CS-E4830 - Kernel Methods in Machine Learning

Homework Assignment 4 – Pen and Paper

Student: Quoc Tuan Vinh, Ngo (704526)

Question 1. *Regularization Requirement in Kernel CCA*

We have the kernel CCA optimization problem as:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \langle \boldsymbol{K_a\alpha}, \boldsymbol{K_b\beta} \rangle$$

$$\text{subject to } \|\boldsymbol{K_a\alpha}\|_2 = \sqrt{\boldsymbol{\alpha}^T\boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha}} = 1 \text{ and } \|\boldsymbol{K_b\beta}\|_2 = \sqrt{\boldsymbol{\beta}^T\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta}} = 1$$

Our optimization problem can be reformulated as:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^T\boldsymbol{K}_{\boldsymbol{a}}^{\boldsymbol{T}}\boldsymbol{K_b\beta}$$

subject to $\boldsymbol{\alpha}^T\boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^T\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = 1$ *(can be expressed in squared form)*

Apply Lagrange method, we have:

$$L = \boldsymbol{\alpha}^T\boldsymbol{K}_{\boldsymbol{a}}^{\boldsymbol{T}}\boldsymbol{K_b\beta} - \frac{\rho_1}{2}(\boldsymbol{\alpha}^T\boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha} - 1) - \frac{\rho_2}{2}(\boldsymbol{\beta}^T\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} - 1) \quad \textbf{(1)}$$

Where ρ_1, ρ_2 are Lagrange multipliers. Taking the derivatives of L wrt. vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$:

$$\frac{\delta L}{\delta\boldsymbol{\alpha}} = \boldsymbol{K_a}\,\boldsymbol{K_b\beta} - \rho_1\boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha} \qquad \textbf{(2)}$$

$$\frac{\delta L}{\delta\boldsymbol{\beta}} = \boldsymbol{K_b}\,\boldsymbol{K_a\alpha} - \rho_2\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} \qquad \textbf{(3)}$$

Setting both (2) and (3) to zero, and multiplying (2) from the left by $\boldsymbol{\alpha}^T$ and (3) from the left by $\boldsymbol{\beta}^T$, we have:

$$\boldsymbol{\alpha}^T\boldsymbol{K_a}\,\boldsymbol{K_b\beta} - \rho_1\boldsymbol{\alpha}^T\boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha} = 0 \quad \textbf{(4)}$$

$$\boldsymbol{\beta}^T\boldsymbol{K_b}\,\boldsymbol{K_a\alpha} - \rho_2\boldsymbol{\beta}^T\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = 0 \iff \boldsymbol{\alpha}^T\boldsymbol{K_a}\,\boldsymbol{K_b\beta} - \rho_2\boldsymbol{\beta}^T\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = 0 \quad \textbf{(5)}$$

Since $\boldsymbol{\alpha}^T\boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^T\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = 1$, from (4) and (5), we get that $\rho_1 = \rho_2 = \rho$. Therefore, from (2), we have:

$$\boldsymbol{K_a}\,\boldsymbol{K_b\beta} - \rho_1\boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha} = \boldsymbol{K_a}\,\boldsymbol{K_b\beta} - \rho\ \ \boldsymbol{K}_{\boldsymbol{a}}^2\boldsymbol{\alpha} = 0 \iff \boldsymbol{\alpha} = \frac{1}{\rho}\boldsymbol{K}_{\boldsymbol{a}}^{-1}\boldsymbol{K}_{\boldsymbol{a}}^{-1}\boldsymbol{K_a}\,\boldsymbol{K_b\beta} = \frac{1}{\rho}\boldsymbol{K}_{\boldsymbol{a}}^{-1}\boldsymbol{K_b\beta} \quad \textbf{(6)}$$

Substitute (6) into (3), we obtain:

$$\boldsymbol{K_b}\,\boldsymbol{K_a\alpha} - \rho_2\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = \frac{1}{\rho}\boldsymbol{K_b}\,\boldsymbol{K_a}\boldsymbol{K}_{\boldsymbol{a}}^{-1}\boldsymbol{K_b\beta} - \rho\ \ \boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = \frac{1}{\rho}\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} - \rho\ \ \boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = 0$$

$$\iff \boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} = \rho^2\boldsymbol{K}_{\boldsymbol{b}}^2\boldsymbol{\beta} \iff \boldsymbol{I\beta} = \rho^2\boldsymbol{\beta} \quad \textbf{(7)}$$

One problem when obtaining **(6)** and **(7)** is that kernels $\boldsymbol{K_a}, \boldsymbol{K_b}$ must be invertible. To guarantee this, we can regularize these two by adding a regularized constant c_1, c_2 respectively to the diagonal of kernel $\boldsymbol{K_a}, \boldsymbol{K_b}$. Therefore, our restrictions now become:

$$\boldsymbol{\alpha}^T(\boldsymbol{K_a} + c_1\boldsymbol{I})^2\boldsymbol{\alpha} = 1 \text{ and } \boldsymbol{\beta}^T(\boldsymbol{K_b} + c_2\boldsymbol{I})^2\boldsymbol{\beta} = 1 \quad \textbf{(8)}$$

The solutions then can be found by solving the standard eigenvalue problem:

$$(\boldsymbol{K_b} + c_2\boldsymbol{I})^{-2}\boldsymbol{K_b}\,\boldsymbol{K_a}(\boldsymbol{K_a} + c_1\boldsymbol{I})^{-2}\boldsymbol{K_a}\,\boldsymbol{K_b}\boldsymbol{\alpha} = \rho^2\boldsymbol{\alpha} \quad \textbf{(9)}$$

Similarly to CCA case, the formulation can be explained as:

$$\begin{pmatrix} 0 & \boldsymbol{K_a}\,\boldsymbol{K_b} \\ \boldsymbol{K_b}\,\boldsymbol{K_a} & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \rho\begin{pmatrix} (\boldsymbol{K_a} + c_1\boldsymbol{I})^2 & 0 \\ 0 & (\boldsymbol{K_b} + c_2\boldsymbol{I})^2 \end{pmatrix}\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}$$

Question 2. *Kernel CCA is CCA on Hilbert Space Objects*

We need to prove:

$$\widehat{cov}(\langle\phi_a(\boldsymbol{x}_a), \boldsymbol{w}_a\rangle, \langle\phi_b(\boldsymbol{x}_b), \boldsymbol{w}_b\rangle) = \frac{1}{n}\,\boldsymbol{\alpha}^T\boldsymbol{K_a^T}\boldsymbol{K_b}\boldsymbol{\beta}$$

Let S_a and S_b represent the linear spaces spanned by the images of the data points. For any $\boldsymbol{w}_a \in \boldsymbol{\mathcal{H}}_a, \boldsymbol{w}_b \in \boldsymbol{\mathcal{H}}_b$, we have $\boldsymbol{w}_a = \boldsymbol{w}_a^{||} + \boldsymbol{w}_a^{\perp}$, where $\boldsymbol{w}_a^{||} = \sum_{k=1}^{n}\alpha_k\phi_a(\boldsymbol{x}_a^k) \in S_a$ and $\boldsymbol{w}_a^{\perp}$ is orthogonal to all objects $\phi_a \in S_a$. The same logic applies to $\boldsymbol{w}_b$.

Therefore, we can derive the empirical covariance of the transformations in the feature space as below:

$$\begin{aligned}
\widehat{cov}\ (\langle\phi_a(\boldsymbol{x}_a), \boldsymbol{w}_a\rangle, \langle\phi_b(\boldsymbol{x}_b), \boldsymbol{w}_b\rangle) &= \frac{1}{n}\sum_{k=1}^{n}\langle\phi_a(\boldsymbol{x}_a^k), \boldsymbol{w}_a\rangle\langle\phi_b(\boldsymbol{x}_b^k), \boldsymbol{w}_b\rangle \\
&= \frac{1}{n}\sum_{k=1}^{n}\langle\phi_a(\boldsymbol{x}_a^k), \boldsymbol{w}_a^{||} + \boldsymbol{w}_a^{\perp}\rangle\langle\phi_b(\boldsymbol{x}_b^k), \boldsymbol{w}_b^{||} + \boldsymbol{w}_b^{\perp}\rangle \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(\langle\phi_a(\boldsymbol{x}_a^k), \boldsymbol{w}_a^{||}\rangle + \langle\phi_a(\boldsymbol{x}_a^k), \boldsymbol{w}_a^{\perp}\rangle\right)\left(\langle\phi_b(\boldsymbol{x}_b^k), \boldsymbol{w}_b^{||}\rangle + \langle\phi_b(\boldsymbol{x}_b^k), \boldsymbol{w}_b^{\perp}\rangle\right) \quad \textbf{(1)}
\end{aligned}$$

Since $\langle\phi_a(\boldsymbol{x}_a^k), \boldsymbol{w}_a^{\perp}\rangle = \langle\phi_b(\boldsymbol{x}_b^k), \boldsymbol{w}_b^{\perp}\rangle = 0$, **(1)** can be simplified as:

$$\begin{aligned}
\textbf{(1)} &= \frac{1}{n}\sum_{k=1}^{n}\left(\langle\phi_a(\boldsymbol{x}_a^k), \boldsymbol{w}_a^{||}\rangle\right)\left(\langle\phi_b(\boldsymbol{x}_b^k), \boldsymbol{w}_b^{||}\rangle\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\langle\phi_a(\boldsymbol{x}_a^k), \sum_{i=1}^{n}\alpha_i\phi_a(\boldsymbol{x}_a^i)\rangle\langle\phi_b(\boldsymbol{x}_b^k), \sum_{j=1}^{n}\beta_j\phi_b(\boldsymbol{x}_b^j)\rangle \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\langle\phi_a(\boldsymbol{x}_a^i), \phi_a(\boldsymbol{x}_a^k)\rangle\langle\phi_b(\boldsymbol{x}_b^j), \phi_b(\boldsymbol{x}_b^k)\rangle\beta_j \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\boldsymbol{K_a}(\boldsymbol{x}_a^i, \boldsymbol{x}_a^k)\boldsymbol{K_b}(\boldsymbol{x}_b^j, \boldsymbol{x}_b^k)\beta_j = \frac{1}{n}\,\boldsymbol{\alpha}^T\boldsymbol{K_a^T}\boldsymbol{K_b}\boldsymbol{\beta}
\end{aligned}$$

Which is what we need to prove.