Exercise Sheet 3

Machine Learning 2, SS16

May 16, 2016

Sascha Huk, 321249; Viktor Jeney, 348969; Mario Tambos, 380599; Jan Tinapp, 0380549

Exercise 1

(a) For data-span constructed $w_x = X\alpha_x$ and $w_y = Y\alpha_y$ the primal problem is:

$$\max_{\alpha_x,\alpha_y} {\alpha_x}^\top X^\top C_{xy} Y \alpha_y$$

$$\text{s.t. } {\alpha_x}^\top X^\top C_{xx} X \alpha_x - 1 = 0 \qquad , \qquad {\alpha_y}^\top Y^\top C_{yy} Y \alpha_y - 1 = 0$$

Lagrangian (the factor 1/2 is introduced just for convenience):

$$\mathcal{L} = \alpha_x^\top X^\top C_{xy} Y \alpha_y - \frac{1}{2}\lambda_x(\alpha_x^\top X^\top C_{xx} X \alpha_x - 1) - \frac{1}{2}\lambda_y(\alpha_y^\top Y^\top C_{yy} Y \alpha_y - 1)$$

Partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial \alpha_x^\top} = X^\top C_{xy} Y \alpha_y - \lambda_x X^\top C_{xx} X \alpha_x \overset{!}{=} 0 \qquad , \qquad \frac{\partial \mathcal{L}}{\partial \alpha_y^\top} = Y^\top C_{yx} X \alpha_x - \lambda_y Y^\top C_{yy} Y \alpha_y \overset{!}{=} 0$$

We now multiply with $\alpha_x^\top$, $\alpha_y^\top$

$$\alpha_x^\top X^\top C_{xy} Y \alpha_y = \lambda_x \alpha_x^\top X^\top C_{xx} X \alpha_x \qquad , \qquad \alpha_y^\top Y^\top C_{yx} X \alpha_x = \lambda_y \alpha_y^\top Y^\top C_{yy} Y \alpha_y$$

$$\Longrightarrow \alpha_x^\top X^\top C_{xy} Y \alpha_y = \lambda_x \alpha_x^\top X^\top C_{xx} X \alpha_x \qquad , \qquad \alpha_x^\top X^\top C_{xy} Y \alpha_y = \lambda_y \alpha_y^\top Y^\top C_{yy} Y \alpha_y$$

From the auto-cov constraints follows

$$\alpha_x^\top X^\top C_{xy} Y \alpha_y = \lambda_x \underbrace{\alpha_x^\top X^\top C_{xx} X \alpha_x}_{=1} = \lambda_y \underbrace{\alpha_y^\top Y^\top C_{yy} Y \alpha_y}_{=1} \quad \Longrightarrow \quad \lambda_x = \lambda_y$$

Now the derivatives can be rewritten as follows:

$$X^\top C_{xy} Y \alpha_y \overset{!}{=} \lambda_x X^\top C_{xx} X \alpha_x \qquad , \qquad Y^\top C_{yx} X \alpha_x \overset{!}{=} \lambda_x Y^\top C_{yy} Y \alpha_y$$

The same in blockmatrix form:

$$\begin{bmatrix} 0 & X^\top C_{xy} Y \\ Y^\top C_{yx} X & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} \overset{!}{=} \lambda_x \begin{bmatrix} X^\top C_{xx} X & 0 \\ 0 & Y^\top C_{yy} Y \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

$$\Longrightarrow \begin{bmatrix} X^\top C_{xx} X & 0 \\ 0 & Y^\top C_{yy} Y \end{bmatrix}^{-1} \begin{bmatrix} 0 & X^\top C_{xy} Y \\ Y^\top C_{yx} X & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} \overset{!}{=} \lambda_x I \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

(a) $X^\top C_{xx} X$ and $Y^\top C_{xx} Y$ are positive semi-definite. At least after regularizing one of these blocks, the diagonal block matrix becomes positive definite / invertible, which leads to a non-trivial solution.
(b) In comparison, the same conditions occur when viewing the Jacobian of $\mathcal{L}$. The Jacobian has to be negative definite. Since the Jacobian is symmetric, this is true iff the determinants of the principle minors alternate. We already know that the first principle minor $X^\top C_{xx} X$ or $Y^\top C_{yy} Y$ can only be positive. Then, the second principle minor should be negative, which means $-A^2 - B^2 - AB - BA < 0$. This is true as this is the quadratic form $-(A - B)^2 < 0$. Ultimately, positive $X^\top C_{xx} X$ or $Y^\top C_{yy} Y$ is solely necessary for a solution likewise.
(c) After solving the eigenvalue problem above the solution w_x, w_y to the original problem can be obtained by the given identity $w_x = X\alpha_x$ and $w_y = Y\alpha_y$. By finding the solutions ${\alpha_x}^*$ and ${\alpha_y}^*$, the dual variable λ_x is determined (eigenvalue problem). Each eigenvalue λ_x corresponds to an eigenvector $[\alpha_x, \alpha_y]^\top$. Therefore, the dual program does not depend on λ_x anymore, which means $\forall \lambda_x . \mathcal{L}({\alpha_x}^*, {\alpha_y}^*, \lambda_x) = \mathcal{L}({\alpha_x}^*, {\alpha_y}^*)$. We therefore find $\min_{\lambda_x} \max_{\alpha_x,\alpha_y} \mathcal{L}(\alpha_x, \alpha_y, \lambda_x) = \min_{\lambda_x} \mathcal{L}({\alpha_x}^*, {\alpha_y}^*, \lambda_x) = \mathcal{L}({\alpha_x}^*, {\alpha_y}^*)$

Exercise 2

(a) Let Φ be a general symbol for an appropriate feature mapping and define $\Phi(X) := [\Phi(x^{(1)}, \Phi(x^{(2)}), \dots, \Phi(x^{(N)}]$ for each dataset X. As in exercise (1), by starting with $w_x = \Phi(X)\alpha_x$ and $w_y = \Phi(Y)\alpha_y$ one ends up with the eigenvalue problem stated in the task description of exercise (1b), where $A = \Phi(X)^\top \Phi(X) = K_x$ and $B = \Phi(X)^\top \Phi(X) = K_y$. The inner products in the Gramian matrices don't need to be computed via Φ explicitly but rather via the kernels' definitions.

(b) The results α_x, α_y are linear combinations of the N vectors in the respective feature spaces, which is spanned by $\Phi(X)$, $\Phi(Y)$. The solutions can be interpreted as directions in the input space if Φ is a conformal map, otherwise not.

Exercise 3

(a) Taking the derivative of the new objective wrt. $w_x^\top$ and $w_y^\top$ and setting to zero leads to this eigenvalue quation:

$$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} \overset{!}{=} 2\alpha \begin{bmatrix} \{C_{xx}\} & 0 \\ 0 & \{C_{yy}\} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

where

$$\{C_{xx}\} := \begin{cases} C_{xx} & \text{if } w_x^\top C_{xx} w_x > 1 \\ 0 & \text{else} \end{cases} \qquad \{C_{yy}\} := \begin{cases} C_{yy} & \text{if } w_y^\top C_{yy} w_y > 1 \\ 0 & \text{else} \end{cases}$$

For positive definite C_{xx} and C_{yy} solving this equation is possible iff $w_x^\top C_{xx} w_x > 1$ and $w_y^\top C_{yy} w_y > 1$. One could state a corresponding optimization problem as follows:

$$\max_{w_x, w_y} w_x^\top C_{xy} w_y$$
$$\text{s.t. } w_y^\top C_{yy} w_y = w_x^\top C_{xx} w_x > 1$$

(Remark: So what... I'm completely unsure what we should have learned here.)

(b) We express the gradient of the new objective J wrt. θ_x as a function of the Jacobian $\frac{\partial \phi_x}{\partial \theta_x}$:

$$grad_{\theta_x} J = w_x^\top \mathbb{E}[\frac{\partial \phi_x}{\partial \theta_x} \otimes \phi_y^\top] w_y - \alpha \begin{cases} w_x^\top \mathbb{E}[\frac{\partial \phi_x}{\partial \theta_x} \otimes \phi_x^\top + \phi_x \otimes \frac{\partial \phi_x^\top}{\partial \theta_x}] w_x & \text{if } w_x^\top C_{xx} w_x > 1 \\ 0 & \text{else} \end{cases}$$

Since $\phi_x \phi_y^\top$ is a tensor product we find $\frac{\partial \phi_x \phi_y^\top}{\partial \theta_x} = \frac{\partial \phi_x \otimes \phi_y^\top}{\partial \theta_x} = \frac{\partial \phi_x}{\partial \theta_x} \otimes \phi_y^\top + \phi_x \otimes \frac{\partial \phi_y^\top}{\partial \theta_x} = \frac{\partial \phi_x}{\partial \theta_x} \otimes \phi_y^\top$.
Explanation why α became $-\alpha$ in the gradient: Let $f(x) = min(0, 1-x)$ and $g(\theta) = w_x^\top C_{xx} w_x$. Now take the derivative of $f(g(\theta))$ by using the chain rule. So, the minus comes from the derivative of f.