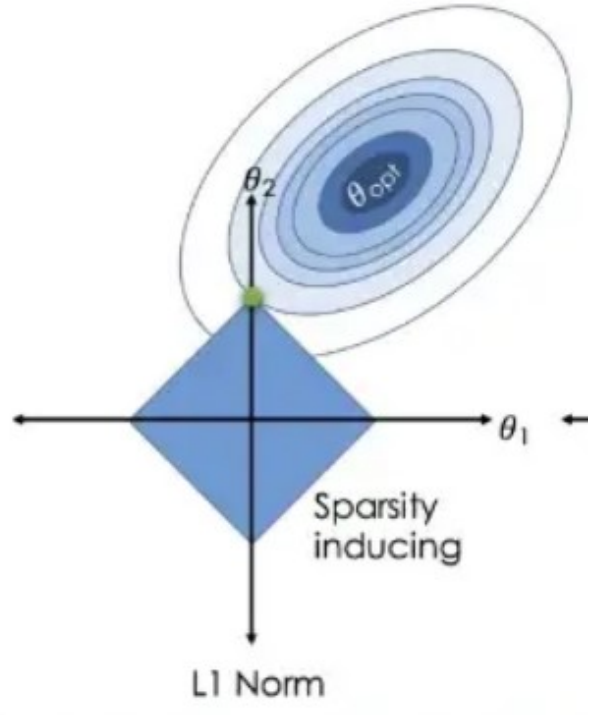
Lasso Regression

Lasso (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) ist konzeptionell der Ridge-Regression sehr ähnlich.

Sie fügt ebenfalls eine Bestraffung für Nicht-Null-Koeffizienten hinzu (L1-Penalty).

$$\sum \left(w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n - y\right)^2 + \lambda \sum |w_i|$$

Dies hat zur Folge, dass bei hohen Werten von λ viele Koeffizienten unter Lasso **genau auf Null** gesetzt werden, was bei der Ridge-Regression nie der Fall.



Quelle: Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net.

Lasso Regression

Schritte

1. Berechnung der Korrelationsmatrix und die VIF-Werte für die Prädiktorvariablen.
2. Anpassung des Lasso-Regressionsmodells und Auswahl des Werts für λ.

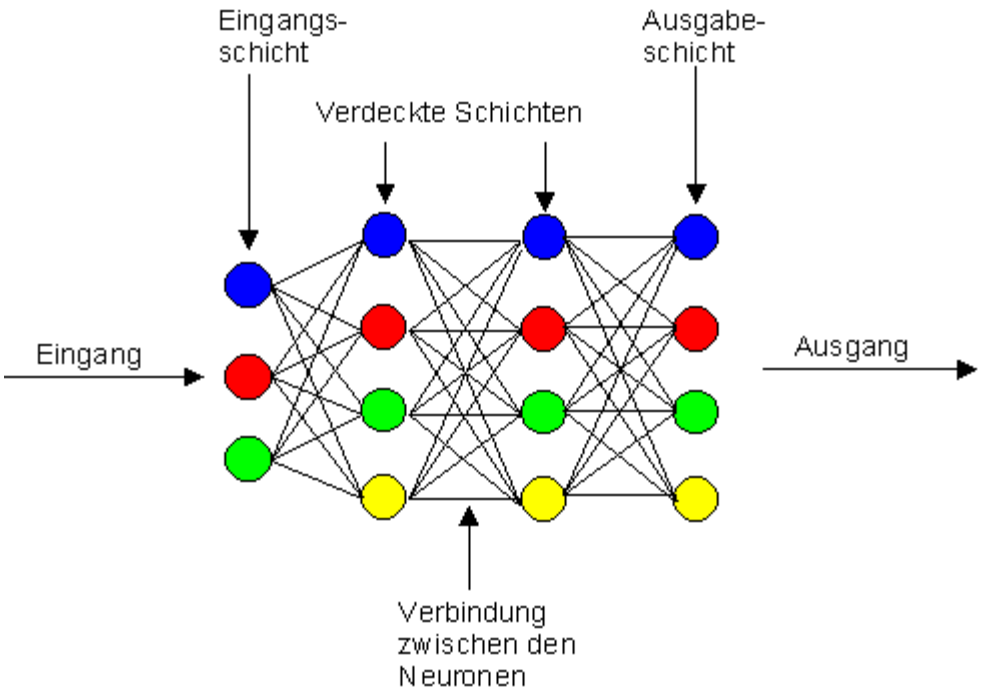
Vorteile

Erstellung von parametersparsame Modell, die somit eine verbesserte **Interpretiebarkeit** und **Prädiktionsfähigkeit** aufweisen.

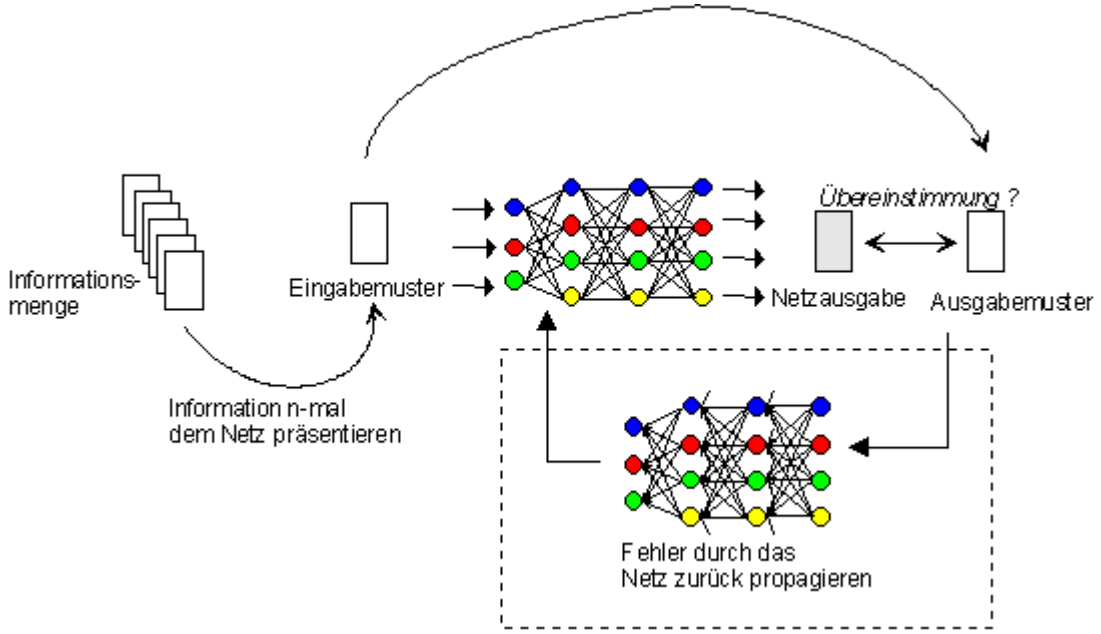
Nachteile

Bei gruppierten Variablen beziehungsweise **stark korrelierten Variablen** tendiert Lasso dazu, aus einer Gruppe **eine beliebige Variable zu wählen** und ignoriert die anderen Variablen der Gruppe.

Deep Learning



Deep Learning



Vielen Dank

Für Ihre

Aufmerksamkeit!