Übung zur Vorlesung

# Maschinelles Lernen 2

Sommersemester 2011

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

## Blatt 2

Abgabe bis Montag, **2. Mai 2011**, 12:00 Uhr, Briefkasten bei Raum FR6061

In diesem Übungsblatt werden die Themen von zwei Vorlesungen (19. und 26. April) behandelt: Canonical Correlation Analysis (CCA) und Stationary Subspace Analysis (SSA). Es können so 50 Punkte erzielt werden.

1. **(Kernel Canonical Correlation Analysis, 20 Punkte)** Gegeben seien zwei Datenmatrizen $X \in \mathbb{R}^{U \times N}$, $Y \in \mathbb{R}^{V \times N}$ ($N$ sei die Anzahl der Datenpunkte, $U$, $V$ die Dimensionalität der jeweiligen Variable). Die *CCA* Lösung maximiert die Kreuz-Korrelation und beschränkt die Auto-Korrelation in jeder Variable auf 1:

$$\begin{aligned} \operatorname{argmax}_{w_x, w_y} \quad & w_x^\top C_{xy} w_y \\ \text{s.t.} \quad & w_x^\top C_{xx} w_x &= 1 \\ & w_y^\top C_{yy} w_y &= 1. \end{aligned} \tag{1}$$

Hier bezeichnen $C_{xx} = XX^\top \in \mathbb{R}^{U \times U}$ und $C_{yy} = YY^\top \in \mathbb{R}^{V \times V}$ die Auto-Kovarianz Matrizen und $C_{xy} = XY^\top \in \mathbb{R}^{U \times V}$ die Kreuz-Kovarianz Matrix zwischen $x$ und $y$. Bei *Kernel CCA* werden die Lösungen $w_x \in \mathbb{R}^U, w_y \in \mathbb{R}^V$ als Linearkombination von Datenpunkten dargestellt:

$$w_x = X\alpha, w_y = Y\beta. \tag{2}$$

Zeigen Sie, dass die Koeffizienten $\alpha$, $\beta$ gegeben sind durch die Lösung der generalisierten Eigenwertgleichung

$$\begin{bmatrix} 0 & K_X K_Y \\ K_Y K_X & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \rho \begin{bmatrix} K_X^2 & 0 \\ 0 & K_Y^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \tag{3}$$

wobei $K_X = X^\top X$ und $K_Y = Y^\top Y$ die *linearen Kerne* der Datenmatrizen bezeichnen.

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

$$\mu = E[X] \text{ und} \tag{4}$$

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top]. \tag{5}$$

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 $\mu$ und $\Sigma$.

3. **(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 $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_N \in \mathbb{R}^D$ 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

$$\sum_{i=1}^{N} \hat{\boldsymbol{\mu}}_i = 0 \text{ und } \sum_{i=1}^{N} \hat{\Sigma}_i = I. \tag{6}$$

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

$$\hat{\boldsymbol{\mu}}_i^{\mathfrak{s}} = I^d R \hat{\boldsymbol{\mu}}_i \text{ und } \hat{\Sigma}_i^{\mathfrak{s}} = I^d R \hat{\Sigma}_i (I^d R)^\top, \tag{7}$$

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

$$\begin{aligned} L(R) &= \sum_{i=1}^{N} D_{\mathrm{KL}} \left[ \mathcal{N}(\hat{\boldsymbol{\mu}}_i^{\mathfrak{s}}, \hat{\Sigma}_i^{\mathfrak{s}}) \mid\mid \mathcal{N}(0, I) \right] \\ &= \frac{1}{2} \sum_{i=1}^{N} \left( -\log \det \hat{\Sigma}_i^{\mathfrak{s}} + \hat{\boldsymbol{\mu}}_i^{\mathfrak{s}\top} \hat{\boldsymbol{\mu}}_i^{\mathfrak{s}} \right) - \frac{N-1}{2} d. \end{aligned} \tag{8}$$

die angegebene Form (8) hat. Verwenden Sie dazu die Definition der Kullback-Leibler Divergenz zwischen zwei Multivariaten Gaußverteilungen in $d$ Variablen,

$$\begin{aligned} &D_{\mathrm{KL}} \left( \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) \mid\mid \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1) \right) \\ &\qquad = \frac{1}{2} \left( \log \left( \frac{\det \Sigma_1}{\det \Sigma_0} \right) + \operatorname{tr} \left( \Sigma_1^{-1} \Sigma_0 \right) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \Sigma_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - d \right), \end{aligned} \tag{9}$$

und die Voraussetzungen (6).

4. **(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.

---

Für Fragen zum Übungsblatt bitte in der Google Group http://groups.google.com/group/ml-tu registrieren und die Fragen an die Mailingliste stellen.