Exercises for the course

Machine Learning 1

Winter semester 2020/21

Abteilung 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 3

Exercise 1: Lagrange Multipliers (10 + 10 P)

Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ 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 $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$. 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

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

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. *(Hint: to show this, it is useful to rewrite the scatter matrix in terms of eigenvalues and eigenvectors, i.e.* $\boldsymbol{S} = \sum_{i=1}^d \boldsymbol{u}_i \boldsymbol{u}_i^\top \lambda_i$*.)*

Exercise 5: Programming (30 P)

Download the programming files on ISIS and follow the instructions.

$\boxed{1}$ a) $f(\vec{\theta}) = \sum_{k=1}^{N} \|\vec{\theta} - \vec{x}_k\|^2$

Without contraint: $\underset{\theta}{\operatorname{argmin}} f(\vec{\theta}) \Rightarrow \vec{m} = \frac{1}{N} \sum_{k=1}^{N} \vec{x}_k = \vec{\theta}$

With contraint $\vec{\theta}^T \vec{b} = 0$: Application of Lagrange multiplier.

$$\frac{\partial \mathcal{L}(\vec{\theta}, \lambda)}{\partial \vec{\theta}} = \frac{\partial}{\partial \theta}\left(f(\vec{\theta}) + \lambda g(\vec{\theta})\right) = \frac{\partial}{\partial \theta}\left(f(\vec{\theta}) + \lambda \vec{\theta}^T b\right) = \frac{\partial}{\partial \theta} \sum_{k=1}^{N} \|\vec{\theta} - \vec{x}_k\|^2 + \lambda \vec{\theta} \vec{b}$$

$$= \frac{\partial}{\partial \theta} \sum_{k=1}^{N} \|\vec{\theta}\|^2 - 2\vec{\theta}^T \vec{x}_k + \|\vec{x}_k\|^2 + \lambda \vec{\theta} \vec{b} = \frac{\partial}{\partial \theta} \|\vec{\theta}\|^2 - 2\vec{\theta}^T \vec{m} + \|\vec{m}\|^2 + \lambda \vec{\theta} \vec{b}$$

$$= \frac{\partial}{\partial \vec{\theta}} \|\theta - m\|^2 + \lambda \vec{\theta} \vec{b} = 2(\vec{\theta} - \vec{m}) + \lambda \vec{b} \overset{!}{=} 0$$

$$\Rightarrow \vec{\theta} = \frac{2\vec{m} - \lambda \vec{b}}{2}$$

$$2\vec{\theta} - 2\vec{m} + \lambda \vec{b} = 0 \Leftrightarrow 2\underbrace{\vec{b}^T \vec{\theta}}_{0} - 2\vec{b}^T \vec{m} + \lambda \underbrace{\vec{b}^T \vec{b}}_{1} = \underbrace{\vec{b}\vec{0}}_{0}$$

$$\Leftrightarrow -2\vec{b}^T \vec{m} + \lambda = 0 \Leftrightarrow \lambda = 2\vec{b}^T \vec{m}$$

$$\Rightarrow \theta = \frac{2\vec{m} - \lambda \vec{b}}{2} = \frac{2\vec{m} - 2\vec{b}^T \vec{m}\, \vec{b}}{2} = \vec{m} - \vec{b}\vec{b}^T \vec{m}$$

The optimized parameter θ corresponds to the mean $\vec{m}$ with the subtracted projection of the mean on the vector $\vec{b}$.

b) $\mathcal{L}(\vec{\theta},\lambda) = J(\vec{\theta}) + \lambda\, g(\vec{\theta}) = \|\vec{\theta}-\vec{m}\| + \lambda \|\vec{\theta}-\vec{c}\|^2 - \lambda$

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

$$= \cancel{2}(\vec{\theta}-\vec{m}) + \cancel{2}\lambda(\vec{\theta}-\vec{c}) \overset{!}{=} 0$$

$$\Leftrightarrow (1+\lambda)(\vec{\theta}-\vec{c}) - (\vec{m}-\vec{c}) = 0$$

1) $\Leftrightarrow (1+\lambda)(\vec{\theta}-\vec{c}) = \vec{m}-\vec{c}$

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

2) $\Leftrightarrow |1+\lambda| = \pm\|\vec{m}-\vec{c}\|$

2) in 1): $\pm\|\vec{m}-\vec{c}\|(\vec{\theta}-\vec{c}) = \vec{m}-\vec{c}$

$$\Leftrightarrow \vec{\theta} = \vec{c} \pm \frac{\vec{m}-\vec{c}}{\|\vec{m}-\vec{c}\|}$$

Only one of these two solutions minimizes the objective $J(\vec{\theta})$.
The constraint $g(\vec{\theta}) = \|\vec{\theta}-\vec{c}\|^2 - 1 = 0$ corresponds to a circle ~~around~~ with center $\vec{c}$. One solution maximizes the distance to the mean $\vec{m}$, the other one minimizes it. But both solution lie on the same circle.

$\boxed{2}$

a) $\underset{u}{\arg\max} \frac{1}{N} \sum_{k=1}^{N} \left[\vec{u}^T \vec{x}_k - \frac{1}{N} \left(\sum_{l=1}^{N} \vec{u}^T \vec{x}_l \right) \right]^2$

1) $\frac{1}{N} \sum_{l=1}^{N} \vec{x}_l = \vec{m}$

$$\overset{1)}{=} \underset{u}{\arg\max} \frac{1}{N} \sum_{k=1}^{N} \left[\vec{u}^T \vec{x}_k - \vec{u}^T \cdot \vec{m} \right]^2$$

$$= \underset{u}{\arg\max} \frac{1}{N} \sum_{k=1}^{N} (\vec{u}^T \vec{x}_k - \vec{u}^T \cdot \vec{m})(\vec{u}^T \vec{x}_k - \vec{u}^T \vec{m})^T$$

$$= \underset{u}{\arg\max} \frac{1}{N} \sum_{k=1}^{N} \vec{u}^T (\vec{x}_k - \vec{m})(\vec{x}_k - \vec{m})^T \vec{u}$$

$$= \underset{u}{\arg\max} \frac{1}{N} \vec{u}^T \vec{S} \vec{u}$$

$$= \arg\max \vec{u}^T \vec{S} \vec{u} \quad \square$$

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

$$\frac{\partial \mathcal{L}(\vec{u}, \vec{S})}{\partial \vec{u}} = \vec{S} \vec{u} + \vec{S}^T \vec{u} \mp 2 \lambda \vec{u}$$

$$\overset{\vec{S} = \vec{S}^T}{=} 2 \cdot \vec{S} \vec{u} \mp 2 \lambda \vec{u} \overset{!}{=} 0$$

1) $\Leftrightarrow \vec{S} \vec{u} = + \lambda \vec{u}$

In a) we have $\underset{u}{\arg\max}\, \vec{u}^T \vec{S} \vec{u}$ which becomes

$\underset{u}{\arg\max}\, \vec{u}^T \vec{S} \vec{u} = \underset{u}{\arg\max}\, \vec{u}^T \lambda \vec{u}$ using 1)

$= \underset{u}{\arg\max}\, \underbrace{\vec{u}^T \vec{u}}_{=1} \lambda = \underset{u}{\arg\max}\, \lambda = \max \lambda_i$

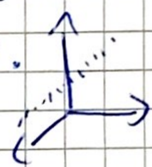
which corresponds to eigenvector $\vec{u}$ with maximum eigenvalue λ_i

$\boxed{3}$

a) $\sum_{i=1}^{d} S_{ii} = tr(\vec{S}) \geq \max_i \lambda_i$ since the trace of a matrix corresponds to the sum of its eigenvalues:

$$tr(\vec{A}) = \sum_{i=1}^{N} \lambda_i$$

b) The condition is tight if $\sum_{i=1}^{N} \lambda_i = \max_i \lambda_i$: the sum of all eigenvalues corresponds to the maximum eigenvalue. In that case all other eigenvalues must be 0. This is the case if the data is one-dimensional.



c) According to the previous exercise

$$\lambda_1 = \vec{u}_1^T \vec{S} \vec{u}_1 \text{ which corresponds to } \arg\max_u \vec{u}^T \vec{S} \vec{u}$$

This maximized $\vec{u}$ vector can always be represented as a of unit vectors:
Therefore, $\arg\max_u \vec{u}^T \vec{S} \vec{u} \geq \arg\max_i^d \vec{e}_i^T \vec{S} \vec{e}_i = \arg\max_i^d S_{ii}$
since $\vec{e}_i^T$ and $\vec{e}_i$ select the rows and columns of $\vec{S}$.

d) The vector $\vec{u}$ can always be represented as a linear combination of unit vectors: $\alpha_1 \vec{e}_1 + \alpha_2 \vec{e}_2 + \dots + \alpha_d \vec{e}_d$
If ~~the~~ $\vec{u}$ corresponds to a unit vector, meaning $\forall \alpha_i = 0$ except for one α_i, then the bound is tight.

$\boxed{4}$ a)

$$\varepsilon_k(w^{(t+1)}) = \left|\frac{w^{(t+1)} \mu_k}{w^{(t+1)} \mu_1}\right|$$

$$= \left|\frac{(S w^{(t)} / \cancel{\|S w^{(t)}\|})^T \cdot \mu_k}{(S w^{(t)} / \cancel{\|S w^{(t)}\|})^T \cdot \mu_1}\right|$$

$$= \left|\frac{w^{(t)T} \cdot S \cdot \mu_k}{w^{(t)T} S \cdot \mu_1}\right| = \left|\frac{w^{(t)T}\left(\sum_{i=1}^{d} \mu_i \overbrace{\mu_i^T \mu_k}^{\delta_{ik}} \lambda_i\right)}{w^{(t)T}\left(\sum_{i=1}^{d} \mu_i \underbrace{\mu_i^T \mu_1}_{\delta_{i1}} \lambda_i\right)}\right|$$

$$= \left|\frac{w^{(t)T} \mu_k \lambda_k}{w^{(t)T} \mu_1 \lambda_1}\right| = \underbrace{\left|\frac{w^{(t)T} \mu_k}{w^{(t)T} \mu_1}\right|}_{\varepsilon_k} \left|\frac{\lambda_k}{\lambda_1}\right| = \left|\frac{\lambda_k}{\lambda_1}\right|^T \cdot \varepsilon_k(w^{(T)}$$

$\square$