Florian Rösel, Frauke Liers Winter semester 2021/2022

Mathematics of Learning – Worksheet 2

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 λ 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.

Exercise 1 [Get some literature sources].
Get the two books recommended for reading in this course (see module manual) -

- Goodfellow et al., Deep learning. e.g. `https://www.deeplearningbook.org/`, maybe you find a better source; if so, let me know.
- Hastie et al., The Elements of Statistical Learining (available as full text pdf in our library)

Read one (you choose which) subsection dealing with "unsupervised learning" (in the Hastie book). Explain it to a fellow student.

Exercise 2 [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.
2. In which cases is $\langle y, Cy \rangle = 0$ for $y \in \mathbb{R}^M / \{\vec{0}\}$. What does that mean for the given data?

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`.

[1] There are lots of nice tutorial books for linear algebra and analysis available in our library. For a less formal introduction, you can, e.g., also consult wikipedia ;)

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?

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.

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 peculiarly 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.

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

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