Florian Rösel, Prof. Dr. Frauke Liers Winter semester 2021/2022

Mathematics of Learning – Worksheet 4

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

Solution.

$$\lambda_1\langle v_1, v_2\rangle = \langle \lambda_1 v_1, v_2\rangle = \langle Av_1, v_2\rangle \overset{\text{Symmetry}}{=} \langle v_1, Av_2\rangle = \langle v_1, \lambda_2 v_2\rangle = \lambda_2\langle v_1, v_2\rangle$$

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

Exercise 1 [Reading assignment: Spectral Clustering].
Read chapter 14.5.3 regarding Spectral Clustering in the Hastie book. Spectral Clustering is a method which can be applied to data with some radial structure, for example. At some point in the chapter, the Laplacian of graphs will be of importance. If you do not know about graph Laplacians, inform yourself about it (it is going to be important later in the lecture). Peculiarly ambitious students can implement their version of Spectral Clustering and apply it on various data sets (extract some data from the internet or use the data sets already uploaded or which you generated on your own - be creative). Discuss the contents of the chapter with a fellow student for at least half an hour.

Exercise 2 [Equivalence of eigenvalue problems].
Let $x^{(1)}, \dots, x^{(N)}$ 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:

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

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

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

   is also a solution of the eigenvalue equation:

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

   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 (**??**) but not (**??**). Then we now that:

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

We can rewrite (**??**) 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 $\lambda > 0$:

$$\begin{gathered} \mathbf{v} \in \mathcal{H} \text{ is eigenvector of } \mathbf{C} \text{ with respect to eigenvalue } \lambda \\ \Leftrightarrow \\ \vec{\alpha} \in \mathbb{R}^m \text{ is eigenvector of } K \text{ with respect to eigenvalue } N\lambda \end{gathered}$$

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 $i = 1, \ldots, N$. Using this notation we apply this operator to the let side of the eigenvalue equation and can deduce for all $i = 1, \ldots, N$:

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

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 (**??**) 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 (**??**), 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 $\vec{v} \in \mathcal{H}$ 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 3 [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.

Acknowledgement, for generating a previous version of this sheet, to Leon Bungert and Daniel Tenbrinck.