Maschinelles Lernen
Übungsblatt 3

Prof. Dr. Dr. Lars Schmidt-Thieme, Zeno Gantner
Wirtschaftsinformatik und Maschinelles Lernen (ISMLL)
Universität Hildesheim

18. November 2009
Abgabe bis 25. November

Aufgabe 1: Multiple Lineare Regression (5 Punkte)

Auf einer Website, die DVD-Bewertungen sammelt und mit den gesammelten Bewertungen ihren Benutzern neue DVDs empfiehlt, haben unter anderem zwei Benutzer die folgenden Bewertungen (1 Stern ist die schlechteste, 5 Sterne die beste) abgegeben:

Index	**Benutzer**	**Film**	**Bewertung**
1	A	*The Big Lebowski*	4 Sterne
2	A	*Brazil*	2 Sterne
3	A	*Titanic*	5 Sterne
4	B	*Brazil*	3 Sterne
5	B	*The Godfather*	4 Sterne
6	B	*Toy Story*	4 Sterne

Drei verschiedene Vorhersageverfahren würden anhand der restlichen gesammelten Bewertungen folgende Vorhersagen treffen:

Index	$\hat{r}_s$	$\hat{r}_r$	$\hat{r}_k$
1	3.7	3.8	3.9
2	2.4	2.5	2.3
3	2.2	3.0	4.1
4	3.2	3.1	2.9
5	4.7	4.4	4.2
6	4.1	3.9	4.2

a)

Berechnen Sie für jedes Verfahren den durchschnittlichen absoluten und den durchschnittlichen quadratischen Fehler im Vergleich zu den tatsächlichen Bewertungen.

b)

Ein Modell für die Kombination der ersten beiden Verfahren ist

$$r(x) = \beta_0 + \beta_1 \cdot \hat{r}_s(x) + \beta_2 \cdot \hat{r}_r(x) + \epsilon$$

Berechnen Sie die Schätzungen $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ mit Hilfe der in der Vorlesung vorgestellten Methode. Verwenden Sie zur Lösung des auftretenden linearen Gleichungssystems das Gaußsche Eliminationsverfahren.

Geben Sie die Gleichungen des zu lösenden Systems explizit an. Geben Sie die Zwischenschritte (gerundet auf zwei Nachkommastellen) in Matrixschreibweise an.

Hinweise:

- Sie können die Matrixmultiplikationen und Zeilenoperationen mit Rechnerunterstützung (bspw. in R) durchführen.
- Sie können Ihr Ergebnis mit einem Solver für lineare Gleichungssysteme, z.B. der `solve()`-Funktion in R, überprüfen.

c)

Berechnen Sie für das kombinierte Verfahren die Residuenquadratsumme, den durchschnittlichen absoluten und den durchschnittlichen quadratischen Fehler zu den bekannten Daten. Wie aussagekräftig sind die so berechneten Fehlermaße? Begründen Sie!

d)

Berechnen Sie die kombinierte Vorhersage für $\hat{r}_s(x) = 3.0$ und $\hat{r}_r(x) = 4.6$. Welcher negative Umstand fällt Ihnen dabei auf? Was passt nicht mit den Parametern $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$? Wie könnte man solche Ergebnisse verhindern?

Aufgabe 2: Variablenselektion (3 Punkte)

a)

Warum ist bei linearen Regressionsmodellen der Betrag der Koeffizienten kein gutes Maß für den Einfluss einer Variable?

b)

Was ist der Hauptunterschied zwischen Akaikes Informationskriterium (AIC) und dem Bayes-Schwarz-Informationskriterium (BIC)?

c)

Berechnen Sie für die Daten aus Aufgabe 1 ein lineares Regressionsmodell, das alle drei Vorhersageverfahren kombiniert (Darstellung der Zwischenschritte ist nicht notwendig). Berechnen Sie AIC und BIC für dieses Modell und das Modell aus Aufgabe 1. Welches Modell ist nach diesen Kriterien vorzuziehen?

Aufgabe 3: R (2 Punkte)

Lesen Sie Kapitel 4 und 5 von „An Introduction to R“.

a)

Was sind „Faktoren“ in R, wie werden sie erzeugt und wie können sie eingesetzt werden?

b)

Was ist der Unterschied zwischen einem Array und einem Vektor in R? Nennen Sie jeweils drei Operationen auf Arrays und Matrizen in R.