Assignment 4 : CS-E4830 Kernel Methods in Machine Learning 2019

The **deadline** for this assignment is **Thursday 04.04.2019 at 4pm**.

If you have **questions** about the assignment, you can ask them in the 'General discussion' section on MyCourses.

We will have a tutorial session regarding the **solutions** of this assignment on 28.03.19 at 4:15 pm in TU1(1017), TUAS, Maarintie 8. The solutions will also be available in MyCourses.

Please follow the **submission instructions** given in MyCourses: `https://mycourses.aalto.fi/course/view.php?id=20602&section=2`.

Pen & Paper exercise

Question 1: Regularization Requirement in Kernel CCA (**2 points**)

The kernel CCA optimization problem can be formulated as

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \langle \mathbf{K}_a\boldsymbol{\alpha}, \mathbf{K}_b\boldsymbol{\beta}\rangle \\ \text{subject to} \quad & \|\mathbf{K}_a\boldsymbol{\alpha}\|_2 = 1 \text{ and } \|\mathbf{K}_b\boldsymbol{\beta}\|_2 = 1. \end{aligned}$$

Using the equality $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$ apply the Lagrange multiplier technique to solve the kernel CCA optimization problem.

Solution 1:

$$L = \boldsymbol{\alpha}^\top \mathbf{K}_a^\top \mathbf{K}_b\boldsymbol{\beta} - \frac{\rho_1}{2}(\boldsymbol{\alpha}^\top \mathbf{K}_a^2\boldsymbol{\alpha} - 1) - \frac{\rho_2}{2}(\boldsymbol{\beta}^\top \mathbf{K}_b^2\boldsymbol{\beta} - 1) \tag{1}$$

where ρ_1 and ρ_2 denote the Lagrange multipliers. Differentiating L with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ gives

$$\frac{\delta L}{\delta \boldsymbol{\alpha}} = \mathbf{K}_a\mathbf{K}_b\boldsymbol{\beta} - \rho_1\mathbf{K}_a^2\boldsymbol{\alpha} = \mathbf{0} \tag{2}$$

$$\frac{\delta L}{\delta \boldsymbol{\beta}} = \mathbf{K}_b\mathbf{K}_a\boldsymbol{\alpha} - \rho_2\mathbf{K}_b^2\boldsymbol{\beta} = \mathbf{0} \tag{3}$$

Multiplying (2) from the left by $\boldsymbol{\alpha}^\top$ and (3) from the left by $\boldsymbol{\beta}^\top$ gives

$$\boldsymbol{\alpha}^\top \mathbf{K}_a\mathbf{K}_b\boldsymbol{\beta} - \rho_1\boldsymbol{\alpha}^\top \mathbf{K}_a^2\boldsymbol{\alpha} = 0 \tag{4}$$

$$\boldsymbol{\beta}^\top \mathbf{K}_b\mathbf{K}_a\boldsymbol{\alpha} - \rho_2\boldsymbol{\beta}^\top \mathbf{K}_b^2\boldsymbol{\beta} = 0. \tag{5}$$

Since $\boldsymbol{\alpha}^\top K_a^2 \boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^\top K_b^2 \boldsymbol{\beta} = 1$, we obtain that

$$\rho_1 = \rho_2 = \rho. \tag{6}$$

Substituting (6) into Equation (2) we obtain

$$\boldsymbol{\alpha} = \frac{\mathbf{K}_a^{-1}\mathbf{K}_a^{-1}\mathbf{K}_a\mathbf{K}_b\boldsymbol{\beta}}{\rho} = \frac{\mathbf{K}_a^{-1}\mathbf{K}_b\boldsymbol{\beta}}{\rho}. \tag{7}$$

Substituting (7) into (3) we obtain

$$\frac{1}{\rho}\mathbf{K}_b\mathbf{K}_a\mathbf{K}_a^{-1}\mathbf{K}_b\boldsymbol{\beta} - \rho\mathbf{K}_b^2\boldsymbol{\beta} = 0 \tag{8}$$

which is equivalent to the generalized eigenvalue problem of the form

$$\mathbf{K}_b^2\boldsymbol{\beta} = \rho^2\mathbf{K}_b^2\boldsymbol{\beta}. \tag{9}$$

If $\mathbf{K}_b^2$ is invertible, the problem reduces to a standard eigenvalue problem of the form

$$\mathbf{I}\boldsymbol{\beta} = \rho^2\boldsymbol{\beta}. \tag{10}$$

Clearly, in the kernel space, if the Gram matrices are invertible the resulting canonical correlations are all equal to one. Regularization is therefore needed to solve the kernel CCA problem.

Question 2: Kernel CCA is CCA on Hilbert Space Objects (**3 points**)

Let the data matrices $\mathbf{X}_a$ and $\mathbf{X}_b$, of sizes $n \times p$ and $n \times q$, denote the views a and b respectively. The row vectors $\mathbf{x}_a^k \in \mathbb{R}^p$ and $\mathbf{x}_b^k \in \mathbb{R}^q$ for $k = 1, 2, \ldots, n$ denote the sets of empirical observations of X_a and X_b respectively and the column vectors $\mathbf{a}_i \in \mathbb{R}^n$ for $i = 1, 2, \ldots, p$ and $\mathbf{b}_j \in \mathbb{R}^n$ for $j = 1, 2, \ldots, q$ denote centered variable vectors of the n samples respectively. The empirical covariance matrix $\mathbf{C}_{ab}$ between the variable column vectors in $\mathbf{X}_a$ and $\mathbf{X}_b$ is $\mathbf{C}_{ab} = \mathbf{X}_a^\top\mathbf{X}_b$. The empirical variance matrices between the variables in $\mathbf{X}_a$ and in $\mathbf{X}_b$ are given by $\mathbf{C}_{aa} = \mathbf{X}_a^\top\mathbf{X}_a$ and $\mathbf{C}_{bb} = \mathbf{X}_b^\top\mathbf{X}_b$ respectively. The objective in CCA is to maximize the canonical correlation ρ between the variables in $\mathbf{X}_a$ and $\mathbf{X}_b$, obtained by transforming $\mathbf{X}_a$ and $\mathbf{X}_b$ by the vectors $\mathbf{w}_a$ and $\mathbf{w}_b$, respectively, such that the inner product, denoted by $\langle\cdot,\cdot\rangle$, between the two transformations is maximized. Hence

$$\rho = \max_{\mathbf{w}_a,\mathbf{w}_b} \frac{\mathbf{w}_a^\top\mathbf{C}_{ab}\mathbf{w}_b}{\sqrt{\mathbf{w}_a^\top\mathbf{C}_{aa}\mathbf{w}_a \cdot \mathbf{w}_b^\top\mathbf{C}_{bb}\mathbf{w}_b}} \tag{11}$$

or equivalently

$$\rho = \max_{\mathbf{w}_a,\mathbf{w}_b} \frac{\frac{1}{n}\sum_{k=1}^n \langle\mathbf{a}_k,\mathbf{w}_a\rangle\langle\mathbf{b}_k,\mathbf{w}_b\rangle}{\sqrt{\frac{1}{n}\sum_{k=1}^n \langle\mathbf{a}_k,\mathbf{w}_a\rangle\langle\mathbf{a}_k,\mathbf{w}_a\rangle}\sqrt{\frac{1}{n}\sum_{k=1}^n \langle\mathbf{b}_k,\mathbf{w}_b\rangle\langle\mathbf{b}_k,\mathbf{w}_b\rangle}}. \tag{12}$$

In kernel canonical correlation analysis (KCCA), CCA is performed by mapping the original observations $\{\mathbf{x}_a^1, \mathbf{x}_a^2, \ldots, \mathbf{x}_a^n\}$ and $\{\mathbf{x}_b^1, \mathbf{x}_b^2, \ldots, \mathbf{x}_b^n\}$ to Reproducing Kernel Hilbert Spaces (RKHS) through feature maps $\boldsymbol{\phi}_a : \mathbf{x}_a^i \mapsto \boldsymbol{\phi}_a(\mathbf{x}_a^i) \in \mathcal{H}_a$ and $\boldsymbol{\phi}_b : \mathbf{x}_b^i \mapsto \boldsymbol{\phi}_b(\mathbf{x}_b^i) \in \mathcal{H}_b$ for $i = 1, 2, \ldots, n$. The mapping to a Hilbert space ensures that the similarity of the mapped objects can be represented by symmetric positive semi-definite kernel functions $\mathbf{K} : X \times X \mapsto \mathbb{R}$ corresponding to inner products in the respective Hilbert spaces: $\mathbf{K}_a(\mathbf{x}_a^i, \mathbf{x}_a^j) = \langle\boldsymbol{\phi}_a(\mathbf{x}_a^i), \boldsymbol{\phi}_a(\mathbf{x}_a^j)\rangle_{\mathcal{H}_a}$ and $\mathbf{K}_b(\mathbf{x}_b^i, \mathbf{x}_b^j) = \langle\boldsymbol{\phi}_b(\mathbf{x}_b^i), \boldsymbol{\phi}_b(\mathbf{x}_b^j)\rangle_{\mathcal{H}_b}$.

Similarly to (12), for fixed Hilbert space objects $\mathbf{w}_a \in \mathcal{H}_a$ and $\mathbf{w}_b \in \mathcal{H}_b$, the empirical covariance of the transformations in the feature space can be written as

$$\hat{\text{cov}}(\langle \boldsymbol{\phi}_a(\mathbf{x}_a), \mathbf{w}_a \rangle, \langle \boldsymbol{\phi}_b(\mathbf{x}_b), \mathbf{w}_b \rangle) = \frac{1}{n} \sum_{k=1}^{n} \langle \boldsymbol{\phi}_a(\mathbf{x}_a^k), \mathbf{w}_a \rangle \langle \boldsymbol{\phi}_b(\mathbf{x}_b^k), \mathbf{w}_b \rangle.$$

Now, let S_a and S_b represent the linear spaces spanned by the images of the data points. For any $\mathbf{w}_a \in \mathcal{H}_a$ and $\mathbf{w}_b \in \mathcal{H}_b$, we can write $\mathbf{w}_a = \mathbf{w}_a^{\|} + \mathbf{w}_a^{\perp}$ where $\mathbf{w}_a^{\|} = \sum_{k=1}^{n} \alpha_k \boldsymbol{\phi}_a(\mathbf{x}_a^k) \in S_a$ and $\mathbf{w}_a^{\perp}$ is orthogonal to all objects $\boldsymbol{\phi}_a \in S_a$, ensuring $\langle \boldsymbol{\phi}_a, \mathbf{w}_a^{\perp} \rangle = 0$. Similarly, $\mathbf{w}_b = \mathbf{w}_b^{\|} + \mathbf{w}_b^{\perp}$ where $\mathbf{w}_b^{\|} = \sum_{k=1}^{n} \beta_k \boldsymbol{\phi}_b(\mathbf{x}_b^k) \in S_b$ and $\mathbf{w}_b^{\perp}$ is orthogonal to every object in S_b.

Show that

$$\hat{\text{cov}}(\langle \boldsymbol{\phi}_a(\mathbf{x}_a), \mathbf{w}_a \rangle, \langle \boldsymbol{\phi}_b(\mathbf{x}_b), \mathbf{w}_b \rangle) = \frac{1}{n} \boldsymbol{\alpha}^\top \mathbf{K}_a \mathbf{K}_b \boldsymbol{\beta}. \qquad \textbf{(3 points)}$$

Same holds for the variances $\hat{var}(\langle \boldsymbol{\phi}_a(\mathbf{x}_a), \mathbf{w}_a \rangle) = \frac{1}{n} \boldsymbol{\alpha}^\top \mathbf{K}_a \mathbf{K}_a \boldsymbol{\alpha}$ and $\hat{var}(\langle \boldsymbol{\phi}_b(\mathbf{x}_b), \mathbf{w}_b \rangle) = \frac{1}{n} \boldsymbol{\beta}^\top \mathbf{K}_b \mathbf{K}_b \boldsymbol{\beta}$. As a result, we can write the (K)CCA objective in dual form

$$\rho = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^\top \mathbf{K}_a \mathbf{K}_b \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^\top \mathbf{K}_a^2 \boldsymbol{\alpha} \cdot \boldsymbol{\beta}^\top \mathbf{K}_b^2 \boldsymbol{\beta}}}$$

where covariances between variables are replaced with equivalent representations expressed in terms of kernel matrices of the two views.

Solution 2:

$$\begin{aligned}
&\hat{\text{cov}}(\langle \boldsymbol{\phi}_a(\mathbf{x}_a), \mathbf{w}_a \rangle, \langle \boldsymbol{\phi}_b(\mathbf{x}_b), \mathbf{w}_b \rangle) \\
&= \frac{1}{n} \sum_{k=1}^{n} \langle \boldsymbol{\phi}_a(\mathbf{x}_a^k), \mathbf{w}_a^{\|} + \mathbf{w}_a^{\perp} \rangle \langle \boldsymbol{\phi}_b(\mathbf{x}_b^k), \mathbf{w}_b^{\|} + \mathbf{w}_b^{\perp} \rangle \\
&= \frac{1}{n} \sum_{k=1}^{n} [\langle \boldsymbol{\phi}_a(\mathbf{x}_a^k), \mathbf{w}_a^{\|} \rangle + \underbrace{\langle \boldsymbol{\phi}_a(\mathbf{x}_a^k), \mathbf{w}_a^{\perp} \rangle}_{=0}][\langle \boldsymbol{\phi}_b(\mathbf{x}_b^k), \mathbf{w}_b^{\|} \rangle + \underbrace{\langle \boldsymbol{\phi}_b(\mathbf{x}_b^k), \mathbf{w}_b^{\perp} \rangle}_{=0}] \\
&= \frac{1}{n} \sum_{k=1}^{n} \langle \boldsymbol{\phi}_a(\mathbf{x}_a^k), \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_a(\mathbf{x}_a^i) \rangle \langle \boldsymbol{\phi}_b(\mathbf{x}_b^k), \sum_{j=1}^{n} \beta_i \boldsymbol{\phi}_b(\mathbf{x}_b^j) \rangle \\
&= \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \mathbf{K}_a(\mathbf{x}_a^i, \mathbf{x}_a^k) \mathbf{K}_b(\mathbf{x}_b^j, \mathbf{x}_b^k) \beta_j \\
&= \frac{1}{n} \boldsymbol{\alpha}^\top \mathbf{K}_a \mathbf{K}_b \boldsymbol{\beta}
\end{aligned}$$

Computer Exercise

Solve the computer exercise in JupyterHub (`https://jupyter.cs.aalto.fi`). The instructions for that are given in MyCourses: `https://mycourses.aalto.fi/course/view.php?id=20602&section=3`.