UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning Work 2. Dimensionality Reduction and Visualization using PCA and t-SNE

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona

Contents

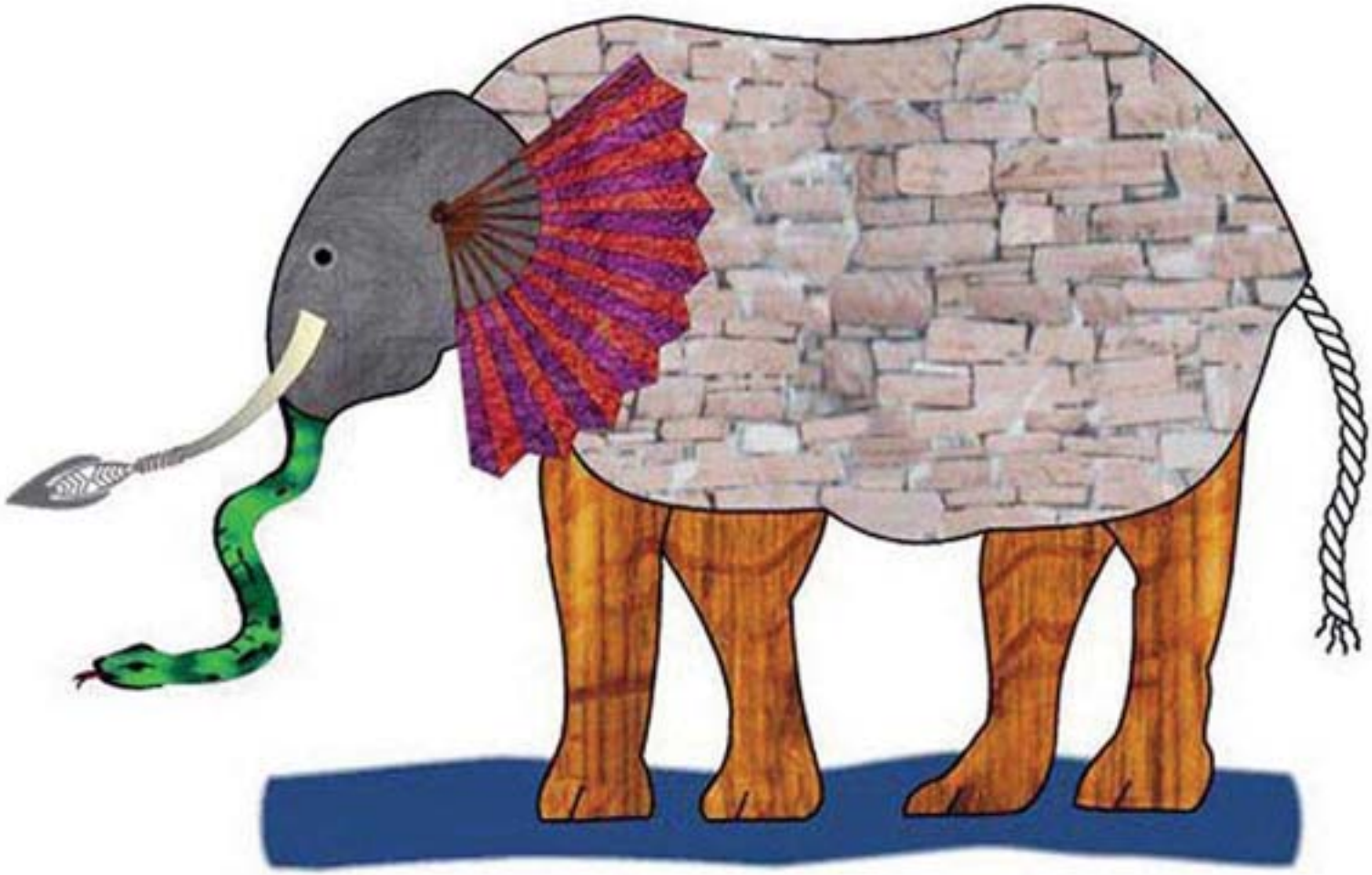
Introduction

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.

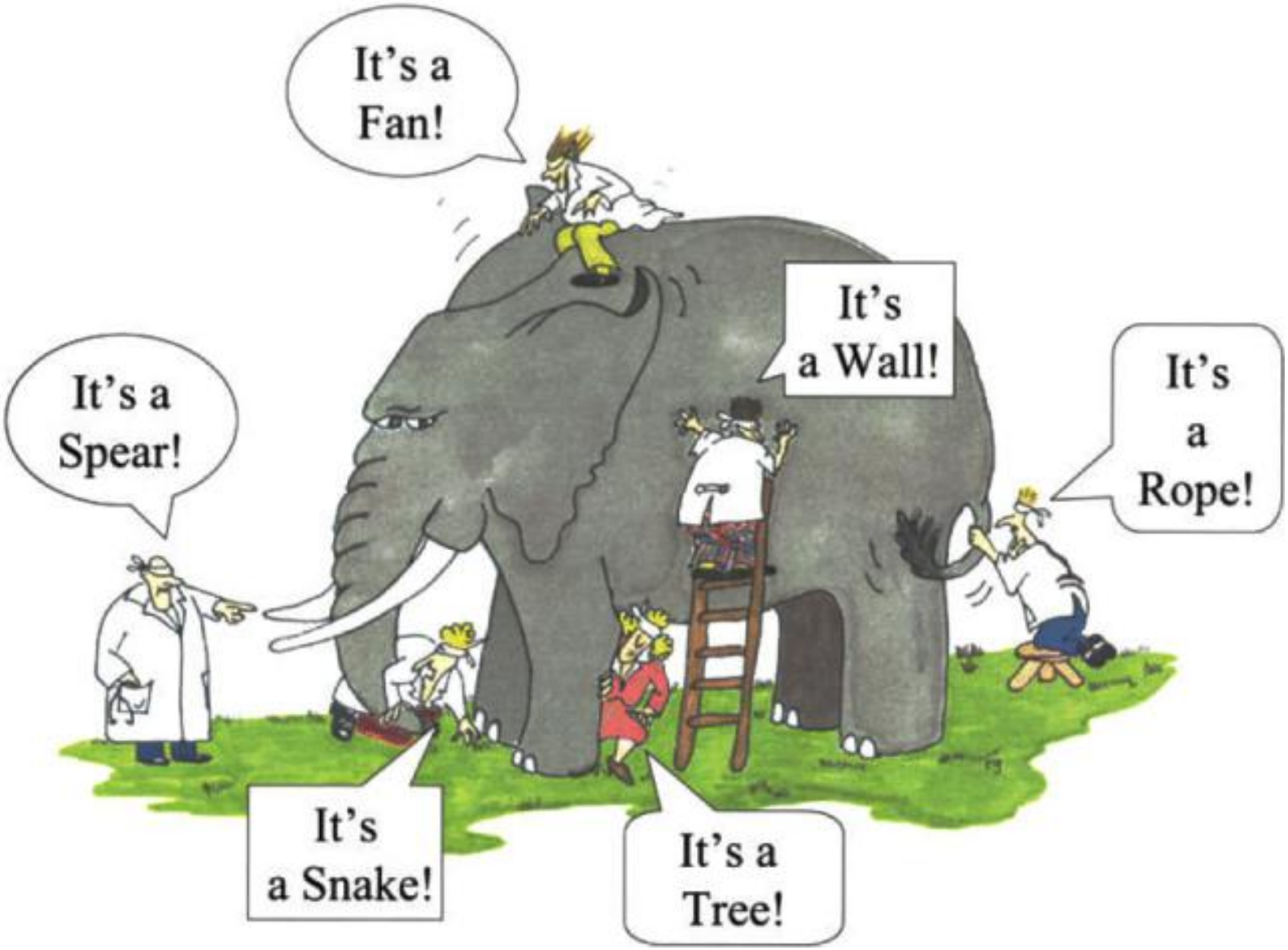
Introduction



Parable: *The blind men and an elephant*

UNIVERSITAT DE BARCELONA

Introduction



Introduction

The **goal** of Work 2 is to…

1. Reduce dimensionality with PCA
2. Analyse K-Means with and without dimensionality reduction
3. Analyse different low-dimensional visualization algorithms
4. Compare PCA and t-SNE

Principal Component Analysis

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 works by identifying the hyperplane which lies closest to the data and then projects the data on that hyperplane while retaining most of the variation in the data 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 possible, and each succeeding 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)

t-SNE

Main approaches for Dimensionality Reduction

The two main approaches:

- **Projection**: This technique deals with projecting every data point which is in high dimension, onto a subspace suitable lower-dimensional space in a way which approximately preserves the distances between the points
- **Manifold Learning:** Many dimensionality reductions algorithm work by modelling the manifold on which the training instance lie; this is called *Manifold learning*.
  - It relies on the manifold hypothesis or assumption, which *holds that most real-world high-dimensional datasets lie close to a much lower-dimensional manifold,* this assumption in most of the cases is based on observation or experience rather than theory or pure logic

t-SNE

- t-distributed Stochastic Neighbor Embedding
  - It is an **unsupervised non-linear** dimensionality reduction and data visualization technique
  - Proposed Laurens van der Maaten and Geoffrey Hinton in 2008
  - The math behind it is quite complex but the idea is simple
    - It embeds the points from a higher dimension to a lower dimension trying to preserve the neighborhood of that point

t-SNE

- t-distributed Stochastic Neighbor Embedding
  - Unlike PCA it tries to preserve the local structure of data by minimizing the **Kullback-Leibler divergence (KL divergence)** between two distributions with respect to locations of the points in the map.
  - However, it ignores almost completely the global structure
  - Applications
    - Security research, music analysis, cancer research, bioinformatics, and biomedical signal processing, among others

Comparison

PCA	T-SNE
It is a **linear** Dimensionality reduction technique	It is a **non-linear** Dimensionality reduction technique
It tries to preserve the **global structure** of the data	It tries to preserve the **local structure** (cluster) of data
It does not work well as compared to t-SNE	It is one of the best dimensionality reduction technique
It does not involve Hyperparameters	It involves Hyperparameters such as *perplexity*, *learning rate* and *number of steps*
It gets highly **affected by outliers**	It can **handle outliers**
PCA is a **deterministic** algorithm	It is a **non-deterministic**
It works by **rotating the vectors for preserving variance**	It works by **minimising the distance between the point in a guassian**
We can find decide on how much variance to preserve using eigen values	We cannot preserve variance instead we can preserve distance using hyperparameters

Links

- **https://mlexplained.com/2018/09/14/paper-dissected-visualizing-data-using-t-sne-explained/**

- **https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html**

- ORIGINAL PAPER (optional) https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning Work 2. Dimensionality Reduction and Visualization using PCA and t-SNE

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona