Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Matthias Schubert
Gregor Jossé

München, 13.06.2016

Managing Massive Multiplayer Online Games
SS 2016

Übungsblatt 7: Knowledge Discovery and Data Mining II

Besprechung: 16.06.2016

Aufgabe 7-1 *Lineare Regression*

Der Mietpreis y_i einer Wohnung i hängt stark von der Quadratmeterzahl x_i ab. Andere Faktoren haben natürlich auch Einfluss, aber vereinfachend kann der Zusammenhang zwischen Mietpreis und Größe über ein lineares Regressionsmodell dargestellt werden, d.h.

$$y_i = w_0 + w_1 x_i$$

Als Trainingsmenge stehen folgende Daten zur Verfügung.

Wohnfläche in m^2	Kaltmiete in Euro
30	600
60	966
100	1640
55	992
93	1790
195	2925
21	469
61	840
62	1400

(a) Berechnen Sie die Regressionsgerade, die den mittleren quadratischen Fehler MSE zwischen den Mietpreis-Vorhersagen $\hat{y}_i$ und den tatsächlichen Preisen y_i minimiert

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Wir suchen w_0, w_1, sodass $f(w_0, w_1) = \sum_{i=1}^{n}(\hat{y}_i - (w_0 + w_1 x_i))^2$ minimal wird.

$$\frac{\partial f}{\partial w_0}\sum_{i=1}^{n}(\hat{y}_i - (w_0 + w_1 x_i))^2 = -2\sum(y_i - (w_0 + w_1 x_i))$$

$$\frac{\partial f}{\partial w_1}\sum_{i=1}^{n}(\hat{y}_i - (w_0 + w_1 x_i))^2 = -2\sum x_i(y_i - (w_0 + w_1 x_i))$$

Erstere Gleichung lässt sich leicht berechnen:

$$\sum(y_i - (w_0 + w_1 x_i)) = 0 \Leftrightarrow -nw_0 + \sum y_i - w_1 \sum x_i = 0$$
$$\Leftrightarrow w_0 = \bar{y} - w_1 \bar{x}$$

Zweitere Gleichung lässt sich nur umständlich über ein System von Normalengleichungen unter Verwendung geometrischer Eigenschaften (Hesse-Matrix) berechnen.
Die optimalen Gewichte w_0 und w_1 lassen sich deutlich einfacher über die Matrix-Darstellung berechnen. Dazu sei $w = (w_0, w_1)^T$, y der Output-Vektor der Trainingsdaten (Mietpreis) und x der um eine konstante 1-Spalte erweiterte Input-Vektor (Wohnfläche):

$$\begin{pmatrix} 1 & 30 \\ 1 & 60 \\ \vdots & \\ 1 & 62 \end{pmatrix}$$

also eine 9×2 Matrix (die Einsen werden für die Multiplikation mit w_0 benötigt: $y = 1w_0 + xw_1$). Dann lässt sich die Fehlerfunktion als Skalarprodukt schreiben: $f(w) = (y - Xw)^T(y - Xw)$. Es gilt:

$$\frac{\partial f(w)}{w} = -2X^T(y - Xw)$$

da nach Matrixableitung $\frac{\partial Ax}{x} = A^T$.
Aufgelöst nach w:

$$w = (X^T X)^{-1} X^T y$$

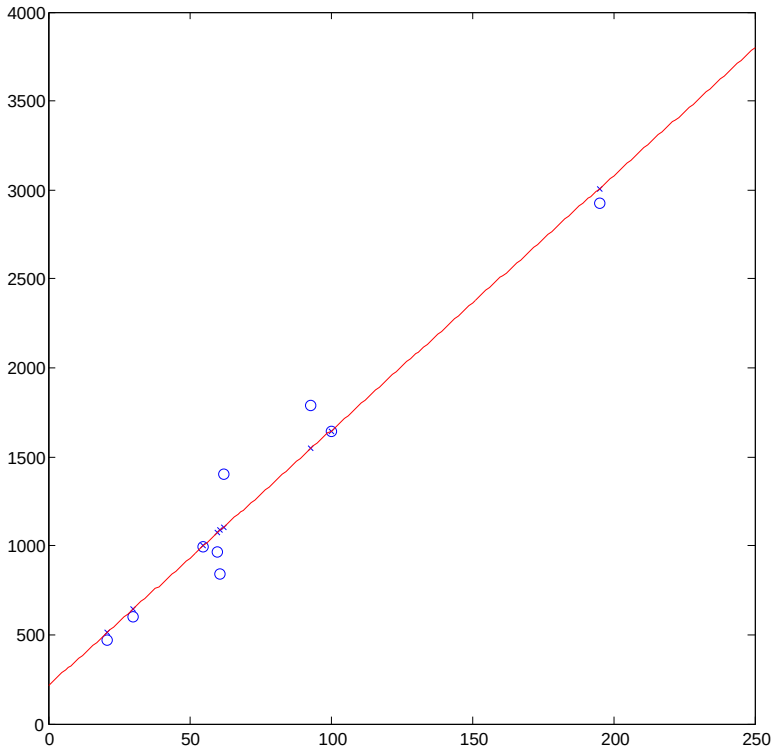
Falls $X^T X$ nicht invertierbar, addiere $\varepsilon > 0$ auf X. Also ergibt sich hier $w_0 = 213,10$ und $w_1 = 14,33$.

(b) Bestimmen Sie den quadratischen Fehler, um abzuschätzen wie gut das Modell den Zusammenhang beschreibt.

$$\sum(\hat{y}_i - y_i)^2 = 232051$$

MSE: 25783. MAE: 120.

(c) Berechnen Sie mithilfe der Regressionsgeraden den vorraussichtlichen Preis für eine Wohnung mit 120 Quadratmetern Fläche.
$\hat{y} = w_0 + w_1 120 = 1932$.


4000
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250