

Maschinelles Lernen und Data Mining
Sommersemester 2010
Übungsblatt 2

Besprechung des Übungsblattes am 20.05.2010

Aufgabe 2-1 Lineare Regression
schriftlich bearbeiten

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

x	23	24	25	26	27	28
y	150	155	150	170	160	175

- a) Nehmen Sie an, dass das Modell folgenden linearen Zusammenhang aufweist:
$$y_i = a + \beta x_i = x^T w$$

Bestimmen Sie w mit Hilfe des LS-Schätzers aus der Vorlesung.
- b) Nehmen Sie nun den nichtlinearen Zusammenhang
$$y_i = a + \beta_1 x_i + \beta_2 x_i^2 = x^T w$$

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? (nicht schriftlich zu bearbeiten)

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

Aufgabe 2-2 Regularisierung / Overfitting
schriftlich bearbeiten

- 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?