Machine learning 2
Exercise sheet 4

FLEISCHMANN Kay, Matrnr: 352247
ROHRMANN Till, Matrnr: 343756

May 8, 2013

6 Kernel Canonical Correlation Analysis

(a) Given some training data $X \in \mathbb{R}^{d1 \times N}$ and $Y \in \mathbb{R}^{d2 \times N}$. The idea behind CCA is to find two features w_x and w_y in input space such that their correlation is maximised. Let $C_{xx} = XX^T$, $C_{yy} = YY^T$, $C_{xy} = XY^T$ and $C_{yx} = YX^T$.

Formally: Find $w_x \in \mathbb{R}^{d_1}$, $w_y \in \mathbb{R}^{d_2}$ which maximize

$$w_x^T C_{xy} w_y \tag{1}$$

with subject to

$$w_x^T C_{xx} w_x = 1 \tag{2}$$
$$w_y^T C_{yy} w_y = 1 \tag{3}$$

Show that it is always possible to find an optimal solution in the span of the data, that is $w_x = X\alpha_x$, $w_y = Y\alpha_y$:

Proof by contradiction. Let's assume

$$\max_{\alpha_x, \alpha_y \in \mathbb{R}^N} \alpha_x^T X^T C_{xy} Y \alpha_y \quad < \quad \max_{v \in \mathbb{R}^{d_1}, w \in \mathbb{R}^{d_2}} v^T C_{xy} w \tag{4}$$

We can seperate the space $\mathbb{R}^{d_1} = \text{span}\{X\} \cup \text{span}\{X\}^\perp$ into the span of the column vectors of X and its orthogonal space. The same holds for $\mathbb{R}^{d_2} = \text{span}\{Y\} \cup \text{span}\{Y\}^\perp$. Thus every $v \in \mathbb{R}^{d_1}$ can be represented by its projection v_X into $\text{span}\{X\}$ and its projection $v_{X^\perp}$ in $\text{span}\{X\}^\perp$.

$$v \quad = \quad v_X + v_{X^\perp}$$

We can now deduce the following:

$$\max_{v \in \mathbb{R}^{d_1}, w \in \mathbb{R}^{d_2}} v^T C_{xy} w \quad = \quad \max_{v \in \mathbb{R}^{d_1}, w \in \mathbb{R}^{d_2}} (v_X + v_{X^\perp})^T XY^T (w_Y + w_{Y^\perp}) \tag{5}$$

Because $v_{X^\perp}$ belongs to the orthogonal space $\text{span}\{X\}^\perp$, the term $X^T v_{X^\perp} = 0$. The same holds for $w_{Y^\perp}$, that is to say $Y^T w_{Y^\perp} = 0$. This leads to:

$$\begin{aligned}(5) \quad &= \max_{v_X \in \text{span}\{X\}, w_Y \in \text{span}\{Y\}} v_X^T XY^T w_Y \\ &= \max_{\alpha_x, \alpha_y \in \mathbb{R}^N} \alpha_x^T X^T C_{xy} Y \alpha_y \end{aligned} \tag{6}$$

Where the equation (6) is just another form to express that v_X lies in the space $\text{span}\{X\}$ and w_Y lies in the space $\text{span}\{Y\}$. By equating equation (4) with (6) we get

$$\max_{\alpha_x, \alpha_y \in \mathbb{R}^N} \alpha_x^T X^T C_{xy} Y \alpha_y \quad < \quad \max_{\alpha_x, \alpha_y \in \mathbb{R}^N} \alpha_x^T X^T C_{xy} Y \alpha_y$$

Which is obviously a contradiction. Thus we can always find an optimal solution for the equation (1) in the span of the data X and Y respectively. □

Derive the dual optimization problem:

$$\mathcal{L}(\alpha, \beta) = w_x^T C_{xy} w_y - \frac{1}{2}\alpha(w_x^T C_{xx} w_x - 1) - \frac{1}{2}\beta(w_y^T C_{yy} w_y - 1)$$

$$\frac{\partial L}{\partial w_x^T} = XY^T w_y - \alpha(XX^T w_x) = 0 \Leftrightarrow XY^T w_y = \alpha(XX^T w_x) \tag{7}$$

$$\frac{\partial L}{\partial w_y^T} = YX^T w_x - \beta(YY^T w_y) = 0 \Leftrightarrow YX^T w_x = \beta(YY^T w_y) \tag{8}$$

Multiplication w_x^T with equation (7) and w_y^T with Equation (8) results to:

$$w_x^T XY^T w_y = \alpha(w_x^T XX^T w_x)$$
$$w_y^T YX^T w_x = \beta(w_y^T YY^T w_y)$$

Because of constraints (2) and (3)

$$\alpha(w_x^T XX^T w_x) = \beta(w_y^T YY^T w_y) \Rightarrow \alpha = \beta \tag{9}$$

Next we combine equations (7), (8) and (9).

$$C_{xy} w_y = \alpha C_{xx} w_x$$
$$C_{yx} w_x = \alpha C_{yy} w_y$$

Written in matrix form, we finally get a generalized eigenvalue problem:

$$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \alpha \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} \tag{10}$$

6.1 Dual optimization problem

Since we know that there exists an optimal solution in the data span, we can represent our w_x and w_y by:

$$w_x = X\alpha_x \tag{11}$$
$$w_y = Y\alpha_y \tag{12}$$

for some $\alpha_x, \alpha_y \in \mathbb{R}^N$. By substituting equations (11) and (12) into equation (10) with a subsequent left multiplication of the matrix

$$\begin{bmatrix} X^T & 0 \\ 0 & Y^T \end{bmatrix}$$

we obtain the following:

$$\begin{bmatrix} 0 & X^T XY^T \\ Y^T YX^T & 0 \end{bmatrix} \begin{bmatrix} X\alpha_x \\ Y\alpha_y \end{bmatrix} = \rho \begin{bmatrix} X^T XX^T & 0 \\ 0 & Y^T YY^T \end{bmatrix} \begin{bmatrix} X\alpha_x \\ Y\alpha_y \end{bmatrix}$$

Which is equivalent to

$$\begin{bmatrix} 0 & X^T XY^T Y \\ Y^T YX^T X & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = \rho \begin{bmatrix} X^T XX^T X & 0 \\ 0 & Y^T YY^T Y \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

With $K_X = X^T X$ and $K_Y = Y^T Y$ we finally obtain

$$\begin{bmatrix} 0 & K_X K_Y \\ K_Y K_X & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = \rho \begin{bmatrix} K_X^2 & 0 \\ 0 & K_Y^2 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

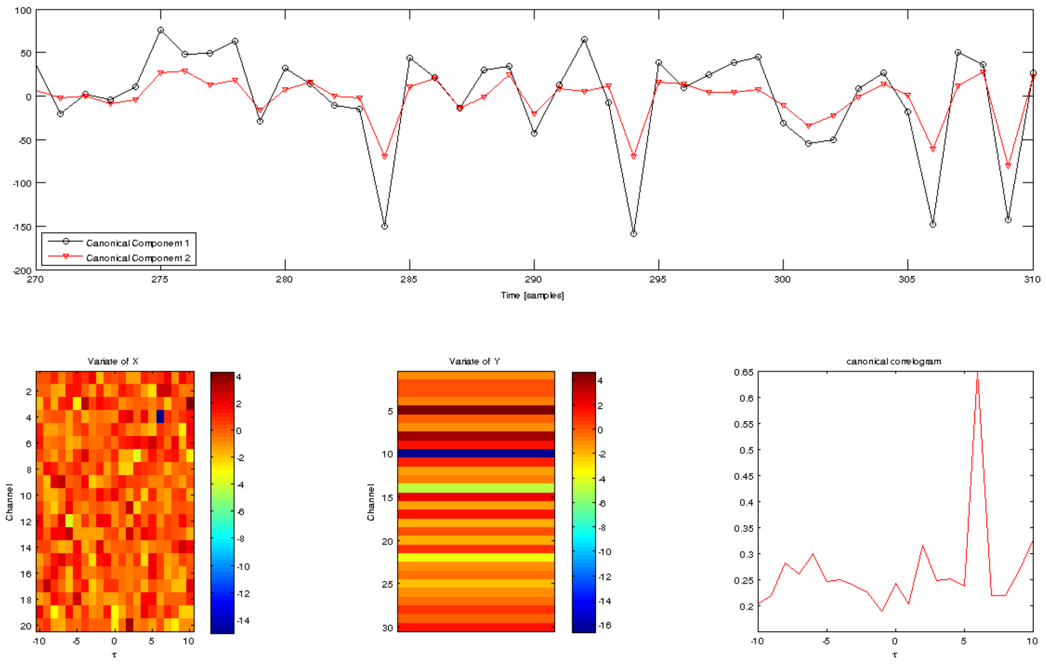

Figure 1: Results of the tkCCA.

6.2 Describe how the generalized eigenvalue problem from exercise (a) - and thus CCA - can be kernelized.

To apply the CCA to a feature space without explicitely calculating the mapping between input and feature space, we can easily adapt the existing method. Given that we have a kernel function $k(\cdot,\cdot)$ representing the inner product in our feature space, we only have to substitute K_X by $(\tilde{K}_X)_{i,j} = k(x_i, x_j)$ and K_Y by $(\tilde{K}_Y)_{i,j} = k(y_i, y_j)$ where x_i is the i-th column of X and y_i the i-th column of Y. This can easily be done because the original problem has been transformed into the dual problem which uses only scalar products. The resulting generalized eigenvalue problem is then:

$$\begin{bmatrix} 0 & \tilde{K}_X\tilde{K}_Y \\ \tilde{K}_Y\tilde{K}_X & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = \rho \begin{bmatrix} \tilde{K}_X^2 & 0 \\ 0 & \tilde{K}_Y^2 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

7 tkCCA

We can clearly see in the canonical correlogram that there exists a strong correlation between X and Y at the time shift $\tau = 6$. Since there exists no other maximum which has a comparable magnitude, we can conclude that the hidden one-dimensional signal occurs probably with a delay of 6 time units in the data set Y.