Exercises for the course
Machine Learning 1
Winter semester 2024/25

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 4

Exercise 1: Lagrange Multipliers (10 + 10 P)

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ be a dataset of N data points. We consider the objective function

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{N} \|\boldsymbol{\theta} - \boldsymbol{x}_k\|^2$$

to be minimized with respect to the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. In absence of constraints, the parameter $\boldsymbol{\theta}$ that minimizes this objective is given by the empirical mean $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$. However, this is generally not the case when the parameter $\boldsymbol{\theta}$ is constrained.

(a) Using the method of Lagrange multipliers, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{\theta}^\top \boldsymbol{b} = 0$, with $\boldsymbol{b}$ some unit vector in $\mathbb{R}^d$. Give a geometrical interpretation to your solution.

(b) Using the same method, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to $\|\boldsymbol{\theta} - \boldsymbol{c}\|^2 = 1$, where $\boldsymbol{c}$ is a vector in $\mathbb{R}^d$ different from $\boldsymbol{m}$. Give a geometrical interpretation to your solution.

Exercise 2: Principal Component Analysis (10 + 10 P)

We consider a dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$. Principal component analysis searches for a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that projecting the data on that vector produces a distribution with maximum variance. Such vector can be found by solving the optimization problem:

$$\arg\max_{\boldsymbol{u}} \; \frac{1}{N}\sum_{k=1}^{N}\Big[\boldsymbol{u}^\top \boldsymbol{x}_k - \frac{1}{N}\Big(\sum_{l=1}^{N} \boldsymbol{u}^\top \boldsymbol{x}_l\Big)\Big]^2 \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

(a) *Show* that the problem above can be rewritten as

$$\arg\max_{\boldsymbol{u}} \; \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

where $\boldsymbol{S} = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^\top$ is the scatter matrix, and $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$ is the empirical mean.

(b) *Show* using the method of Lagrange multipliers that the problem above can be reformulated as solving the eigenvalue problem

$$\boldsymbol{S}\boldsymbol{u} = \lambda \boldsymbol{u}$$

and retaining the eigenvector $\boldsymbol{u}$ associated to the highest eigenvalue λ.

Exercise 3: Bounds on Eigenvalues (5 + 5 + 5 + 5 P)

Let λ_1 denote the largest eigenvalue of the matrix $\boldsymbol{S}$. The eigenvalue λ_1 quantifies the variance of the data when projected on the first principal component. Because its computation can be expensive, we study how the latter can be bounded with the diagonal elements of the matrix $\boldsymbol{S}$.

(a) *Show* that $\sum_{i=1}^{d} \boldsymbol{S}_{ii}$ is an upper bound to the eigenvalue λ_1.

(b) *State* the conditions on the data for which the upper bound is tight.

(c) *Show* that $\max_{i=1}^{d} \boldsymbol{S}_{ii}$ is a lower bound to the eigenvalue λ_1.

(d) *State* the conditions on the data for which the lower bound is tight.

Exercise 4: Iterative PCA (10 P)

When performing principal component analysis, computing the full eigendecomposition of the scatter matrix $\boldsymbol{S}$ is typically slow, and we are often only interested in the first principal components. An efficient procedure to find the first principal component is *power iteration*. It starts with a random unit vector $\boldsymbol{w}^{(0)} \in \mathbb{R}^d$, and iteratively applies the parameter update

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{S}\boldsymbol{w}^{(t)} \big/ \|\boldsymbol{S}\boldsymbol{w}^{(t)}\|$$

until some convergence criterion is met. Here, we would like to show the exponential convergence of power iteration. For this, we look at the error terms

$$\mathcal{E}_k(\boldsymbol{w}) = \left| \frac{\boldsymbol{w}^\top \boldsymbol{u}_k}{\boldsymbol{w}^\top \boldsymbol{u}_1} \right| \qquad \text{with} \quad k = 2, \ldots, d,$$

and observe that they should all converge to zero as $\boldsymbol{w}$ approaches the eigenvector $\boldsymbol{u}_1$ and becomes orthogonal to other eigenvectors.

(a) Show that $\mathcal{E}_k(\boldsymbol{w}^{(T)}) = |\lambda_k/\lambda_1|^T \cdot \mathcal{E}_k(\boldsymbol{w}^{(0)})$, i.e. the convergence of the algorithm is exponential with the number of time steps T.

Exercise 5: Programming (30 P)

Download the programming files on ISIS and follow the instructions.

Exercise 1: Lagrange Multipliers (10 + 10 P)

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ be a dataset of N data points. We consider the objective function

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{N} \|\boldsymbol{\theta} - \boldsymbol{x}_k\|^2$$

to be minimized with respect to the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. In absence of constraints, the parameter $\boldsymbol{\theta}$ that minimizes this objective is given by the empirical mean $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$. However, this is generally not the case when the parameter $\boldsymbol{\theta}$ is constrained.

(a) Using the method of Lagrange multipliers, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{\theta}^\top \boldsymbol{b} = 0$, with $\boldsymbol{b}$ some unit vector in $\mathbb{R}^d$. Give a geometrical interpretation to your solution.

Solution:

First the objective function can be writen as

$$J(\theta) = \sum_{k=1}^{N} \|\theta - x_k\|^2$$

not related to x_k

$$= \sum_{k=1}^{N} \theta^T\theta - 2\theta^T x_k + x_k^T x_k$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad \sum_{k=1}^{N} \theta^T\theta - 2\theta^T x_k$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad N\theta^T\theta - 2\theta^T \sum_{k=1}^{N} x_k$$

$$\Downarrow$$

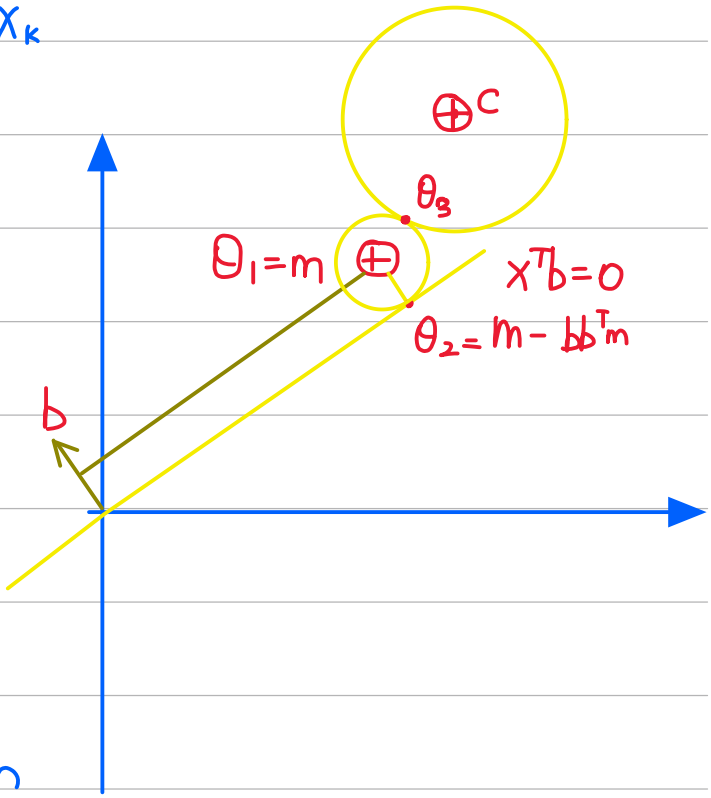
$$\underset{\theta}{\operatorname{argmin}} \quad \theta^T\theta - 2\theta^T m$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad \theta^T\theta - 2\theta^T m + m^T m$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad \|\theta - m\|^2 \qquad \theta_1^* = m$$



So when this objective is constraint by $\theta^T b = 0$ using Lagrange multiplier we can form a Lagrange function:

日期: /

$$\max_{\lambda}\min_{\theta} \mathcal{L}(\theta,\lambda) = \frac{1}{2}\|\theta-m\|^2 + \lambda\theta^T b$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \theta^T} = (\theta - m) + \lambda b = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = \theta^T b = 0 \end{cases}$$

$$\theta - m + \lambda b = 0 \qquad ①$$

$$\therefore \quad \theta^* = m - \lambda b$$

Then we need to find the expression of λ.

let ① be left multiplied by b^T.

$$b^T\theta - b^T m + \lambda b^T b = 0$$

Since $b^T\theta = 0$ and b is a unit vector, i.e. $b^T b = 1$

$$\therefore \quad \lambda = b^T m$$

$$\therefore \quad \theta_2^* = m - b^T m b$$

$$= m - b b^T m$$

(b) Using the same method, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to $\|\boldsymbol{\theta} - \boldsymbol{c}\|^2 = 1$, where $\boldsymbol{c}$ is a vector in $\mathbb{R}^d$ different from $\boldsymbol{m}$. Give a geometrical interpretation to your solution.

Solution:

The Lagrange function is

$$\mathcal{L}(\theta,\lambda) = \frac{1}{2}\|\theta - m\|^2 + \frac{1}{2}\lambda\left(\|\theta - c\|^2 - 1\right)$$

日期: /

$$\frac{\partial \mathcal{L}}{\partial \theta^T} = (\theta - m) + \lambda(\theta - c) = 0$$

$$\therefore \quad (\theta - c) + \lambda(\theta - c) = m - c$$

$$(1+\lambda)(\theta - c) = m - c \longrightarrow \theta = c + \frac{m-c}{1+\lambda}$$

$$\therefore \quad \Downarrow \text{ take 2nd Norm}$$

$$(1+\lambda)^2 \underbrace{\|\theta - c\|^2}_{=1} = \|m - c\|^2$$

$$\Downarrow$$

$$(1+\lambda)^2 = \|m - c\|^2$$

$$\therefore \quad \lambda = -1 \pm \|m - c\|$$

$$\therefore \quad \theta^* = c \pm \frac{m-c}{\|m-c\|}$$

We can compare the objective value:

$$J\left(c + \frac{m-c}{\|m-c\|}\right) = \left\| c + \frac{m-c}{\|m-c\|} - m \right\|^2 \longrightarrow \left\| -(m-c) + \frac{m-c}{\|m-c\|} \right\|^2$$

$$J\left(c - \frac{m-c}{\|m-c\|}\right) = \left\| c - \frac{m-c}{\|m-c\|} - m \right\|^2 \longrightarrow \left\| -(m-c) - \frac{m-c}{\|m-c\|} \right\|^2$$

$$\left\| (m-c) \times \left(\frac{1}{\|m-c\|} - 1 \right) \right\|^2 \leq \left\| (m-c) \times \left(-\frac{1}{\|m-c\|} - 1 \right) \right\|^2$$

$$\Downarrow$$

$$\theta_3^* = c + \frac{m-c}{\|m-c\|}$$

Exercise 2: Principal Component Analysis (10 + 10 P)

We consider a dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$. Principal component analysis searches for a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that projecting the data on that vector produces a distribution with maximum variance. Such vector can be found by solving the optimization problem:

$$\arg\max_{\boldsymbol{u}} \frac{1}{N}\sum_{k=1}^{N}\Big[\boldsymbol{u}^\top \boldsymbol{x}_k - \frac{1}{N}\Big(\sum_{l=1}^{N}\boldsymbol{u}^\top \boldsymbol{x}_l\Big)\Big]^2 \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

(a) *Show* that the problem above can be rewritten as

$$\arg\max_{\boldsymbol{u}} \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

where $\boldsymbol{S} = \sum_{k=1}^{N}(\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^\top$ is the scatter matrix, and $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k$ is the empirical mean.

Solution:

We can derive the function as.

$$\frac{1}{N}\sum_{k=1}^{N}\Big[u^T x_k - \frac{1}{N}\Big(\sum_{l=1}^{N} u^T x_l\Big)\Big]^2$$

$$= \frac{1}{N}\sum_{k=1}^{N}\Big[u^T x_k - u^T m\Big]^2$$

$$= \frac{1}{N}\sum_{k=1}^{N}(u^T x_k - u^T m)^T(u^T x_k - u^T m)$$

$$= \frac{1}{N}\sum_{k=1}^{N}\big(u^T(x_k - m)\big)^T\big(u^T(x_k - m)\big)$$

$$= \frac{1}{N}\sum_{k=1}^{N}(x_k - m)^T u u^T (x_k - m)$$

$$= \frac{1}{N}\sum_{k=1}^{N} u^T (x_k - m)(x_k - m)^T u$$

$$\Downarrow \text{ delete } \frac{1}{N}$$

$$u^T\Big(\sum_{k=1}^{N}(x_k - m)(x_k - m)^T\Big)u$$

$$\Downarrow$$

$$\boxed{\arg\max_{u} u^T S u \quad \text{with} \quad \|u\|^2 = 1}$$

(b) *Show* using the method of Lagrange multipliers that the problem above can be reformulated as solving the eigenvalue problem

$$\boldsymbol{S}\boldsymbol{u} = \lambda \boldsymbol{u}$$

and retaining the eigenvector $\boldsymbol{u}$ associated to the highest eigenvalue λ.

Solution:

We can form the Lagrange function as:

$$\mathcal{L}(\lambda, u) = u^T S u - \lambda(\|u\|^2 - 1)$$

$$\frac{\partial \mathcal{L}}{\partial u^T} = Su - \lambda u = 0$$

$$\therefore \quad Su = \lambda u \qquad (1)$$

So we have proofed that it can be solved by solving an Eigenvalue problem.

Then we left multiply u^T on both sides.

$$u^T S u = \lambda u^T u = \lambda \|u\|^2 = \lambda$$

$$\therefore \quad \underset{u}{\operatorname{argmax}}\ u^T S u \iff \underset{\lambda}{\operatorname{argmax}}\ \lambda$$

$\therefore$ the optimal results is the eigenvector u^* that correspond to the largest eigenvalue λ_1.

Exercise 3: Bounds on Eigenvalues (5 + 5 + 5 + 5 P)

Let λ_1 denote the largest eigenvalue of the matrix $\boldsymbol{S}$. The eigenvalue λ_1 quantifies the variance of the data when projected on the first principal component. Because its computation can be expensive, we study how the latter can be bounded with the diagonal elements of the matrix $\boldsymbol{S}$.

(a) *Show* that $\sum_{i=1}^{d} \boldsymbol{S}_{ii}$ is an upper bound to the eigenvalue λ_1.

Solution:

$$\lambda_i = u_i^T S u_i = \sum_{k=1}^{N} (u^T(x_k - m))^2 \geq 0$$

$$\sum_i S_{ii} = \text{trace}(S)$$

From Matrix cookbook, Page 6. formular (12)
we have:

$$\sum_i S_{ii} = \text{trace}(S) = \sum_i \lambda_i \geq \lambda_1$$

∴ the largest eigenvalue is upperbounded by $\sum_i S_{ii}$

(b) *State* the conditions on the data for which the upper bound is tight.

Solution:

When all other eigenvalues $\lambda_2, \lambda_3 \cdots \lambda_d$ are 0.
Then

$$\sum_i S_{ii} = \lambda_i$$

(c) *Show* that $\max_{i=1}^{d} \boldsymbol{S}_{ii}$ is a lower bound to the eigenvalue λ_1.

Solution:

every element ≥ 0

$$\lambda = \max_{\|u\|=1} u^T S u \geq \max_{u \in \{e_1 \cdots e_d\}} u^T S u = \max_{i=1 \cdots d} e_i^T S e_i$$

$$= \{e_1^T S e_1, e_2^T S e_2 \cdots e_d^T S e_d\}$$

[$e_1 \cdots e_d$ are canonical coordinate vectors.]

$$\therefore \quad \lambda \geq \max_{i=1}^{d} S_{ii}$$

(d) *State* the conditions on the data for which the lower bound is tight.

Solution:

If we want $\lambda = \max_{i=1}^{d} S_{ii} = \max_{i=1}^{d} e_i^T S e_i$

then $u \in \{e_1 \cdots e_d\}$

So that it can pick out the largest diagonal element

Exercise 4: Iterative PCA (10 P)

When performing principal component analysis, computing the full eigendecomposition of the scatter matrix $\boldsymbol{S}$ is typically slow, and we are often only interested in the first principal components. An efficient procedure to find the first principal component is *power iteration*. It starts with a random unit vector $\boldsymbol{w}^{(0)} \in \mathbb{R}^d$, and iteratively applies the parameter update

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{S}\boldsymbol{w}^{(t)} / \|\boldsymbol{S}\boldsymbol{w}^{(t)}\|$$

until some convergence criterion is met. Here, we would like to show the exponential convergence of power iteration. For this, we look at the error terms

$$\mathcal{E}_k(\boldsymbol{w}) = \left| \frac{\boldsymbol{w}^\top \boldsymbol{u}_k}{\boldsymbol{w}^\top \boldsymbol{u}_1} \right| \qquad \text{with} \quad k = 2, \ldots, d,$$

and observe that they should all converge to zero as $\boldsymbol{w}$ approaches the eigenvector $\boldsymbol{u}_1$ and becomes orthogonal to other eigenvectors.

(a) Show that $\mathcal{E}_k(\boldsymbol{w}^{(T)}) = |\lambda_k/\lambda_1|^T \cdot \mathcal{E}_k(\boldsymbol{w}^{(0)})$, i.e. the convergence of the algorithm is exponential with the number of time steps T.

Solution:

$$\mathcal{E}_k(w^{(T+1)}) = \left| \frac{w^{(T+1)^T} u_k}{w^{(T+1)^T} u_1} \right|$$

$$= \left| \frac{\frac{(Sw^{(T)})^T}{\|Sw^{(T)}\|} u_k}{\frac{(Sw^{(T)})^T}{\|Sw^{(T)}\|} u_1} \right| = \left| \frac{(Sw^{(T)})^T u_k}{(Sw^{(T)})^T u_1} \right|$$

Since $S = S^T$

$$= \left| \frac{w^{(T)^T} S u_k}{w^{(T)^T} S u_1} \right| = \left| \frac{w^{(T)^T} \lambda_k u_k}{w^{(T)^T} \lambda_1 u_1} \right|$$

$$= \left| \frac{w^{(T)^T} u_k}{w^{(T)^T} u_1} \right| \cdot \left| \frac{\lambda_k}{\lambda_1} \right|$$

$$= \mathcal{E}_k(w^{(T)}) \left| \frac{\lambda_k}{\lambda_1} \right| = \mathcal{E}_k(w^{(T-1)}) \left| \frac{\lambda_k}{\lambda_1} \right|^2$$

$$= \mathcal{E}_k(w^{(0)}) \left| \frac{\lambda_k}{\lambda_1} \right|^T$$