Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Fachbereich Informatik
Dr. Marco Hülsmann

Numerische Mathematik 1
Übungsblatt 12, WS 2019/20
– Musterlösungen –

Aufgabe 1 (Regressionsgerade)

Gegeben seien die Punkte $(x_i, y_i) \in \mathbb{R}^2$ für $i = 1, ..., n$, $n \in \mathbb{N}$. Die *Regressionsgerade* ist diejenige Gerade

$$g(x) = ax + b$$

welche die Abstände $||y_i - g(x_i)||$ für alle $i = 1, ..., n$ minimiert. Zeigen Sie, daß die Koeffizienten der Regressionsgeraden gegeben sind durch

$$\begin{aligned} b &= \bar{y} - a\bar{x}, \\ a &= \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \end{aligned}$$

Dabei sind $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i$ und $\bar{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n y_i$ die Mittelwerte der Daten, $\mathrm{Cov}(x,y)$ die *Kovarianz* und $\mathrm{Var}(x)$ deren *Varianz*.

Musterlösung:

Zu minimieren ist die Fehlerfunktion

$$E(a,b) := \sum_{i=1}^n (y_i - (ax_i + b))^2$$

Die binomischen Formeln liefern

$$\begin{aligned} E(a,b) &= \sum_{i=1}^n \left(y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2\right) \\ &= \sum_{i=1}^n \left(y_i^2 - 2ax_iy_i - 2by_i + a^2x_i^2 + 2abx_i + b^2\right) \end{aligned}$$

Setze nun den Gradienten von E gleich Null:

$$\frac{\partial E}{\partial b}(a,b) = \sum_{i=1}^{n}(-2y_i + 2ax_i + 2b) = 0$$

Dies ist genau dann der Fall, wenn

$$\sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b - \sum_{i=1}^{n} y_i = 0$$

also wenn

$$nb = n\bar{y} - an\bar{x}$$

bzw.

$$b = \bar{y} - a\bar{x}$$

Weiterhin gilt

$$\begin{aligned}
\frac{\partial E}{\partial a}(a,b) &= \sum_{i=1}^{n}\left(-2x_i y_i + 2ax_i^2 + 2bx_i\right) = 0 \\
\Leftrightarrow\ 0 &= \sum_{i=1}^{n}(ax_i^2 + x_i(\bar{y} - a\bar{x}) - x_i y_i) \\
\Leftrightarrow\ 0 &= \sum_{i=1}^{n}(ax_i^2 - a\bar{x}x_i + x_i\bar{y} - x_i y_i)
\end{aligned}$$

Dies ist genau dann der Fall, wenn

$$\sum_{i=1}^{n}(ax_i^2 - a\bar{x}x_i) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y}$$

bzw.

$$a\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i) = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i$$

Löse nach a auf:

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i)} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}n\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Multipliziert man Zähler und Nenner mit $\frac{1}{n}$, so folgt die Behauptung.

Aufgabe 2 (Plotten einer Regressionsgerade mit Octave)

Ein überbestimmtes lineares Gleichungssystem $A\theta = b$ mit $A \in \mathbb{R}^{m\times n}$, wobei $m > n$, $\theta \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, kann in das quadratisches System der Form

$$A^T A\theta = A^T b$$

überführt werden. Die Lösung θ^* minimiert den Abstand $||A\theta - b||$. Schreiben Sie ein *octave*-Programm, welches dieses Gleichungssystem löst, wobei $A = (e\ x), e = (1, ..., 1)^T \in \mathbb{R}^m$, $b = y$ (x und y seien die x- bzw. y-Koordinaten der betrachteten Datenpunkte). Benutzen Sie dabei die Funktion `pinv` zur Berechnung der sog. *Pseudoinversen* von $A^T A$. Plotten Sie anschließend die Regressionsgerade, deren Koeffizienten durch den Lösungsvektor θ^* gegeben sind. Hinweis: Der Befehl `hold on` sorgt dafür, daß ein Plot beibehalten wird und etwas in den Plot hinzugefügt werden kann (ansonsten wird ein komplett neuer Plot erzeugt).

Musterlösung:

Das folgende *octave*-Programm liefert die gewünschte Regressionsgerade für zufällig gewählte y-Werte:

```
>> x = 1:10
x =

    1    2    3    4    5    6    7    8    9   10

>> y = rand(10)(1,:)
y =

   0.305264   0.120211   0.498296   0.210333   0.155803
   0.525931   0.025622   0.475793   0.286737   0.811778


>> plot(x,y,'rx')
>> X = [ones(length(x), 1) x']
X =

    1    1
    1    2
    1    3
    1    4
    1    5
    1    6
    1    7
    1    8
    1    9
    1   10
```

```
>> theta = (pinv(X'*X))*X'*y'
theta =

   0.160651
   0.032896

>> hold on
>> legend('Training data', 'Linear regression')
>> plot(X(:,2), X*theta, '-')
>> hold off
```