Maschinelles Lernen
Übungsblatt 1

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

08. November 2011
Abgabe bis 15. November

Aufgabe 1: Lineare Regression (5 Punkte)

a)

Seien die folgenden Dateninstanzen aus dem Beispiel aus der Vorlesung (Gasverbrauch) gegeben: $D = \{(1, 6.25), (2, 6), (4, 5.5)\}$. Berechnen Sie die Zielvariable $\hat{y}$ für $x = 10$ mit der Methode der kleinsten Quadrate. Der tatsächliche Wert sei $y = 2$. 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)

Was kann unternommen werden, wenn das lineare Modell nicht die Verteilung der Daten widerspiegelt? Welche Probleme können auftreten?

c)

In der Vorlesung wurde für die einfache lineare Regression bewiesen, dass

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

die Residuenquadratsumme (RSS) minimiert. Betrachten Sie den Beweis dafür und geben Sie die Zwischenschritte bei der partiellen Ableitung an. Das Nullsetzen der Ableitung ist ein notwendiges Kriterium für die Existenz eines Extremums. Begründen Sie, dass es sich bei der vorliegenden Lösung um ein globales Minimum handelt.

Aufgabe 2: R (3 Punkte)

a)

Lesen Sie Kapitel 2 und 3 von „An Introduction to R“. Schreiben Sie 3 Beispiele in R-Code, wie man Vektoren auf unterschiedliche Weise erzeugen kann. Schreiben Sie 3 Sätze über Objekte/Klassen in R.

b)

Erstellen Sie ein lineares Regressionsmodell für den in R integrierten Datensatz cars. Gesucht ist ein Schätzer für die Variable `cars$dist` gegeben `cars$speed`. Lesen Sie die Koeffizienten aus, plotten Sie die Daten, und fügen Sie dem Plot die Regressionslinie hinzu. Geben Sie den verwendeten R-Code an.

Hinweis: Die hierfür nötigen Befehle finden Sie in Appendix A, den Sie im letzten Übungsblatt durchgearbeitet haben.

c)

Sind Sie zufrieden mit diesem linearen Modell? Wie könnte man das Verhältnis von Geschwindigkeit und Bremsweg noch modellieren?

Aufgabe 3: Weka (1 Punkt)

a)

Laden Sie den Datensatz `lymph.arff` in Weka. Wandeln Sie die nominalen Attribute des Datensatzes mit Hilfe des Filters `unsupervised/attributes/NominalToBinary` in binäre Variablen um und speichern Sie die Daten als `lymph-bin.arff`. Vergleichen Sie nun die beiden ARFF-Dateien. Welche Unterschiede fallen Ihnen auf?