Maschinelles Lernen
Übungsblatt 4

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

17. November 2010
Abgabe bis 26. November

Aufgabe 1: Logistische Regression

Das logistische Regressionsmodell ist wiefolgt definiert:

$$\log\left(\frac{p(Y=1|x,\beta)}{p(Y=0|x,\beta)}\right) = x^T\beta$$

a)

Worin liegt der Unterschied zur klassischen Regression. Warum ist diese Anpassung/Erweiterung sinnvoll?

b)

Zur Berechnung der Wahrscheinlichkeit $p(Y = 1|x,\beta)$ wird im logistischen Regressionsmodell folgende Formel verwendet:

$$p(Y=1|x,\beta) = \sigma(x^T\beta) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$$

Zeigen Sie zum einen, dass

$$\frac{e^{x^T\beta}}{1+e^{x^T\beta}} = \frac{1}{1+e^{-x^T\beta}}$$

ist und zum anderen, dass diese Vorhersagefunktion aus der Modellgleichung des logistischen Modells folgt.

Hinweis: Erinnern Sie sich an die Modellgleichung des logistischen Modells und berücksichtigen Sie, dass $p(Y = 1|x,\beta) + p(Y = 0|x,\beta) = 1$ ist.

c)

Beweisen Sie die in der Vorlesung als Zwischenschritt zum Beweis des IRLS-Verfahrens (Folie 18) verwendete Identität:

$$\frac{\partial\sigma(x^T\beta)}{\partial\beta} = \sigma(x^T\beta)(1-\sigma(x^T\beta))x$$

Hinweis: Nutzen Sie den Umstand, dass $\sigma(x^T\beta) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$ ist.

Aufgabe 2: IRLS

Es seien folgende Daten gegeben:

y	x	y	x	y	x	y	x
0	9.5	0	10.5	1	11.1	1	11.9
0	9.6	0	11.0	1	11.1	1	12.1
0	9.7	0	11.2	1	11.1	1	12.2
0	9.8	0	11.5	1	11.5	1	12.5
0	9.9	0	11.7	1	11.8	1	12.6
0	12.1	-	-	1	12.6	-	-

a)

Berechnen Sie die lineare Regression $y = \beta_0 + \beta_1 x$ für die Zielvariable y (in R, Excel,...). Berechnen Sie den durchschnittlichen quadratischen Fehler des Modells auf den gegebenen Daten.

b)

Berechnen Sie ein logistisches Regressionsmodell $logit(y) = \beta_0 + \beta_1 x$ für die Zielvariable y. Wenden Sie (auf Papier, die auftretenden Gleichungssystem können Sie mit dem Rechner lösen) hierzu den Algorithmus *Iteratively Reweighted Least Squares* an. Stoppen Sie nach der zweiten Iteration und vergleichen Sie die Parameterschätzer nach Iteration 1 und 2 mit den endgültigen Parameterschätzern des logistischen Regressionsmodells $\hat{\beta}_0 = -23.35,\ \hat{\beta}_1 = 2.064$.

c)

Plotten Sie die Daten sowie die geschätzten Funktionen aus (a) und (b). Diskutieren Sie die Ergebnisse. Warum ist der Verlauf der Vorhersagefunktion des logistischen Regressionsmodells nicht linear?

Aufgabe 3: Maximum Likelihood Schätzer

In der Vorlesung wurde das Maximum-Likelihood Prinzip schon bei der Schätzung der Parameter des linearen sowie des logistischen Regressionsmodells benutzt. Im Allgemeinen und bei genügend großen Datenmengen ist das Maximum-Likelihood (ML) Prinzip das beste Verfahren, um aus einer gegebenen Datenmenge die Parameter des vermuteten Modells zu schätzen. Voraussetzung ist aber eine probabilistische Modellannahme mit endlicher Zahl von Parametern wie sie beim klassischen linearen oder logistischen Regressionsmodell getroffen wurde.

a)

Wiederholen Sie die probabilistischen, parametrischen Modellannahmen des klassischen linearen und logistischen Regressionsmodells. Beschreiben Sie den Unterschied. Wieviele Parameter haben die jeweiligen Modelle?

b)

Das ML-Schätzverfahren dient also der Parameterschätzung eines beliebigen parametrischen Modells. Es sei nun folgendes Modell angenommen: Die beobachteten Daten $y_i \in \mathbb{R}\ \forall i = 1, \ldots, n$ sind n unabhängige Realisierungen eines normalverteilten Zufallsprozesses $y_i \sim N(\mu, \sigma^2)$, wobei dessen Varianz σ^2 bekannt jedoch dessen zweiter Parameter, der Mittelwert μ, unbekannt ist. Leiten Sie den ML-Schätzer für den Mittelwert für dieses Modell und die gegebenen Daten vom Umfang n her.

Hinweis: Stellen Sie die gemeinsame Dichte der Daten unter der Modellannahme auf. Leiten Sie dann die logarithmierte gemeinsame Dichte nach dem gesuchten Parameter μ ab, setzen sie das Ergebnis gleich 0 und lösen sie nach μ auf.