Prof. Dr. Jan Rolfes, Ehsan Waiezi, Lars Weidne Winter term 23/24

*Mathematics of Learning* – Worksheet 3 - – Discussion on Novervmber 2/3th, 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 [Analytically calculating eigenvalues and eigenvectors].**[1]

Let $A \in \mathbb{R}^{n \times n}$ be a quadratic matrix. Whenever for a vector $v \in \mathbb{R}^n$ and for a $\lambda \in \mathbb{R}$ the equation

$$Av = \lambda v$$

holds, we call $\lambda$ and eigenvalue of $A$ and $v$ the corresponding eigenvector. Find out how to calculate eigenvalues and eigenvectors analytically and calculate the eigenvalues and eigenvectors of the matrix

$$A := \frac{1}{3} \cdot \begin{pmatrix} 5 & -2 & -1 \\ -2 & 5 & 1 \\ -1 & 1 & 8 \end{pmatrix}.$$

Hint: The eigenvalues of this matrix are integers; if you get some fractional values, you made some mistake.

**Solution.**

To calculate the eigenvalues analytically, it is (at least for $3 \times 3$ matrices) feasible to calculate the characteristic polynomial, i.e., $\chi_A(\lambda) := \det(A - \lambda I_{3\times 3})$. The values for which the polynomial evaluates to zero are the eigenvalues of the matrix. For our matrix, the characteristic polynomial is

$$\det(\frac{1}{3} \cdot \begin{pmatrix} 5-3\lambda & -2 & -1 \\ -2 & 5-3\lambda & 1 \\ -1 & 1 & 8-3\lambda \end{pmatrix}) = \frac{1}{27} \cdot \det(\begin{pmatrix} 5-3\lambda & -2 & -1 \\ -2 & 5-3\lambda & 1 \\ -1 & 1 & 8-3\lambda \end{pmatrix}) =$$

$$\frac{1}{27}((5-3\lambda)(5-3\lambda)(8-3\lambda) + 2 + 2 - (5-3\lambda) - 4(8-3\lambda) - (5-3\lambda)) =$$

$$\frac{1}{27} \cdot (-27\lambda^3 + (45+45+72)\lambda^2 + (-75-120-120+3+12+3)\lambda + 200+2+2-5-32-5) =$$

$$\frac{1}{27} \cdot (-27\lambda^3 + 162\lambda^2 - 297\lambda + 162) = -\lambda^3 + 6\lambda^2 - 11\lambda + 6 = (1-\lambda)(2-\lambda)(3-\lambda).$$

[1] There are lots of nice tutorial books for linear algebra and analysis available in our library, one of them I put in the forum. For a less formal introduction, you can also consult wikipedia (with caution).

This evaluates to zero for $\lambda_1 = 1, \lambda_2 = 2$ and $\lambda_3 = 3$. The next step is to solve linear equation systems $A - \lambda_i I_{3\times 3} = 0$ to obtain the eigenvectors. For example for $\lambda_1$ (we compute $3A - 3\lambda_1 I_{3\times 3}$ for easier reading):

$$\begin{pmatrix} 2 & -2 & -1 \\ -2 & 2 & 1 \\ -1 & 1 & 5 \end{pmatrix} \overset{\text{add I to II and III, multiply and swap}}{\rightarrow} \begin{pmatrix} 2 & -2 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

This is (non-trivially) solvable by the vector $(1,1,0)^T$ (or any multiple of this vector). The solutions for $\lambda_2$ and $\lambda_3$ are $(-1,1,-1)^T$ and $(-1,1,2)^T$.

**Exercise 1 [Definiteness of the covariance matrix].**
Let $y^{(1)}, \ldots, y^{(N)} \in \mathbb{R}^M$ be centered input data and let $C$ be the respective covariance matrix.

1. Show that $C$ is always positive semi-definite.

   **Solution.** We have to show that

   $$x^T C x \;\geq\; 0 \quad \forall x \in \mathbb{R}^M.$$

   Hence, let $x \in \mathbb{R}^M$ be an arbitrary vector. Then we have:

   $$\begin{aligned} x^T C x &= \langle x, Cx \rangle \;=\; \langle x, \frac{1}{N} \sum_{i=1}^N y^{(i)} y^{(i)T} x \rangle \\ &= \frac{1}{N} \sum_{i=1}^N \langle x, y^{(i)} y^{(i)T} x \rangle \;=\; \frac{1}{N} \sum_{i=1}^N \langle x, y^{(i)} \langle y^{(i)}, x \rangle \rangle \\ &= \frac{1}{N} \sum_{i=1}^N \langle y^{(i)}, x \rangle \langle x, y^{(i)} \rangle \;=\; \frac{1}{N} \sum_{i=1}^N \langle x, y^{(i)} \rangle^2 \;\geq\; 0. \end{aligned}$$

2. In which cases is $\langle x, Cx \rangle = 0$ for an $x \in \mathbb{R}^M \setminus \{\vec{0}\}$. What does that mean for the given data?

   **Solution.** Let $x \in \mathbb{R}^M \setminus \{\vec{0}\}$ be an arbitrary vector. We know that $\langle x, Cx \rangle = 0$ only if the covariance matrix $C$ does not have full rank, i.e.,

   $$\langle x, Cx \rangle = 0 \;\Rightarrow\; \text{rank}(C) < M.$$

   We distinguish two cases:

   (a) We assume that $N < M$, i.e., we have less data points $y^{(1)}, \ldots, y^{(N)}$ than the dimension $M$ of the input data space $\mathbb{R}^M$. Due to the subadditivity property of the rank operator we know for two matrices $A, B \in \mathbb{R}^{M\times M}$ that:

   $$\text{rank}(A + B) \;\leq\; \text{rank}(A) + \text{rank}(B).$$

   In case of the covariance matrix $C$ we can then deduce:

   $$\text{rank}(C) \;=\; \text{rank}(\frac{1}{N} \sum_{i=1}^N y^{(i)} y^{(i)T}) \;\leq\; \sum_{i=1}^N \text{rank}(y^{(i)} y^{(i)T}) \;=\; \sum_{i=1}^N 1 \;=\; N \;<\; M.$$

(b) Now we assume that $N \geq M$, $y^{(i)} \neq \vec{0}, 1 \leq i \leq N$, and there is no pair of parallel vectors, i.e.,

$$y^{(i)} \neq \alpha y^{(j)}, \quad 1 \leq i, j < N, i \neq j, \forall \alpha \in \mathbb{R}/\{0\}.$$

Then, the only possible way that $\operatorname{rank}(C) < M$ is that at least one of the eigenvalues is 0. But that means that there exists a direction $v \in \mathbb{R}^M$ in which there is no variance. Hence, the centered input data $y^{(1)}, \ldots, y^{(N)}$ lies in a hyperplane of $\mathbb{R}^M$ (and can be reduced losslessly via PCA).

**Exercise 2 [Prerequisites for PCA].**

Given a set of data vectors $x_1, ..., x_N \in \mathbb{R}^p$ and a matrix $V \in \mathbb{R}^{p \times q}$, $q < p$, with $q$ orthogonal unit vectors as columns. Prove that

$$\tilde{\mu} = \bar{x}, \quad \tilde{\lambda}_i = V^T(x_i - \bar{x})$$

is a minimizer (over $\mu$ and $\lambda_i$) of

$$f(\mu, \lambda_1, ..., \lambda_N) = \sum_{i=1}^{N} ||x_i - \mu - V\lambda_i||^2,$$

where $|| \cdot ||$ denotes the euclidean norm. Furthermore, show that the minimizer $\bar{x}$ for $\mu$ is not unique and find the set of minimizers for $\mu$.

**Solution.** We can interpret the setting as follows: Given data points in $\mathbb{R}^p$, $V^T$ defines a transformation to $\mathbb{R}^q$ with $q < p$ by simply deleting entries of the data points. The function $f$ now determines loss if we are given data points $x^{(i)} \in \mathbb{R}^p$ and are allowed to choose a translation (by $\mu$) and $N$ points in $\mathbb{R}^q$ which are then transformed back to $\mathbb{R}^p$. We want to show that indeed the center $\bar{x}$ and the transformed centered data points are minimizing this loss.

For given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, it is relatively easy to show that the function $g(x, y) = ||x + Ay + b||^2$ is convex. Thus, $f$ is the sum of convex functions and therefore convex. Hence, to find minima it suffices to find critical points, i.e., zeros of the gradient. We observe that $f$ can be rewritten as

$$f(\mu, \lambda_1, ..., \lambda_N) = \sum_{i=1}^{N} \sum_{j=1}^{p} (x_{i,j} - \mu_j - (V\lambda_i)_j)^2.$$

We calculate the gradient (and therefore the first partial derivatives) of $f$ first with respect to $\mu$, i.e.,

$$\frac{\partial}{\partial \mu_k} \sum_{i=1}^{N} \sum_{j=1}^{p} (x_{i,j} - \mu_j - (V\lambda_i)_j)^2 = -\sum_{i=1}^{N} 2(x_{i,k} - \mu_k - (V\lambda_i)_k)$$

for $k = 1, ..., p$. Using $\tilde{\mu}$ and $\tilde{\lambda}_i$, we get

$$-\sum_{i=1}^{N} 2(x_i - \bar{x} - VV^T(x_i - \bar{x})) = -2(\sum_{i=1}^{N} x_i - N\bar{x} - VV^T(\sum_i x_i - N\bar{x})) = 0$$

since $\sum_{i=1}^{N} x_i - N\bar{x}$ is the zero vector.

On the other hand, if we calculate the partial derivative for $\lambda_{\ell,k}$ for $\ell = 1, ..., N$ and $k = 1, ..., q$, we get

$$\begin{aligned} \frac{\partial}{\partial \lambda_{\ell,k}} \sum_{i=1}^{N} \sum_{j=1}^{p} (x_{i,j} - \mu_j - (V\lambda_i)_j)^2 &= -\sum_{j=1}^{p} 2(x_{\ell,j} - \mu_j - (V\lambda_\ell)_j) \cdot V_{j,k} \\ &= -2\langle x_\ell - \mu - V\lambda_\ell, V_{.,k}\rangle \\ &= -2\langle x_\ell - \mu - V\lambda_\ell, Ve_k\rangle \\ &= -2\langle V^T(x_\ell - \bar{x}) - V^T V V^T (x_\ell - \bar{x}), e_k\rangle = 0, \end{aligned}$$

where we used the values for $\tilde{\mu}$ and $\tilde{\lambda}_i$ in the last line.

The minimizer is not unique, since we can choose any vector for $\mu$ such that $\sum_{i=1}^{N} x_i - N\mu$ is an eigenvector corresponding to eigenvalue 1 of the matrix $VV^T$, and accordingly $\lambda_i = V^T(x_i - \mu)$. (Consider the example $x^1 = (1\ \ 1)^T, x^2 = (2\ \ 1)^T$ and $V = (1\ \ 0)^T$. Then $\bar{x} = (1.5\ \ 1)^T$, but also any vector $\mu = (a\ \ 1)^T$ with $a \in \mathbb{R}$ can be used.)

**Exercise 3 [Implementing PCA for data reduction].**
Implement the (linear) principal component analysis algorithm as described on the slides. For the numerical approximation of the eigenvalues and respective eigenvectors of the covariance matrix $C$ you can use the Python function `scipy.linalg.eig`.
Test your algorithm on the Iris data set[2]. This is perhaps the most popular data set to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The data has 5 columns representing the following attributes:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

- Iris Setosa
- Iris Versicolor
- Iris Virginica

You can ignore attribute 5 in the above list for the PCA, but use it for visualization of the different classes. Plot the features after applying the PCA algorithm for $k = 3$ and $k = 2$. What can you observe?

**Solution.** See python code.

**Exercise 4 [Apply clustering algorithms].**
Apply the clustering algorithms ($k$-means and EM) to the Iris data set (before and after PCA). Describe, interpret and visualize your results.

[2] Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936);

**Bonus [Apply clustering algorithms].**
Apply the clustering algorithms (*k*-means and EM) to the data set you created on your own for the last exercise sheet's bonus exercise (before and after PCA). Describe, interpret and visualize your results; concentrate particularly on some differences between your data set and the Iris data set, if there are some, and try to explain the reasons why they occur.