AG Neuroinformatik
Prof. Dr. Helge Ritter

Universität Bielefeld
Technische Fakultät

Übungen zur Vorlesung
Vertiefung Maschinelles Lernen

(14-tägig, Fr 8:30-10:00)

Blatt 1
Abgabe Fr, 9. November 2012

Aufgabe 1 (Additive Mischung von Funktionslernern) Diskutieren Sie folgende Vorschrift zur Verknüpfung der Ausgaben $y_i(x) > 0$ von N identischen Funktions-Lernern:

$$y = \left(\frac{1}{N} \sum_{i=1}^{N} y_i(x)^p \right)^{1/p}$$

Dabei ist $p > 0$ ein vorgegebener Parameter.

a) wie wirkt sich ein Fehler δy_i eines einzelnen Lerners auf das Ergebnis aus?

b) welchen Einfluß hat der Parameter p auf die Robustheit des Resultats gegenüber „Ausreißern“?

c) Betrachten Sie den Fall $p = 2$ und den Fall normalverteilter Antwortfehler (mit Varianz σ^2 um Null verteilt) der einzelnen Klassifikatoren. Welche Varianz hat der Fehler von y?

Aufgabe 2 Für eine „weiche Überlagerung“ von Experten-Ausgaben x_i wird häufig die sog. *Softmax-Funktion* verwendet:

$$F_i(x_1..x_N) = \frac{\exp \beta x_i}{\sum_i \exp \beta x_i}$$

a) wie verhält sich $F(x_1..x_N)$ für $\beta \mapsto \infty$? Wie für $\beta \mapsto 0$? Wie verhalten sich die Antworten F_i unter einer Erhöhung aller x_i um denselben Wert Δ?

b) Erläutern Sie, wieso eine „naive“ numerische Implementierung von $F_i(x_1..x_N)$ (Ausrechnen aller Exponentialfunktionen und Auswertung des Bruchs) für große Werte von β fehlschlägt. Wie kann man besser vorgehen?

c) Zeigen Sie, daß die Jacobi-Matrix $J_{ij} = \boldsymbol{\partial} F_i / \boldsymbol{\partial} F_j$ die Elemente

$$J_{ij} = \beta y_i (\delta_{ij} - y_j)$$

und die Eigenschaft $\sum_i J_{ij} = \sum_j J_{ij} = 0$ $besitzt$.

Aufgabe 3 $(EM - Algorithmus)$

a) Zeigen Sie die Gültigkeit der Ungleichung

$$\log \sum_s P_s(z) \geq \sum_s q_s \log P_s(z) - \sum_s q_s \log q_s$$

für beliebig gewählte Wahrscheinlichkeitsdichten $P_s(z)$ und beliebig gewählte nichtnegative Parameter q_s, die nur der Bedingung $\sum_s q_s = 1$ unterliegen müssen. Hinweis: machen Sie Gebrauch von der Jensen'schen Ungleichung

$$\log \left(\sum_s q_s x_s \right) \geq \sum_s q_s \log(x_s)$$

die für $\sum_s q_s = 1, q_s \geq 0$ gilt.

b) Führen Sie für jeden Datenpunkt i einen eigenen Satz von Wichtungsfaktoren q_{si} ein und wenden Sie das vorstehende Ergebnis auf die log-Likelihood Funktion

$$L_D(\boldsymbol{\theta}) = \sum_i \log P(\boldsymbol{z}_i; \boldsymbol{\theta}) = \sum_i \log \sum_s P_s(\boldsymbol{z}_i, \boldsymbol{\theta})$$

an. Begründen Sie so die Aussage, daß die Funktion

$$-F(\boldsymbol{q}; \boldsymbol{\theta}) = \sum_{s,i} q_{si} \log P_s(\boldsymbol{z}_i; \boldsymbol{\theta}) - \sum_{s,i} q_{si} \log q_{si}$$

eine *untere Schranke* für die log-Likelihood $L_D(\boldsymbol{\theta})$ darstellt, die durch den M-Schritt bezüglich $\boldsymbol{\theta}$ maximiert wird.

c) Für welche Wahl der Parameter q_{si} wird die untere Schranke maximiert? Was folgt daraus für das EM-Verfahren?