UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning
Work 2. PCA and SOM Exercise

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona

Contents

1. Introduction
2. Principal Components Analysis
3. Self-Organizing Maps

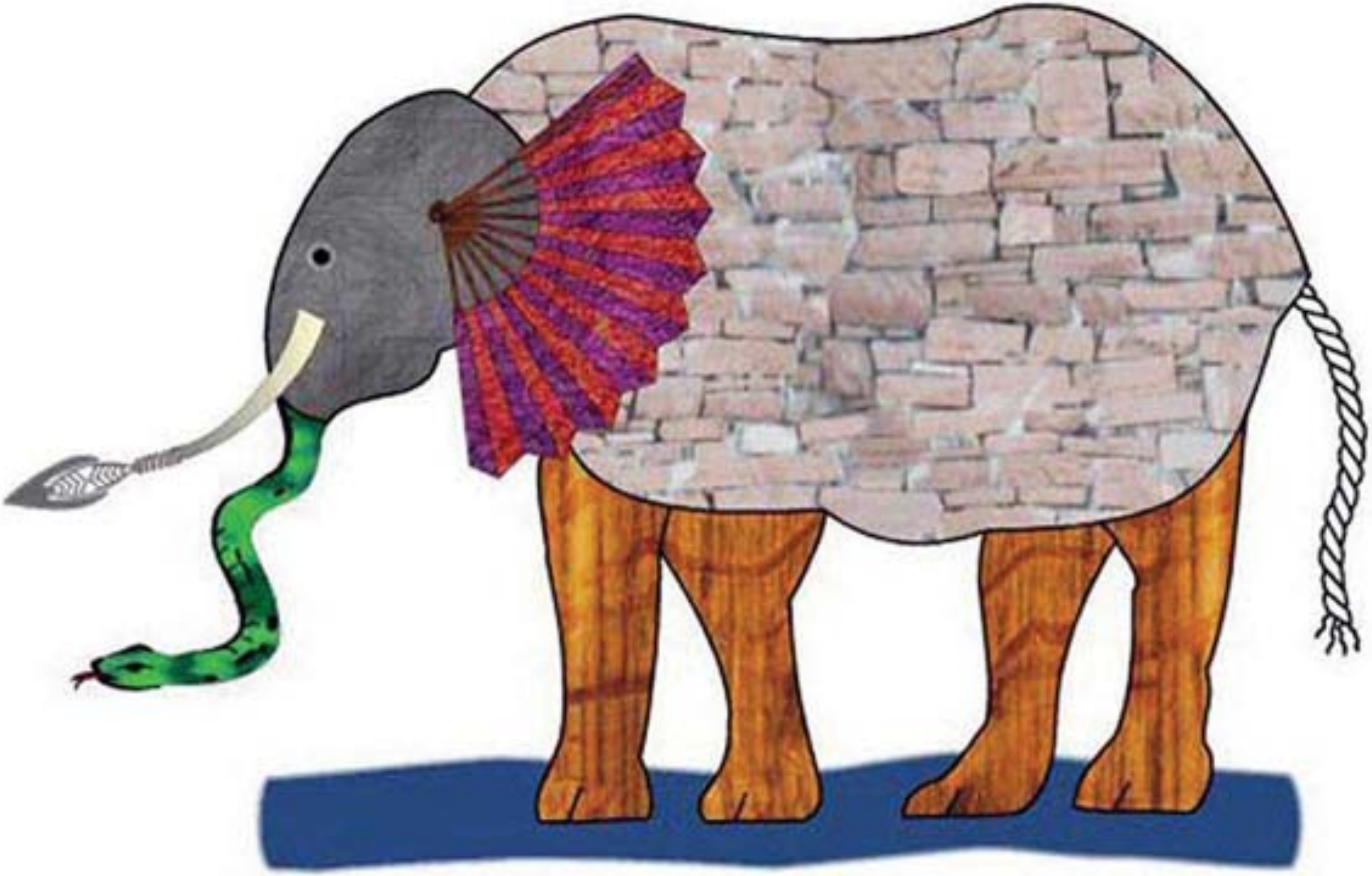
Introduction

Introduction

Unsupervised learning is a class of machine learning algorithms which involves modelling the underlying structure or distribution of the ***"unlabeled"*** data.

Unlabeled data means the classification or categorization is not available in the observations.

Introduction



Six blind brothers and an Elephant

Introduction

The **goal** of Work 2 is to…

1. Reduce dimensionality with PCA
2. Analyse K-Prototype with and without dimensionality reduction
3. Analyse SOM as a clustering algorithm
4. Compare K-Prototype and SOM

Principal Components Analysis

Implement your own code

PCA

- **Principal Component Analysis (PCA)** is a **dimension-reduction** tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

- PCA is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*.

- The first principal component accounts for as much as of the variability in the data as posible, and each succeding component accounts for as much of the remaining variability as possible.

PCA

- Traditionally, PCA is performed on a **square symmetric** matrix. It can be:
  - A **SSCP** matrix (pure sums of squares and cross products),
  - A **Covariance** matrix (scaled sums of squares and cross products), or,
  - A **Correlation** matrix (sums of squares and cross products from standardized data).
- The analysis results for objects of types SSCP and Covariance do not differ, since these objects only differ in a global scaling factor.
- A correlation matrix is used if the variances of individual variates differ much, or if the units of measurement of the individual variates differ.

PCA

- In this work, you have to:
  - Implement your own code of PCA, using a covariance matrix
  - Compare and analyze your results to the ones obtained using:
    - `sklearn.decomposition.PCA` (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html) and,
    - `sklearn.decomposition.IncrementalPCA` (https://scikit-learn.org/stable/auto_examples/decomposition/plot_incremental_pca.html)

Self-Organizing Maps

Use neupy library

SOM

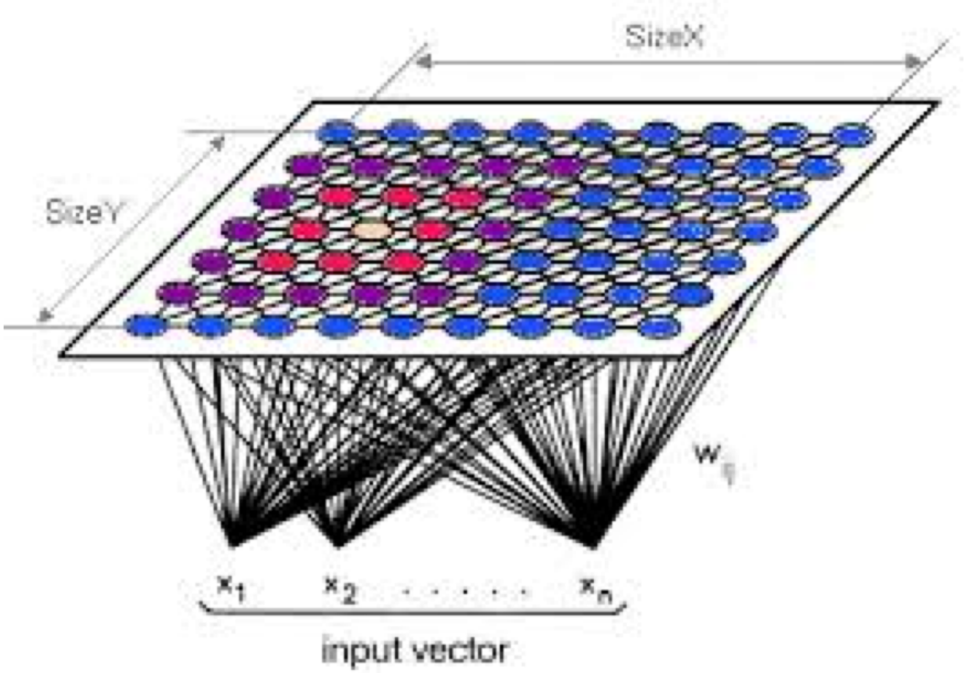
- Self Organizing Maps (SOM) are also known as Kohonen's Map as they were first introduced by the Finnish professor Teuvo Kohonen in the 1980s.

- *The most common definition, "A* ***self-organizing map*** *(****SOM****) or* ***self-organizing feature map*** *(****SOFM****) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a* ***map****, and is therefore a method to do dimensionality reduction."*

- *In fact, SOM can be applied for* ***clustering or dimensionality reduction*** *for various domains.*

SOM

- It is a type of artificial neural network (ANN), a system of computing inspired by biological neural networks of animal brains.
  - So, there are neurons and connections between them



SOM

- Use **NeuPy** library to clusterize the data
- Information of the library is available at http://neupy.com/pages/home.html

- More information at:
  - http://neupy.com/2017/12/09/sofm_applications.html#self-organizing-maps-and-applications