Exercises for the course
Machine Learning for Data Science
Winter Semester 2024/25

G. Montavon
Institute of Computer Science
Department of Mathematics and Computer Science
Freie Universität Berlin

Exercise Sheet 6 (theory part)

Exercise 1: Canonical Correlation Analysis (15 + 5 P)

Recall: For a sample of d_1- and d_2-dimensional data of size N, given as two data matrices $X \in \mathbb{R}^{d_1 \times N}$, $Y \in \mathbb{R}^{d_2 \times N}$ (assumed to be centered), canonical correlation analysis (CCA) finds a one-dimensional projection maximizing the cross-correlation for constant auto-correlation. The optimization problem is:

$$\begin{aligned} \text{Find } w_x \in \mathbb{R}^{d_1}, w_y \in \mathbb{R}^{d_2} \text{ maximizing } \quad & w_x^\top C_{xy} w_y \\ \text{subject to } \quad & w_x^\top C_{xx} w_x &= 1 \\ & w_y^\top C_{yy} w_y &= 1, \end{aligned} \tag{1}$$

where

$$\begin{aligned} C_{xx} &= \tfrac{1}{N} X X^\top \in \mathbb{R}^{d_1 \times d_1} \qquad \text{and} \\ C_{yy} &= \tfrac{1}{N} Y Y^\top \in \mathbb{R}^{d_2 \times d_2} \end{aligned}$$

are the auto-covariance matrices of X resp. Y, and

$$C_{xy} = \tfrac{1}{N} X Y^\top \in \mathbb{R}^{d_1 \times d_2}$$

is the cross-covariance matrix of X and Y. We also define $C_{yx} = \frac{1}{N} Y X^\top = C_{xy}^\top$.

(a) *Show* that a solution of the canonical correlation analysis can be found in some eigenvector of the generalized eigenvalue problem:

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

Let $J(w) = w_x^\top C_{xy} w_y$ be the objective we want to optimize subject to the constraints above. We use the method of Lagrange multipliers to identify a set of possible solutions. We write a Lagrange function:

$$\mathcal{L}(w_x, w_y, \lambda_x, \lambda_y) = w_x^\top C_{xy} w_y + \frac{1}{2}\lambda_x (1 - w_x^\top C_{xx} w_x) + \frac{1}{2}\lambda_y (1 - w_y^\top C_{yy} w_y)$$

and solve $\nabla \mathcal{L} = 0$:

$$\frac{\partial \mathcal{L}}{\partial w_x} = C_{xy} w_y - \lambda_x C_{xx} w_x = 0 \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial w_y} = C_{yx} w_x - \lambda_y C_{yy} w_y = 0 \tag{3}$$

Multiplying (2) and (3) by $w_x^\top$ and $w_y^\top$ on both sides, we get:

$$\begin{aligned} \overbrace{w_x^\top C_{xy} w_y}^{J(w)} - \lambda_x \overbrace{w_x^\top C_{xx} w_x}^{1} &= \overbrace{w_x^\top 0}^{0} \\ \underbrace{w_y^\top C_{yx} w_x}_{J(w)} - \lambda_y \underbrace{w_y C_{yy} w_y}_{1} &= \underbrace{w_x^\top 0}_{0} \end{aligned}$$

which implies that $\lambda_x = \lambda_y \overset{(\text{def})}{=} \lambda$. With this, we can rewrite (2) and (3) in block form as:

$$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} \tag{4}$$

(b) *Show* that among all eigenvectors (w_x, w_y) the solution is the one associated to the highest eigenvalue.

We multiply Equation (4) by $(w_x, w_y)^\top$ on both sides and observe that the left hand side is our objective, and the right hand side is linear with λ:

$$\underbrace{\begin{bmatrix} w_x \\ w_y \end{bmatrix}^\top \begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}}_{2J(w)} = \lambda \underbrace{\begin{bmatrix} w_x \\ w_y \end{bmatrix}^\top \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}}_{2}$$

Hence, the eigenvalue/eigenvector pair that maximizes our objective is the one with highest associated eigenvalue.

Exercise 2: CCA for High Dimensional Data (10 + 10 + 5 + 5 P)

Like for PCA the original problem formulation involving the eigendecomposition of $d \times d$ convariance matrices does not scale well for high-dimensional data. Here, we would like to derive another formulation of the CCA problem that involves instead the eigendecomposition a matrix whose size scales with the number of data points.

(a) *Show*, that it is always possible to find an optimal solution in the span of the data, that is,

$$w_x = X\alpha_x \ , \quad w_y = Y\alpha_y$$

with some coefficient vectors $\alpha_x \in \mathbb{R}^N$ and $\alpha_y \in \mathbb{R}^N$.

Let $w_x = w_x^{(s)} + w_x^{(n)}$ be a decomposition of the solution into a component that is in the span of the data, and another component that is orthogonal to the data, i.e.

$$\begin{aligned} w_x^{(s)} &= X\alpha_x \\ X^\top w_x^{(n)} &= 0 \end{aligned}$$

Similarly for the second modality, we define $w_y = w_y^{(s)} + w_y^{(n)}$ with

$$\begin{aligned} w_y^{(s)} &= Y\alpha_y \\ Y^\top w_y^{(n)} &= 0 \end{aligned}$$

We now show that the objective and the constraints are not affected by $w_x^{(n)}, w_y^{(n)}$. For the objective, we have:

$$\begin{aligned} &w_x^\top C_{xy} w_y \\ &\quad = (w_x^{(s)} + w_x^{(n)})^\top \left(\tfrac{1}{N} XY^\top\right)(w_y^{(s)} + w_y^{(n)}) \\ &\quad = \frac{1}{N}\Big[w_x^{(s)\top} XY^\top w_y^{(s)} + w_x^{(s)\top} X \underbrace{Y^\top w_y^{(n)}}_{0} + \underbrace{w_x^{(n)\top} X}_{0} Y^\top w_y^{(s)} + \underbrace{w_x^{(n)\top} X}_{0} \underbrace{Y^\top w_y^{(n)}}_{0}\Big] \\ &\quad = w_x^{(s)\top} C_{xy} w_y^{(s)} \end{aligned}$$

and we proceed similarly to show that the constraints are not affected either, i.e. we find that

$$\begin{aligned} w_x^\top C_{xx} w_x &= w_x^{(s)\top} C_{xx} w_x^{(s)} = 1 \\ w_y^\top C_{yy} w_y &= w_y^{(s)\top} C_{yy} w_y^{(s)} = 1. \end{aligned}$$

Hence, we can always find an equivalently good solution by setting $w_x^{(n)}, w_y^{(n)}$ to zero, and using $w_x^{(s)} = X\alpha_x$ and $w_y^{(s)} = Y\alpha_y$ in the problem formulation in place of w_x and w_y.

(b) *Show* that the solution of the resulting optimization problem is found in an eigenvector of the generalized eigenvalue problem

$$\begin{bmatrix} 0 & Q_{xy} \\ Q_{yx} & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = \lambda \begin{bmatrix} Q_{xx} & 0 \\ 0 & Q_{yy} \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

where $Q_{xy} = \frac{1}{N} X^\top X Y^\top Y$, $Q_{xx} = \frac{1}{N} X^\top X X^\top X$ and $Q_{yy} = \frac{1}{N} Y^\top Y Y^\top Y$.

We replace w_x and w_y by $X\alpha_x$ and $Y\alpha_y$ in the original problem formulation:

$$\begin{aligned} \text{Find } \alpha_x \in \mathbb{R}^N, \alpha_y \in \mathbb{R}^N \text{ maximizing } \quad & (X\alpha_x)^\top C_{xy} (Y\alpha_y) \\ \text{subject to } \quad & (X\alpha_x)^\top C_{xx} (X\alpha_x) \quad = 1 \\ & (Y\alpha_y)^\top C_{yy} (Y\alpha_y) \quad = 1, \end{aligned} \tag{5}$$

and substituting C matrices by Q matrices, we get the equivalent formulation:

$$\begin{aligned} \text{Find } \alpha_x \in \mathbb{R}^N, \alpha_y \in \mathbb{R}^N \text{ maximizing } \quad & \alpha_x^\top Q_{xy} \alpha_y \\ \text{subject to } \quad & \alpha_x^\top Q_{xx} \alpha_x \quad = 1 \\ & \alpha_y^\top Q_{yy} \alpha_y \quad = 1, \end{aligned} \tag{6}$$

This has exactly the same same structure as the original CCA problem except for the variable over which we optimize and choice of matrix in each term. Applying the same steps as in Exercise 1a, we therefore get the generalized eigenvalue problem:

$$\begin{bmatrix} 0 & Q_{xy} \\ Q_{yx} & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = \lambda \begin{bmatrix} Q_{xx} & 0 \\ 0 & Q_{yy} \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

(c) *Show* that the solution is given by the eigenvector associated to the highest eigenvalue.

We proceed in the same way as in Exercise 1b, specifically, we multiply on both sides of the generalized eigenvalue problem by $(\alpha_x, \alpha_y)^\top$ and find that the eigenvector/eigenvalue pair that maximizes the objective is the one with highest eigenvalue λ.

(d) *Show* how a solution to the original CCA problem can be obtained from the solution of the latter generalized eigenvalue problem.

The solution of the original problem can be obtained simply by observing that we have expressed:

$$\begin{aligned} w_x &= X\alpha_x \\ w_y &= Y\alpha_y \end{aligned}$$

Therefore, once we have obtained α_x, α_y we can easily compute w_x, w_y.