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 7 (theory part)

Exercise 1: CCA and Least Squares Regression (20 P)

The least squares regression problem assumes an input matrix $X \in \mathbb{R}^{d \times N}$ and a vector of outputs $Y \in \mathbb{R}^N$, both centered, and has its solution given by the optimization problem:

$$\min_{\boldsymbol{v} \in \mathbb{R}^d} \quad \|X^\top \boldsymbol{v} - Y\|^2$$

The canonical correlation analysis (CCA), on the other hand. assumes an input matrix $X \in \mathbb{R}^{d_1 \times N}$ and another input matrix $Y \in \mathbb{R}^{d_2 \times N}$, both centered, and the CCA solution is the leading eigenvector of the generalized eigenvalue problem:

$$\begin{bmatrix} 0 & XY^\top \\ YX^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} = \lambda \begin{bmatrix} XX^\top & 0 \\ 0 & YY^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix}$$

We would like to relate the solutions $\boldsymbol{v}$ and $(\boldsymbol{w}_x, \boldsymbol{w}_y)$ of these two analyses for a special case.

(a) *Show* that if feeding to CCA two matrices $X \in \mathbb{R}^{d \times N}$ and $Y \in \mathbb{R}^{1 \times N}$, then the resulting vector $\boldsymbol{w}_x$ is equivalent (up to a scaling factor) to the solution $\boldsymbol{v}$ of a least square regression problem to which we feed the same matrices X and Y, with the latter represented as a vector.

Applying block-wise matrix multiplications, we can rewrite the generalized eigenvalue problem above as a system of two equations:

$$YX^\top \boldsymbol{w}_x = \lambda YY^\top \boldsymbol{w}_y$$
$$XY^\top \boldsymbol{w}_y = \lambda XX^\top \boldsymbol{w}_x$$

Taking the second equation and isolating $\boldsymbol{w}_x$, we get:

$$\boldsymbol{w}_x = \lambda^{-1}(XX^\top)^{-1} XY^\top \boldsymbol{w}_y$$
$$\boldsymbol{w}_x = \lambda^{-1}(NC_{xx})^{-1} NC_{xy} \boldsymbol{w}_y$$
$$\boldsymbol{w}_x = \lambda^{-1}(C_{xx})^{-1} C_{xy} \boldsymbol{w}_y$$

Observe that $\boldsymbol{w}_y$ is one-dimensional and is in effect just a scaling factor. Then, we get:

$$\boldsymbol{w}_x \propto (C_{xx})^{-1} C_{xy}$$

Which is the solution of linear regression.

Exercise 2: Fisher Discriminant (10 + 10 + 10 P)

The objective function to find the Fisher Discriminant has the form

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}}$$

where $\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top$ is the between-class scatter matrix and $\boldsymbol{S}_W$ is within-class scatter matrix, assumed to be positive definite. Because there are infinitely many solutions (multiplying $\boldsymbol{w}$ by a scalar doesn't change the objective), we can extend the objective with a constraint, e.g. that enforces $\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w} = 1$.

(a) *Reformulate* the problem above as an optimization problem with a quadratic objective and a quadratic constraint.

$$\max_{\boldsymbol{w}} \boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w} \qquad \text{s.t.} \quad \boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w} = 1$$

(b) *Show* using the method of Lagrange multipliers that the solution of the reformulated problem is also a solution of the generalized eigenvalue problem:

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

We first state a Lagrange function associated to the constrained optimization problem:

$$\mathcal{L}(\boldsymbol{w}, \lambda) = \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w} + \frac{1}{2} \lambda (1 - \boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w})$$

The solution of the constrained optimization problem is found where the gradient of the Lagrange function is zero:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \boldsymbol{S}_B \boldsymbol{w} - \lambda \boldsymbol{S}_W \boldsymbol{w} \overset{\text{(def)}}{=} 0$$

Therefore, we get the equation

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

which is a generalized eigenvalue problem.

(c) Show that the solution of this optimization problem is equivalent (up to a scaling factor) to

$$\boldsymbol{w} = \boldsymbol{S}_W^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

We inject the form $\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top$ into the equation:

$$(\boldsymbol{m}_2 - \boldsymbol{m}_1) \underbrace{(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top \boldsymbol{w}}_{c} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

where c is a scalar. After multiplying by $(\lambda \boldsymbol{S}_W)^{-1}$ on both sides, we get:

$$\begin{aligned} \boldsymbol{w} &= (\lambda^{-1} c) \boldsymbol{S}_W^{-1} (\boldsymbol{m}_2 - \boldsymbol{m}_1) \\ &= (-\lambda^{-1} c) \cdot \boldsymbol{S}_W^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2) \end{aligned}$$

where $(-\lambda^{-1} c)$ is a scaling factor.