Exercise Sheet 1

Exercise 0

Set-up a python environment that supports the following libraries:

- numpy
- scipy
- matplotlib
- scikit-learn

You can use different libraries in your solutions, but support will only be provided for the libraries mentioned above.

Exercise 1

Show that the Mahalanobis distance fulfils the requirements of a norm:

Mahalanobis distance:

$$d(x, y)_m = ||x - y||_m = \sqrt{(x - y)^T \cdot \Sigma^{-1} \cdot (x - y)}$$

(a) $||\lambda \cdot x||_m = |\lambda| \cdot ||x||_m$

(b) $||x||_m = 0 \leftrightarrow x = 0$

(c) $||x||_m \geq 0$

(d) $||x + y||_m \leq ||x||_m + ||y||_m$

Hint 1: A norm behaves like a distance from the origin. Therefore, if you want to get a norm from a distance, you have to set y to the zero vector:

$$||x||_m = d\left(x, \vec{0}\right)$$

As a result of this, the Mahalanobis distance as a norm can be written as follows:

$$||x||_m = \sqrt{x^T \cdot \Sigma^{-1} \cdot x}$$

Hint 2: To proof (d) (triangle inequality), you can do the following steps:

1. For a norm, Σ^{-1} needs to be a positive-definite matrix → according to the spectral theorem, Σ^{-1} can be decomposed to $\Sigma^{-1} = Q^T \cdot \Lambda \cdot Q$ (SVD, see last semesters lecture *Mathematics & Modeling*).
2. Let $U = sqrt(\Lambda) \cdot Q$. Argue why $\Sigma^{-1} = U^T \cdot U$.
3. Use the fact that the euclidean distance is a norm to show that (d) is also valid for the Mahalanobis distance by setting $\bar{x} = U \cdot x$ and $\bar{y} = U \cdot y$.

Exercise 2

Explore the wine data set contained in the `sklearn` library (https://scikit-learn.org/stable/) of python.

(a) Implement a PCA on your own to extract the first two main components of the data set.

(b) Visualize your results.

Hint 1: For (a) you need to remember what you have learned in last semesters course *Mathematics & Modelling* (*eigenvalues* and *eigenvectors*). See also slide # 25 in the lecture notes.

Exercise 3

Use the data provided in `wine-data_reduced.csv` and:

(a) Estimate the amount of data points contained in the $1-\sigma$, $2-\sigma$ and $3-\sigma$ ellipsoid area. Data points contained in these areas exhibit a Mahalanobis distance ≤ 1, ≤ 2 and ≤ 3.

(b) Compare your results with those of the traditional $3-\sigma$ rule used for the normal distribution.

(c) Visualize your results.

(d) What happens if you normalize the axis by the corresponding eigenvalues?