Übung zur Vorlesung

Maschinelles Lernen 2

Sommersemester 2010

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Dr. Mikio L. Braun, Dr. Konrad Rieck
Email: klaus-robert.mueller@tu-berlin.de

Blatt 3

Abgabe bis Montag, 3. Mai, 2010, 13:00 Uhr bei Dr. Konrad Rieck (Software per email an konrad.rieck@tu-berlin.de).

In diesem Aufgabenblatt wird der Algorithmus Stationary Subspace Analysis (SSA) untersucht.

1. **(Lineare Transformation von Zufallsvariablen, 5 Punkte)** Sei $X \in \mathbb{R}^D$ eine D-variate Zufallsvariable mit Mittelwert μ und Kovarianzmatrix Σ,

$$\begin{aligned}\mu &= E[X] \text{ und}\\ \Sigma &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top].\end{aligned}$$

   Sei $A \in \mathbb{R}^{m \times D}$ eine beliebige Matrix. Bestimmen Sie den Mittelwert und die Kovarianzmatrix der transformierten Zufallsvariablen AX als Funktion von μ und Σ.

2. **(SSA Kostenfunktion, 15 Punkte)** Sei D die Anzahl Eingabedimensionen, $d < D$ die Anzahl der stationären Quellen und N die Anzahl der Epochen. Die geschätzten Mittelwerte und Kovarianzmatrizen der N Epochen bezeichnen wir mit $\hat{\mu}_1, \ldots, \hat{\mu}_N \in \mathbb{R}^D$ und $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_N \in \mathbb{R}^{D \times D}$. Die Epochen wurden gemeinsam zentriert und durch eine Whitening-Transformation dekorreliert, d.h. es gilt

$$\sum_{i=1}^{N} \hat{\mu}_i = 0 \;\text{ und }\; \sum_{i=1}^{N} \hat{\Sigma}_i = I. \tag{1}$$

   Gegeben den Rotationsanteil $R \in \mathbb{R}^{D \times D}$ (mit $R^\top R = I$) der geschätzten Entmischungsmatrix können der Mittelwert und die Kovarianzmatrix der geschätzten stationären Quellen in der i-ten Epoche geschrieben werden als

$$\hat{\mu}_i^{\mathfrak{s}} = I^d R \hat{\mu}_i \;\text{ und }\; \hat{\Sigma}_i^{\mathfrak{s}} = I^d R \hat{\Sigma}_i (I^d R)^\top,$$

   wobei $I^d \in \mathbb{R}^{d \times D}$ die auf die ersten d Zeilen beschnittene Einheitsmatrix ist.

   Zeigen Sie, dass die Kostenfunktion von SSA

$$\begin{aligned}L(R) &= \sum_{i=1}^{N} D_{\mathrm{KL}}\left[\mathcal{N}(\hat{\mu}_i^{\mathfrak{s}}, \hat{\Sigma}_i^{\mathfrak{s}}) \,||\, \mathcal{N}(0, I)\right]\\ &= \frac{1}{2}\sum_{i=1}^{N}\left(-\log\det\hat{\Sigma}_i^{\mathfrak{s}} + \hat{\mu}_i^{\mathfrak{s}\top}\hat{\mu}_i^{\mathfrak{s}}\right) - \frac{N-1}{2}d.\end{aligned} \tag{2}$$

   die angegebene Form (2) hat. Verwenden Sie dazu die Definition der Kullback-Leibler Divergenz zwischen zwei Multivariaten Gaußverteilungen in d Variablen,

$$\begin{aligned}&D_{\mathrm{KL}}\left(\mathcal{N}(\mu_0, \Sigma_0) \,||\, \mathcal{N}(\mu_1, \Sigma_1)\right)\\ &\qquad = \frac{1}{2}\left(\log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^\top\Sigma_1^{-1}(\mu_1 - \mu_0) - d\right),\end{aligned}$$

   und die Voraussetzungen (1).

3. **(Anwendung, 10 Punkte)** Auf der Website finden Sie einen Datensatz und eine Matlab Implementation des SSA Algorithmus. Bestimmen Sie die Anzahl der stationären Quellen mit Hilfe des SSA Algorithmus. Beschreiben Sie Ihr Vorgehen und begründen Sie die Antwort, z.B. auch durch geeignete Plots.