

Maschinelles Lernen und Data Mining
Sommersemester 2014
Übungsblatt 2

Besprechung des Übungsblattes am 08.05.2014

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

- Nehmen Sie an, dass das Modell folgenden linearen Zusammenhang aufweist:
 $y_i = \beta_0 + \beta_1 x_i = x_i^T w$
Bestimmen Sie w mit Hilfe des LS-Schätzers aus der Vorlesung.
- Nehmen Sie nun den nichtlinearen Zusammenhang
 $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 = x_i^T w$
an und bestimmen Sie erneut w
- 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.
- 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 Wdh. Vektor Calculus

Berechnen Sie $\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$ für untenstehende Funktionen. *Hinweis:* Für eine Funktion $g(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ mit $\mathbf{x} \in \mathbb{R}^n$ gilt:

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g(\mathbf{x})}{\partial x_1} \\ \frac{\partial g(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial g(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

- $g(\mathbf{x}) = \sum_{i=1}^n x_i$,
- $g(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle$, das Standard-Skalarprodukt von \mathbf{x} mit sich selbst,
- $g(\mathbf{x}) = (\mathbf{x} - \mu)^2$ für $\mu \in \mathbb{R}^n$.

Optional:

Aufgabe 2-3 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?

Optional:

Aufgabe 2-4 Curse of Dimensionality vs. Kernel Trick

- a) Erklären Sie den Begriff *Curse of Dimensionality*.
Wie tritt er auf, was ist betroffen, wie kann man entgegenwirken?
- b) Erklären Sie den Begriff *Kernel Trick*.
Was ist er, wozu wird er gebraucht, wie ist die Verbindung zum *Curse of Dimensionality*?

Aufgabe 2-5 Basisfunktionen von Neuronalen Netzen

Die Ausgabe eines neuronalen Netzes für einen Testvektor \mathbf{x}_i ist definiert durch

$$f(\mathbf{x}_i) = \sum_{h=0}^{M_\phi-1} w_h \phi_h(\mathbf{x}_i, \mathbf{v}_h).$$

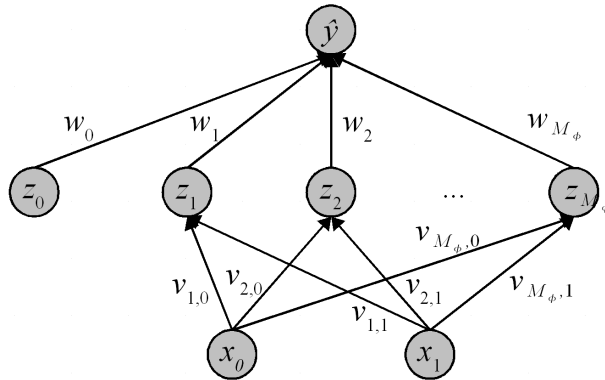
Die Gewichte der einzelnen Neuronen können über die Backpropagation-Regel mit musterbasiertem Gradientenabstieg gelernt werden. In der Vorlesung wurden die neuronalen Netze mit sigmoiden Neuronen vorgestellt. Natürlich können auch andere Basisfunktionen verwendet werden.

- a) Welche Eigenschaften müssen diese Basisfunktionen erfüllen?
- b) Ist eine Linearkombination $\phi(\mathbf{x}_i, \mathbf{v}_h) = z_h = \sum_{j=0}^M v_{h,j} x_{i,j}$ hierfür geeignet? Begründen Sie.
- c) Ist die Anzahl der Parameter für $\phi(\mathbf{x}_i, \mathbf{v}_h)$ beschränkt? Können mehrere verschiedene Basisfunktionen für ein neuronales Netz verwendet werden?

b.w.

Aufgabe 2-6 Ein einfaches neuronales Netz

Unten abgebildet sehen Sie ein zweischichtiges neuronales Netz mit einem Eingabeneuron $x \in \mathbb{R}$ und je einem Biasneuron $x_0 = z_0 = 1$ (d.h. $\mathbf{x}_i = (1, x_{i,1})^T$) in der Eingabeschicht und der versteckten Schicht.



Als Aktivierungsfunktion der versteckten Neuronen verwenden wir einen Sigmoiden, also

$$z_h = \phi(\mathbf{x}_i, \mathbf{v}_h) = \frac{1}{1 + \exp\left(-\sum_{j=0}^M v_{h,j} x_{i,j}\right)},$$

das Ausgangsneuron \hat{y} wird wie üblich über eine Linearkombination gebildet.

- Zeigen Sie, dass gilt: $\frac{\partial z_h}{\partial v_{h,j}} = x_{i,j} \cdot z_h \cdot (1 - z_h)$
- Drücken Sie den maximalen Wert von \hat{y} in Abhängigkeit von \mathbf{w} aus, wenn alle Ausgangsgewichte w_h ($h \in \{0, \dots, M_\phi\}$) positiv sind. Was ist der minimale Wert?
- Wie sieht \hat{y} aus, wenn $v_{h,j} = 0$ für alle $j \in \{0, \dots, M\}, h \in \{1, \dots, M_\phi\}$? Welche Funktion erhalten Sie für \hat{y} , wenn alle $v_{h,j} = c, c \neq 0$?