Prof. Dr. Claus Weihs, Prof. Dr. Katharina Morik
Julia Schiffner, Marco Stolpe

Dortmund, 21. April
Abgabe: bis zum 28.04. um 10:00 Uhr an
marco.stolpe@tu-dortmund.de

Übungen zur Vorlesung

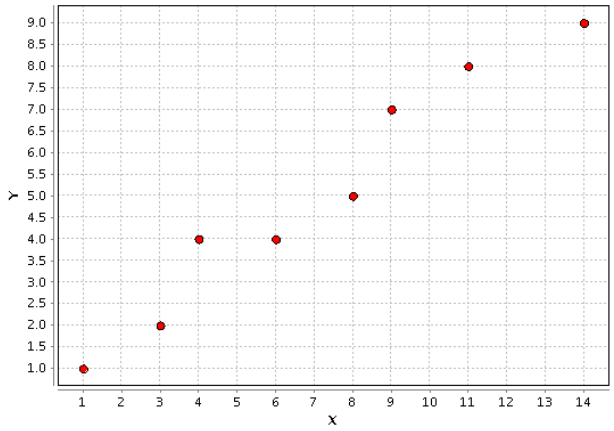
Wissensentdeckung in Datenbanken

Sommersemester 2009

Blatt 2

Aufgabe 2.1 (6 Punkte)
Gegeben seien folgende Datenpunkte im $\mathbb{R}^2$.

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9



Mit Hilfe $(p+1)$-dimensionaler Vektoren $(1, x_1, \ldots, x_p)$ kann β_0 in den Vektor $\vec{\beta}$ kodiert werden. Die Darstellung einer linearen Funktion vereinfacht sich dadurch zu:

$$y = f(\vec{x}) = \sum_{i=0}^{p} \beta_i x_i = \vec{x}^T \vec{\beta} \quad .$$

In der Vorlesung haben Sie gesehen, dass sich der Parameter $\vec{\beta}$ über die Minimierung der quadratischen Fehlersumme $RSS(\vec{\beta})$ bestimmen lässt.

(a) Stellen Sie für die gezeigten Punkte die zu minimierende Funktion bzgl. des RSS-Kriteriums auf - verwenden Sie dabei konkrete Zahlenwerte.

(b) Lösen Sie für diese Werte das Minimierungsproblem.

(c) Wie groß ist der Trainingsfehler Ihres linearen Modells?

(d) Verallgemeinern Sie das Minimierungsproblem, indem Sie Variablen einführen. Leiten Sie her, wie sich die Werte des Vektors $\vec{\beta}$ im $\mathbb{R}^2$ allgemein berechnen lassen.

(e) Berechnen Sie den Korrelationskoeffizienten von X und Y.

Geben Sie bei der Bearbeitung aller Rechenaufgaben den Rechenweg an!

Aufgabe 2.2 (4 Punkte)
In dieser Aufgabe sollen Sie die Software RapidMiner besser kennen lernen. Speichern Sie Ihre Prozesse für die Teilaufgaben b), d) und e) in einer XML-Datei und geben Sie diese Datei(en) bitte mit ab.

Die Daten aus der ersten Aufgabe finden Sie als CSV-Datei unter folgender Adresse:

`http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS09/linreg.csv`

Öffnen Sie RapidMiner und lesen Sie die Daten über den Operator `ExampleSource` ein. Verwenden Sie dazu den *Data Loading Wizard*.

(a) Erstellen Sie für die Daten einen Scatterplot (wie in Aufgabe 1). Speichern Sie den Plot im PNG-Format und geben Sie ihn mit ab.

(b) Hängen Sie den Operator `LinearRegression` an das Experiment an. Führen Sie eine lineare Regression durch.

(c) Wenden Sie das gelernte Modell auf den Datensatz an und berechnen Sie über den Operator `RegressionPerformance` folgende Kennzahlen: `root mean squared error`, `absolute error`, `relative error` und `correlation`.

(d) Stellen Sie das Experiment auf eine Leave-One-Out Kreuzvalidierung um. (Verwenden Sie dazu eine `OperatorChain`). Geben Sie die veränderten Kennzahlen an.

(e) Entfernen Sie die Kreuzvalidierung und stellen Sie den Zustand von Teilaufgabe c) wieder her. Fügen Sie direkt hinter dem Lerner einen `ModelWriter` ein. Speichern Sie das gelernte Modell in einer Datei Ihrer Wahl. Entfernen Sie den Writer und ersetzen Sie den Lerner durch einen `ModelLoader`, der das gespeicherte Modell einliest.