Linnaeus University / 4DV510

# Project Requirements 2:

Dimensionality Reduction

## Introduction

Dimensionality Reduction (DR) techniques reduce the raw number of features of the data into a new, smaller set, which hopefully still retains the main characteristics of the full data set but is easier to handle, has less noise, and highlights the latent factors from the high-dimensional space. We can identify some common properties in most of them: (a) they take structured data as input; (b) they can usually be customized with different hyper-parameters; and (c) the output is a new dataset with the same number of points, but fewer dimensions. Most (but not all) will use distance measures to compare pairs of points, and we can also find many "validation" (or "quality") measures that are independent and external.

## Requirements

You should proceed with the implementation of your data mining framework by focusing on dimensionality reduction techniques, and quality measures for DR.

1. **Create a component for a *Dimensionality Reduction* (or *Projection*) technique.** It should accept as input a *Dataset* and optional hyper-parameters. The hyper-parameters must not be hard-coded; every different technique will have different ones. One parameter should be a *Distance Measure* (even though not all techniques use it, most will). The output is a 2D array of points (rows) and features (columns).
2. **Create a component for a *Quality Measure* for DR.** It should take as input both the results of a DR technique and a Dataset, and output a single value.

## Notes:

Remember to write also at least three different implementations for each component. In order to test your new requirements, make sure you can run a pipeline with them where you programmatically test different combinations of each component. Remember that you will reuse components from the previous week, like the Dataset and the Distance Measure. Finally, run a pipeline where you first reduce the dimensions of the data and then cluster it. Can you get better/faster results than running the Clustering directly in the high-dimensional data?