Maschinelles Lernen
Übungsblatt 2

Dr. Steffen Rendle, Christoph Freudenthaler
Wirtschaftsinformatik und Maschinelles Lernen (ISMLL)
Universität Hildesheim

3. November 2010
Abgabe bis 19. November

Aufgabe 1: Lineare Regression (4 Punkte)

a)

Seien die folgenden Datenreihen zu Außentemperaturen bzw. zum Gasverbrauch von Haushalten bekannt: $D = \{(2, 4), (6, 12.6), (8, 18.2), (4, 9.5)\}$. Berechnen Sie die Zielvariable Y für $x = 10$ mit der Methode der kleinsten Quadrate. Der tatsächliche Wert sei $y = 20$. Berechnen Sie den Fehler. Interpretieren Sie das Ergebnis. Erstellen Sie eine Grafik mit allen Dateninstanzen und zeichnen Sie für jeden Datenpunkt den quadratischen Fehler ein.

b)

Welche Annahmen unterstellt das lineare Modell den Daten? Welche Abweichungen der Daten vom linearen Modell können auftreten?

Aufgabe 2: Lineare Regression mit R (6 Punkte)

a)

Erzeugen Sie in R 2 Datensätze unterschiedlicher Größe aus dem linearen Modell $Y = \beta_0 + \beta_1 X + \epsilon$, wobei die Modellparameter folgende Werte haben: $\beta_0 = 0$, $\beta_1 = 2$, $\epsilon \sim N(0, 1)$ und die erklärende Variable X gleichverteilt aus dem Intervall $[0, 1]$ stammen soll. Datensatz 1 hat einen Umfang $n = 5$ während Datensatz 2 einen Umfang von $n = 10000$ hat.

b)

Erstellen Sie je ein lineares Regressionsmodell mit und ohne Biasparameter β_0 für die simulierten Daten aus a). Plotten Sie jeden Datensatz, lesen Sie pro Datensatz die Koeffizienten der beiden Modelle aus und fügen Sie jedem Datensatz-Plot die 3 linearen Regressionsgeraden der beiden geschätzten und des wahren Modells hinzu. Geben Sie den verwendeten R-Code an. (Hinweis: In den Kapiteln 11 (lineare Modelle) und 12 (Grafiken) der Einführung `http://cran.r-project.org/doc/manuals/R-intro.pdf` finden Sie weitere Informationen).

c)

Berechnen Sie die Abweichung der in den 4 Modellen berechneten Parameterschätzwerte $\hat{\beta}_1$ vom wahren Parameterwert $\beta_1 = 2$. Erklären Sie die verschiedenen Abweichungen.

d)

Geben Sie die Konfidenzintervalle für die Parameterschätzungen $\hat{\beta}_1$ an. Inwiefern stehen die Konfidenzintervalle im Zusammenhang mit der Größe des verwendeten Datensatzes?

Aufgabe 3: Nichtlineare Regression mit R (5 Punkte)

a)

Erzeugen Sie in R einen dritten Datensatz vom Umfang $n = 10000$ nach dem folgenden nicht-linearen Modell $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$, wobei die Modellparameter folgende Werte haben: $\beta_0 = 0$, $\beta_1 = 0$, $\beta_2 = 1$, $\epsilon \sim N(0, 1)$ und die erklärende Variable X gleichverteilt aus dem Intervall $[0, 10]$ stammen soll.

b)

Berechnen Sie ein lineares Regressionsmodell ohne Biasparameter $\beta_0 = 0$ und ohne linearem Parameter $\beta_2 = 0$ für die simulierten Daten aus a). Plotten Sie den nichtlinearen Datensatz, lesen Sie die Koeffizienten des linearen Modells aus und fügen Sie dem Plot die Regressionslinie hinzu. Berechnen Sie als nächstes die Residuen des Modells und plotten sie gegen x. Wie sollte der Plot aussehen und warum ist es mit diesem Modell nicht der Fall?