Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Volker Tresp
Tobias Emrich

10.05.2013

Maschinelles Lernen und Data Mining

Sommersemester 2013

Übungsblatt 2

Besprechung des Übungsblattes am 16.05.2013

Aufgabe 2-1 Lineare Regression

Gegeben die folgenden Daten der Modellvariable X und deren Ausprägungen Y:

x	3	4	5	6	7	8
y	150	155	150	170	160	175

a) Nehmen Sie an, dass das Modell folgenden linearen Zusammenhang aufweist:
$y_i = \beta_0 + \beta_1 x_i = x^T w$
Bestimmen Sie w mit Hilfe des LS-Schätzers aus der Vorlesung.

b) Nehmen Sie nun den nichtlinearen Zusammenhang
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 = x^T w$
an und bestimmen Sie erneut w

c) Wie könnte man den empirischen quadratischen Fehler zwischen Modell und Daten grafisch notieren? Erklären und skizzieren Sie ihren Vorschlag im zwei- und dreidimensionalen Datenraum mit beliebigen Daten.

d) Welches der Modelle a), b) ist besser? Berechnen Sie den mittleren quadratischen Fehler und bewerten Sie die Modelle. Wie ließe sich ein besseres Modell erstellen?

Hinweis: Matrixoperationen (invertieren etc.) müssen nicht von Hand berechnet werden. Im CIP-Pool stehen z.B. `R` oder Maple (Aufruf via `xmaple`) zur Verfügung.

Aufgabe 2-2 Regularisierung / Overfitting

a) Was versteht man unter dem Begriff *Overfitting* und wie kommt es zustande.

b) Wie kann man erkennen, dass ein Modell „overfitted“ ist?

c) Wie kann man Overfitting verhindern?