How To: Lineare Regression, Polynomregression

Hier wird die Methode vorgestellt, eine Linearkombination beliebiger Funktionen $f_j(x) \in \mathbb{R}$ so auszuwählen, dass sie eine Menge von Punkten möglichst genau annähert (Fitfunktion).

Eine solche **Lineare Regression** eignet sich demzufolge insbesondere für **Polynomregressionen**. Verwendet wird die Gaußsche Methode, welche das mittlere **Fehlerquadrat** minimiert. Weiter unten befindet sich ein Beispiel.

Zunächst einigen wir uns auf die zu verwendenden Größen und Indexkonventionen. Gegeben sind:

- n Wertepaare (x_i, y_i)
- m Funktionen $f_\mathbf{j}$, jeweils mit zu ermittelnden Koeffizienten $a_\mathbf{j}$

Die Punkte sollen nun möglichst genau durch die Gesamtfunktion $F(x)$ beschrieben werden. Diese kann, je nach Bedürfnis, ein **Polynom** beliebigen Grades sein, oder eine Linearkombination anderer Funktionen.

$$F(x) = \sum_{\mathbf{j}=1}^{m} a_\mathbf{j} f_\mathbf{j}(x) \tag{1}$$

Wir versuchen, die Koeffizienten a_j so zu wählen, dass dies klappt. Der Trick liegt darin, die Standardabweichung σ zu minimieren.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[F(x_i) - y_i\right]^2} \tag{2}$$

Hierbei genügt es bereits, den Wert der Summe zu minimieren, denn die Wurzel und der konstante Faktor $1/n$ verändern die Extremaleigenschaften nicht.

Da der Wert der Summe ausschließlich über die Koeffizienten a_j variiert wird, müssen die Ableitungen nach den Koeffizienten Null ergeben. Nur in diesem Fall liegt ein Extremum vor.

$$\frac{\partial}{\partial a_\mathbf{k}} \left[\sum_{i=1}^{n} \left(\sum_{\mathbf{j}=1}^{m} a_\mathbf{j} f_\mathbf{j}(x_i) - y_i \right)^2 \right] = 0 \qquad \text{mit } \mathbf{k} = 1, 2, \ldots, m \tag{3}$$

Führen wir innere und äußere Ableitung aus, so ergibt sich das folgende lineare Gleichungssystem:

$$\sum_{\mathbf{j}=1}^{m} \sum_{i=1}^{n} a_\mathbf{j} f_\mathbf{j}(x_i) f_\mathbf{k}(x_i) = \sum_{i=1}^{n} y_i f_\mathbf{k}(x_i) \tag{4}$$

Das sieht zwar ganz furchtbar aus, ist es aber gar nicht. Es handelt sich um m lineare Gleichungen ($k = 1 \ldots m$), die nach den Unbekannten a_j aufgelöst werden müssen.

Dies geschieht ganz einfach mit Hilfe einer Erweiterten Koeffizientenmatrix. Diese hat

- $m + 1$ Spalten ($j = 1 \ldots m$ links sowie die Konstantenspalte rechts)
- m Zeilen ($k = 1 \ldots m$)

Unsere Matrix sieht für $m = 3$ Funktionen konkret folgendermaßen aus:

$$\left(\begin{array}{ccc|c} \sum_i f_{\mathbf{1}}(x_i)f_{\mathbf{1}}(x_i) & \sum_i f_{\mathbf{2}}(x_i)f_{\mathbf{1}}(x_i) & \sum_i f_{\mathbf{3}}(x_i)f_{\mathbf{1}}(x_i) & \sum_i y_i f_{\mathbf{1}}(x_i) \\ \sum_i f_{\mathbf{1}}(x_i)f_{\mathbf{2}}(x_i) & \sum_i f_{\mathbf{2}}(x_i)f_{\mathbf{2}}(x_i) & \sum_i f_{\mathbf{3}}(x_i)f_{\mathbf{2}}(x_i) & \sum_i y_i f_{\mathbf{2}}(x_i) \\ \sum_i f_{\mathbf{1}}(x_i)f_{\mathbf{3}}(x_i) & \sum_i f_{\mathbf{2}}(x_i)f_{\mathbf{3}}(x_i) & \sum_i f_{\mathbf{3}}(x_i)f_{\mathbf{3}}(x_i) & \sum_i y_i f_{\mathbf{3}}(x_i) \end{array}\right) \tag{5}$$

Jedes Matrixelement besteht aus einer Summe über alle gegebenen Wertepaare (x_i, y_i). Mittels **Gauß-Jordan-Verfahren** bringen wir diese Matrix nun auf Diagonalform und erhalten die gesuchten Koeffizienten a_j.

$$\left(\begin{array}{ccc|c} \mathbf{1} & 0 & 0 & a_{\mathbf{1}} \\ 0 & \mathbf{1} & 0 & a_{\mathbf{2}} \\ 0 & 0 & \mathbf{1} & a_{\mathbf{3}} \end{array}\right) \tag{6}$$

Beispiel

Wir wollen die Punkte

$$\begin{aligned} (x_{\mathbf{1}}, y_{\mathbf{1}}) &= (9, 15) \\ (x_{\mathbf{2}}, y_{\mathbf{2}}) &= (18, 24) \end{aligned} \tag{7}$$

durch eine Funktion

$$F(x) = a_{\mathbf{1}}x + a_{\mathbf{2}}\sin(x) + a_{\mathbf{3}} \tag{8}$$

beschreiben. Das heißt, wir haben diese drei Teilfunktionen:

$$\begin{aligned} f_{\mathbf{1}}(x) &= x \\ f_{\mathbf{2}}(x) &= \sin(x) \\ f_{\mathbf{3}}(x) &= 1 \end{aligned} \tag{9}$$

Unsere **Koeffizientenmatrix** sieht deshalb folgendermaßen aus. Beim Berechnen des Sinus aufpassen: wir rechnen in Radianten, nicht in Grad!

$$\left(\begin{array}{ccc|c} 9^2+18^2 & \sin(9)\cdot 9+\sin(18)\cdot 18 & 9+18 & 15\cdot 9+24\cdot 18 \\ 9\sin(9)+18\sin(18) & \sin^2(9)+\sin^2(18) & \sin(9)+\sin(18) & 15\sin(9)+24\sin(18) \\ 9+18 & \sin(9)+\sin(18) & 1+1 & 15+24 \end{array}\right)$$

$$= \left(\begin{array}{ccc|c} 405 & -9.8087 & 27 & 567 \\ -9.8087 & 0.7338 & -0.3389 & -11.8419 \\ 27 & -0.3389 & 2 & 39 \end{array}\right) \tag{10}$$

Diese Matrix muss nun mittels **Gauß-Jordan-Verfahren** auf **Diagonalform** gebracht werden.

405	−9.8087	27	567	$\mid$: **405**
−9.8087	0.7338	−0.3389	−11.8419	
27	−0.3389	2	39	
1	−0.0242	0.0667	1.4	
−9.8087	0.7338	−0.3389	−11.8419	$\mid$ + **9.8087** · Zeile 1
27	−0.3389	2	39	$\mid$ − **27** · Zeile 1
1	−0.0242	0.0667	1.4	
0	0.4964	0.3153	1.8903	$\mid$: **0.4964**
0	0.3145	0.1991	1.2	
1	−0.0242	0.0667	1.4	$\mid$ + **0.0242** · Zeile 2
0	**1**	0.6352	3.8080	
0	0.3145	0.1991	1.2	$\mid$ − **0.3145** · Zeile 2
1	0	0.0821	1.4921	
0	**1**	0.6352	3.8080	
0	0	−0.0007	0.0024	$\mid$: **−0.0007**
1	0	0.0821	1.4921	$\mid$ − **0.0821** · Zeile 3
0	**1**	0.6352	3.8080	$\mid$ − **0.6352** · Zeile 3
0	0	**1**	−3.4286	
1	0	0	**1.7739**	
0	**1**	0	**5.9858**	
0	0	**1**	**−3.4286**	

Damit haben wir die gesuchten Koeffizienten gefunden und können die Fitfunktion $F(x)$ aufstellen.

$a_{\mathbf{1}} = 1.7739$
$a_{\mathbf{2}} = 5.9858$
$a_{\mathbf{3}} = -3.4286$

$\Rightarrow F(x) = 1.7739x + 5.9858\sin(x) - 3.4286$

Als graphische Darstellung sieht unser Ergebnis nun folgendermaßen aus.