Prof. Dr. Jan Rolfes, Ehsan Waiezi, Lars Weidne

Mathematics of Learning – Worksheet 4 - – Discussion on Novermber 14/15th, 2023

- The exercise sheets will be uploaded every Monday. Solution sketches will be uploaded one week later.
- You can hand in your own solutions via StudOn and we correct them - this is not mandatory. Please hand in small groups of 2-3 students.
- For questions, please use the forum on StudOn since other students may have similar questions. If you have a more personal question about the exercises please send an email to ehsan.waiezi@fau.de or lars.weidner@fau.de respectively.

Basics [Eigenvectors of symmetric matrices.] Consider a symmetric matrix $A \in \mathbb{R}^{n \times n}$ (i.e. $A = A^T$). Prove that for two eigenvalues $\lambda_1, \lambda_2 \in \mathbb{R}$ with $\lambda_1 \neq \lambda_2$ and corresponding eigenvectors $v_1 \in \mathbb{R}^n$ and $v_2 \in \mathbb{R}^n$ it holds that $\langle v_1, v_2 \rangle = 0$, i.e., that v_1 and v_2 are orthogonal.

Solution.

$$\lambda_1 \langle v_1, v_2 \rangle = \langle \lambda_1 v_1, v_2 \rangle = \langle A v_1, v_2 \rangle \overset{\text{Symmetry}}{=} \langle v_1, A v_2 \rangle = \langle v_1, \lambda_2 v_2 \rangle = \lambda_2 \langle v_1, v_2 \rangle$$

Since $\lambda_1 \neq \lambda_2$, this implies that the inner product is zero.

Basics. [Singular Value Decomposition].
Calculate the SVD of the matrix

$$A = \begin{pmatrix} 2 & -2 & -2 & 0 \\ -1 & -1 & 3 & 4 \\ 2 & -2 & 2 & -2 \end{pmatrix}.$$

The SVD of a matrix $A \in \mathbb{R}^{m \times n}$ consists of matrices $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ which are orthogonal, i.e., UU^T, VV^T is equal to the corresponding unit matrix, and a matrix $\Sigma \in \mathbb{R}^{m \times n}$ which has only positive, descending entries on the diagonal (it looks like

$$\begin{pmatrix} * & 0 & 0 & 0 \\ 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \end{pmatrix}$$

in our case). $U \cdot \Sigma \cdot V^T = A$ should hold, in case you calculated correctly. Hint: No nice numbers this time.

Solution. SVDs are calculated computationally by sophisticated numerical algorithms, which also work efficient in the case the matrix A is large or ill-conditioned. A possibility to calculate the SVD manually on a sheet of paper is the following: U can be chosen to be the matrix of (normalized) eigenvectors of AA^T. The vectors have to be sorted according to the size of the corresponding eigenvalues. If v is a normalized eigenvector of AA^T, it is not clear beforehand if to take v or $-v$ since both would be normalized eigenvectors. Analogously, V can be chosen to be the matrix of normalized eigenvectors of A^TA. Hence, U and V are equal to (for example, every column might be multiplied by -1)

$$U = \begin{pmatrix} 0.3559 & 0.3110 & -0.8813 \\ -0.9050 & 0.3499 & -0.2420 \\ 0.2331 & 0.8837 & 0.4060 \end{pmatrix}, \quad V = \begin{pmatrix} -0.3811 & 0.5003 & 0.2429 & -0.7385 \\ 0.0499 & -0.6720 & -0.4088 & -0.6155 \\ 0.5417 & 0.5385 & -0.6336 & -0.1231 \\ 0.7476 & -0.0902 & 0.6102 & -0.2462 \end{pmatrix}$$

Σ is a "diagonal" matrix with the squareroots of the eigenvalues of A^TA (or AA^T, respectively; it is the same).

$$\Sigma = \begin{pmatrix} 5.4658 & 0 & 0 & 0 \\ 0 & 4.0762 & 0 & 0 \\ 0 & 0 & 2.9171 & 0 \end{pmatrix}$$

Calculating the product $U\Sigma V^T$, we see that it is not equal to A. We have to set the signs right, using the identity (u_k denoting the k-th column of U, and v_k denoting the k-th column of V, and σ_k denoting the k-th singular value):

$$u_k = \sigma_k^{-1} A v_k$$

Hence, we derive (and technically, this makes it useless that we calculated a candidate for U beforehand)

$$U = \begin{pmatrix} -0.3559 & 0.3110 & 0.8813 \\ 0.9050 & 0.3499 & 0.2420 \\ -0.2331 & 0.8837 & -0.4060 \end{pmatrix}$$

Multiplying $U\Sigma V^T$ we get A.

Exercise 1 [Reading assignment: Association rules].
Read chapter 14.2 of the *Hastie* book, regarding association rules. Discuss the contents with one (or more) fellow student for at least half an hour.

Exercise 2 [Reading assignment: Self organizing maps (SOM)].
Read chapter 14.5 of the *Hastie* book (or 14.4, depending on your version), regarding self organizing maps. Discuss the contents with one (or more) fellow student for at least half an hour.

Exercise 3 [Equivalence of eigenvalue problems].
Let $x^{(1)}, \ldots, x^{(N)}$ be given input data. Furhermore, let $\mathcal{H}$ be a (possibly infinite-dimensional) Hilbert space, $\Psi\colon \mathbb{R}^M \to \mathcal{H}$ a map from the input data, $\mathbf{C}$ the covariance matrix of the transformed data in $\mathcal{H}$ with:

$$\mathbf{C} := \frac{1}{N} \sum_{i=1}^{N} \Psi(x^{(i)}) \Psi(x^{(i)})^T, \tag{1}$$

Furthermore, let $k\colon \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ be the corresponding kernel function and K the associated Kernel matrix with $K_{i,j} = k(x^{(i)}, x^{(j)})$.

1. Show that for any $\lambda \neq 0$ every solution $\vec{\alpha} \in \mathbb{R}^N$ with $\vec{\alpha} \perp \operatorname{kern}(K)$ of the equation

$$N\lambda K\vec{\alpha} = K^2\vec{\alpha} \tag{2}$$

is also a solution of the eigenvalue equation:

$$N\lambda\vec{\alpha} = K\vec{\alpha}. \tag{3}$$

Solution. We proof the statement by contradiction. For this let us assume that there exists a vector $\vec{\alpha} \in \mathbb{R}^N$ with $\vec{\alpha} \perp \operatorname{kern} K$ that solves (2) but not (3). Then we now that:

$$N\lambda\vec{\alpha} - K\vec{\alpha} = \vec{\beta} \neq \vec{0}.$$

We can rewrite (2) and see that

$$\vec{0} = N\lambda K\vec{\alpha} - K^2\vec{\alpha} = K(N\lambda\vec{\alpha} - K\vec{\alpha}) = K\vec{\beta}.$$

Thus, we know that $\vec{\beta} \in \operatorname{kern}(K)$ and thus $\langle \vec{\alpha}, \vec{\beta} \rangle = 0$ since $\vec{\alpha} \perp \operatorname{kern} K$. On the other hand, we can write:

$$0 < \langle \vec{\beta}, \vec{\beta} \rangle = \langle N\lambda\vec{\alpha} - K\vec{\alpha}, \beta \rangle = N\lambda \underbrace{\langle \vec{\alpha}, \vec{\beta} \rangle}_{=0} - \langle K\vec{\alpha}, \beta \rangle = -\langle \vec{\alpha}, \underbrace{K\vec{\beta}}_{=\vec{0}} \rangle = 0.$$

This is clearly a contradiction, which proves the original statement.

2. Use the previous statement to show that the following "equivalence" holds for all $\lambda > 0$

   (a) $\mathbf{v} \in \mathcal{H}$ is eigenvector of $\mathbf{C}$ with respect to eigenvalue $\lambda \Rightarrow$
   $\vec{\alpha} \in \mathbb{R}^N$ defined such that $\mathbf{v} = \sum_{i=1}^N \alpha_i \Psi(x^{(i)})$ is eigenvector of K with respect to eigenvalue $N\lambda$

   (b) $\vec{\alpha} \in \mathbb{R}^N$ is eigenvector of K with respect to eigenvalue $N\lambda \Rightarrow$
   $\mathbf{v} = \sum_{i=1}^N \alpha_i \Psi(x^{(i)}) \in \mathcal{H}$ is eigenvector of $\mathbf{C}$ with respect to eigenvalue λ

   Solution. To prove the statement we have to show both directions separately.

   (a) We show the first direction of this equivalence by assuming that the eigenvalue equation in $\mathcal{H}$ holds:

   $$\lambda\mathbf{v} = \mathbf{C}\mathbf{v}, \text{ for } \lambda \neq 0.$$

   From

   $$\lambda\mathbf{v} = \mathbf{C}\mathbf{v} := \frac{1}{N}\sum_{i=1}^N \Psi(x^{(i)}) \langle \Psi(x^{(i)}), \mathbf{v} \rangle$$

   we see that $\mathbf{v} \in \operatorname{span}(\Psi(x^{(1)}), \ldots, \Psi(x^{(N)}))$ and hence there exists a vector of coefficients $\vec{\alpha} = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ such that:

   $$\mathbf{v} = \sum_{i=1}^N \alpha_i \Psi(x^{(i)}).$$

We define a linear operator $\Psi^T\colon \mathcal{H} \to \mathbb{R}^N$ with $(\Psi^T \mathbf{u})_i = \langle \Psi(x^{(i)}), \mathbf{u} \rangle$ for $i = 1, \dots, N$. Using this notation we apply this operator to the let side of the eigenvalue equation and can deduce for all $i = 1, \dots, N$:

$$
\begin{aligned}
(\Psi^T \lambda \mathbf{v})_i = \lambda \langle \Psi(x^{(i)}), \mathbf{v} \rangle = \lambda \langle \Psi(x^{(i)}), \sum_{j=1}^{N} \alpha_j \Psi(x^{(j)}) \rangle \\
= \lambda \sum_{j=1}^{N} \alpha_j \underbrace{\langle \Psi(x^{(i)}), \Psi(x^{(j)}) \rangle}_{=K_{i,j}} = \lambda (K\vec{\alpha})_i
\end{aligned}
$$

This means that $\Psi^T \lambda \mathbf{v} = \lambda K \vec{\alpha}$. Applying the same transformation on the right side of the eigenvalue equation we get:

$$
\begin{aligned}
(\Psi^T \mathbf{C} \mathbf{v})_i &= \langle \Psi(x^{(i)}), \mathbf{C}\mathbf{v} \rangle = \left\langle \Psi(x^{(i)}), \frac{1}{N} \sum_{j=1}^{N} \Psi(x^{(j)}) \langle \Psi(x^{(j)}), \mathbf{v} \rangle \right\rangle \\
&= \frac{1}{N} \sum_{j=1}^{N} \left\langle \Psi(x^{(i)}), \Psi(x^{(j)}) \langle \Psi(x^{(j)}), \mathbf{v} \rangle \right\rangle \\
&= \frac{1}{N} \sum_{j=1}^{N} \underbrace{\langle \Psi(x^{(i)}), \Psi(x^{(j)}) \rangle}_{K_{i,j}} \langle \Psi(x^{(j)}), \sum_{k=1}^{N} \alpha_k \Psi(x^{(k)}) \rangle \\
&= \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \underbrace{\langle \Psi(x^{(i)}), \Psi(x^{(j)}) \rangle}_{K_{i,j}} \underbrace{\langle \Psi(x^{(j)}), \Psi(x^{(k)}) \rangle}_{=K_{j,k}} \alpha_k = \frac{1}{N} (K^2 \vec{\alpha})_i.
\end{aligned}
$$

Hence, we get that $\Psi^T \mathbf{C} \mathbf{v} = \frac{1}{N} K^2 \vec{\alpha}$. So in total, if $\mathbf{v}$ is an eigenvector of $\mathbf{C}$ for an eigenvalue $\lambda \neq 0$ we can apply the transform Ψ^T to both sides of the eigenvalue equation and get that $\vec{\alpha}$ solves the equation:

$$
N \lambda K \vec{\alpha} = K^2 \vec{\alpha}.
$$

To show that this vector $\vec{\alpha}$ solves the eigenvalue problem (3) in $\mathbb{R}^N$ we analyse two cases.
First, we assume that $\vec{\alpha} \perp \mathrm{kern}(K)$: For this case we have already proven the statement in the first part of this exercise.
Now let us assume that $\vec{\alpha} \not\perp \mathrm{kern}(K)$. This means that we can write $\vec{\alpha}$ as:

$$
\vec{\alpha} = \vec{\alpha}_0 + \vec{\alpha}_\perp,
$$

with $\vec{0} \neq \vec{\alpha}_0 \in \mathrm{kern}(K)$ and $\vec{\alpha}_\perp \perp \mathrm{kern}(K)$. If we plug this $\vec{\alpha}$ into equation (2), we get that

$$
N\lambda K \vec{\alpha}_\perp = N\lambda (K\vec{\alpha}_\perp + \underbrace{K\vec{\alpha}_0}_{=\vec{0}}) = N\lambda K \vec{\alpha} = K^2 \vec{\alpha} = K^2 \vec{\alpha}_\perp + \underbrace{K^2 \vec{\alpha}_0}_{=\vec{0}} = K^2 \vec{\alpha}_\perp.
$$

For the equation $N\lambda K \vec{\alpha}_\perp = K^2 \vec{\alpha}_\perp$ we have already shown that the eigenvalue equation $N\lambda \vec{\alpha}_\perp = K \vec{\alpha}_\perp$ holds in the first part of this exercise.

The only thing left to show is that the vector $\vec{\alpha}_0$ has no influence on the solution $\mathbf{v} \in \mathcal{H}$ of the eigenvalue problem in the Hilbert space $\mathcal{H}$, i.e., we have to show that

$$\mathbf{v} = \sum_{i=1}^{N} \alpha_i \Psi(x^{(i)}) = \sum_{i=1}^{N} (\alpha_\perp)_i \Psi(x^{(i)}) + \sum_{i=1}^{N} (\alpha_0)_i \Psi(x^{(i)}) = \sum_{i=1}^{N} (\alpha_\perp)_i \Psi(x^{(i)}).$$

One sufficient condition for that is obviously to show that

$$\sum_{i=1}^{N} (\alpha_0)_i \Psi(x^{(i)}) = \vec{0}.$$

For this we regard the squared norm of this vector in $\mathcal{H}$:

$$\begin{aligned} || \sum_{i=1}^{N} (\alpha_0)_i \Psi(x^{(i)}) ||^2 &= \langle \sum_{i=1}^{N} (\alpha_0)_i \Psi(x^{(i)}), \sum_{j=1}^{N} (\alpha_0)_j \Psi(x^{(j)}) \rangle \\ &= \sum_{i=1}^{N} (\alpha_0)_i \sum_{j=1}^{N} \underbrace{\langle \Psi(x^{(i)}), \Psi(x^{(j)}) \rangle}_{=K_{i,j}} (\alpha_0)_j \\ &= \sum_{i=1}^{N} (\alpha_0)_i (K\vec{\alpha}_0)_i = \langle \vec{\alpha}_0, \underbrace{K\vec{\alpha}_0}_{=\vec{0}} \rangle = 0. \end{aligned}$$

This concludes the first direction of the equivalence.

(b) To show the other direction of the equivalence, we assume that $\vec{\alpha} \in \mathbb{R}^N$ solves the eigenvector equation $N\lambda\vec{\alpha} = K\vec{\alpha}$. Furthermore, we define a vector $\mathbf{v} = \sum_{i=1}^{N} \alpha_i \Psi(x^{(i)})$. We need to show that $\mathbf{v} \in \mathcal{H}$ solves the eigenvalue equation in the Hilbert space $\mathcal{H}$ for the eigenvalue $\lambda \neq 0$. For this we deduce:

$$\begin{aligned} \lambda \mathbf{v} &= \sum_{i=1}^{N} \lambda \alpha_i \Psi(x^{(i)}) = \sum_{i=1}^{N} \frac{1}{N} (K\vec{\alpha})_i \Psi(x^{(i)}) = \frac{1}{N} \sum_{i=1}^{N} \Psi(x^{(i)}) \sum_{j=1}^{N} K_{i,j} \alpha_j \\ &= \frac{1}{N} \sum_{i=1}^{N} \Psi(x^{(i)}) \sum_{j=1}^{N} \langle \Psi(x^{(i)}), \Psi(x^{(j)}) \rangle \alpha_j = \frac{1}{N} \sum_{i=1}^{N} \Psi(x^{(i)}) \langle \Psi(x^{(i)}), \underbrace{\sum_{j=1}^{N} \alpha_j \Psi(x^{(j)})}_{=\mathbf{v}} \rangle \\ &= \frac{1}{N} \sum_{i=1}^{N} \Psi(x^{(i)}) \langle \Psi(x^{(i)}), \mathbf{v} \rangle = \mathbf{C}\mathbf{v}. \end{aligned}$$

Thus, we have shown that any eigenvector $\vec{\alpha} \in \mathbb{R}^N$ of an eigenvalue $N\lambda \neq 0$ of K induces an eigenvector $\mathbf{v} \in \mathcal{H}$ of the eigenvalue $\lambda \neq 0$ of the covariance matrix $\mathbf{C}$. This concludes the proof.

Exercise 4 [Implementing Kernel PCA for data reduction].
Implement the Kernel principal component analysis algorithm as described on the

slides. For the numerical approximation of the eigenvalues and respective eigenvectors of the Kernel matrix K you can use the Python function `scipy.linalg.eig`.
Test your algorithm on the "Circle" data set. Each line of the data file has to be interpreted as a single data point with `[x, y, label]`. Compare the effect of using an inhomogeneous polynomial kernel of degree 2 and a Gaussian kernel by plotting the respective first two principal components. Choose a good value for $\sigma^2 > 0$ and $a \in \mathbb{R}$ in case of the Gaussian kernel and the polynomial kernel, respectively.

Solution. See python code.