Politecnico di Bari

Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione

Corso di Laurea Triennale in Ingegneria dei Sistemi Medicali

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE

Bioinformatics and Big Data Analytics

Dimensionality Reduction

Eng. Nicola **Altini**, *Ph.D. Student*
Eng. Giacomo Donato **Cascarano**, *Ph.D. Student*
Prof. *Eng*. Vitoantonio **Bevilacqua**, *Ph.D.*

INDUSTRIAL INFORMATICS LAB
LABORATORIO DI INFORMATICA INDUSTRIALE

Anno Accademico 2019/2020

apulian bioengineering company

The Curse of Dimensionality

- The **curse of dimensionality** refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. The expression was coined by Richard E. Bellman when considering problems in dynamic programming.
- **Dimensionality reduction** is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. Approaches can be divided into feature selection and feature extraction.

The Curse of Dimensionality

- Cursed phenomena occur in domains such as numerical analysis, sampling, combinatorics, machine learning, data mining and databases.
- The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality.
- Also, organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data, however, all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient.

Dimensionality Reduction

- Feature selection approaches try to find a subset of the input variables (also called features or attributes). The three strategies are: the filter strategy (e.g. information gain), the wrapper strategy (e.g. search guided by accuracy), and the embedded strategy (selected features add or are removed while building the model based on prediction errors).
- Feature projection (also called Feature extraction) transforms the data in the high-dimensional space to a space of fewer dimensions. The data transformation may be linear, as in principal component analysis (PCA), but many nonlinear dimensionality reduction techniques also exist. For multidimensional data, tensor representation can be used in dimensionality reduction through multilinear subspace learning.

PCA

- **Principal component analysis** (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.
- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.
- The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

PCA – MATLAB

`coeff = pca(X)`

- returns the principal component coefficients, also known as loadings, for the n-by-p data matrix `X`.
- Rows of `X` correspond to observations and columns correspond to variables. There are n observations and p variables.
- The coefficient matrix is p-by-p. Each column of `coeff` contains coefficients for one principal component, and the columns are in descending order of component variance. By default, `pca` centers the data and uses the singular value decomposition (SVD) algorithm.

PCA – MATLAB

- `[coeff,score,latent] = pca(X)` also returns the principal component scores in score and the principal component variances in latent.
- Principal component scores are the representations of `X` in the principal component space. Rows of score correspond to observations, and columns correspond to components.
- The principal component variances are the eigenvalues of the covariance matrix of `X`.

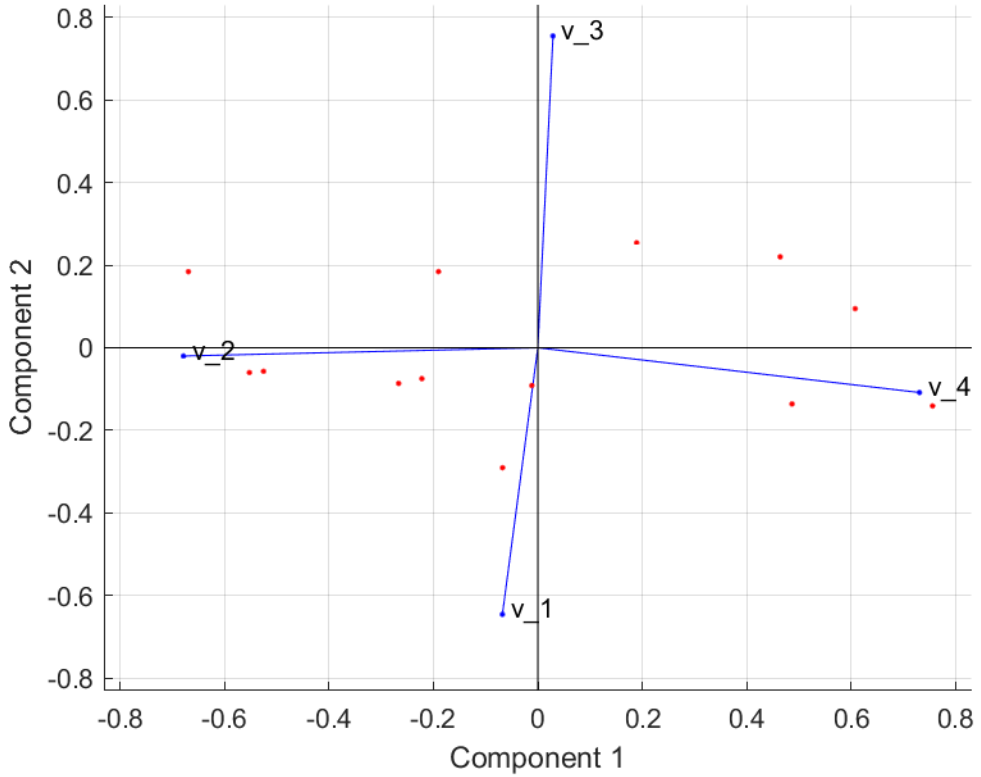
Bi-plot

- Biplots are a type of exploratory graph used in statistics, a generalization of the simple two-variable scatterplot. A biplot allows information on both samples and variables of a data matrix to be displayed graphically.
- Samples are displayed as points while variables are displayed either as vectors, linear axes or nonlinear trajectories. In the case of categorical variables, category level points may be used to represent the levels of a categorical variable.
- A generalised biplot displays information on both continuous and categorical variables.

PCA and bi-plot with MATLAB

Visualize both the orthonormal principal component coefficients for each variable and the principal component scores for each observation in a single plot.

```
[coeff,score,latent] = pca(ingredients);
biplot(coeff(:,1:2),'scores',score(:,1:2),'varlabels',{'v_1','v_2','v_3','v_4'});
```



t-SNE

- It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of 2 or 3 dimensions.
- Specifically, it models each high-dimensional object by a 2D or 3D point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

t-SNE

- The t-SNE algorithm comprises two main stages:
  1. t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked while dissimilar points have an extremely small probability of being picked.
  2. t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map.
- Hyper-parameter: **`perplexity`**
  - It is basically the effective number of neighbors for any point, and t-SNE works relatively well for any value between 5 and 50. Larger perplexities will take more global structure into account, whereas smaller perplexities will make the embeddings more locally focused.

t-SNE – MATLAB

`Y = tsne(X)`

returns a matrix of two-dimensional embeddings of the high-dimensional rows of `X`.

`X - Data points`

specified as an `n`-by-`m` matrix, where each row is one `m`-dimensional point.

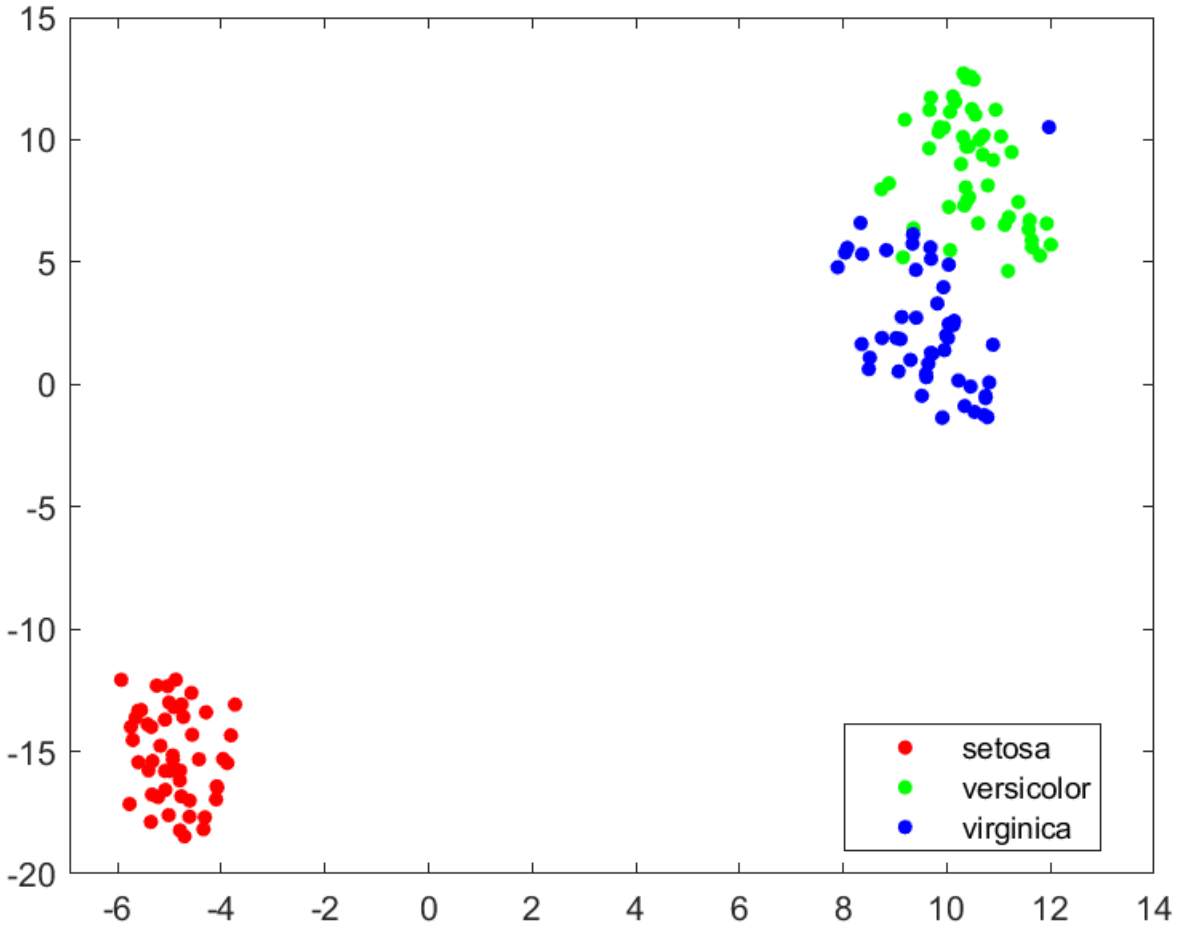
`Y – Embedded points`

returned as an `n`-by-`NumDimensions` matrix. Each row represents one embedded point.

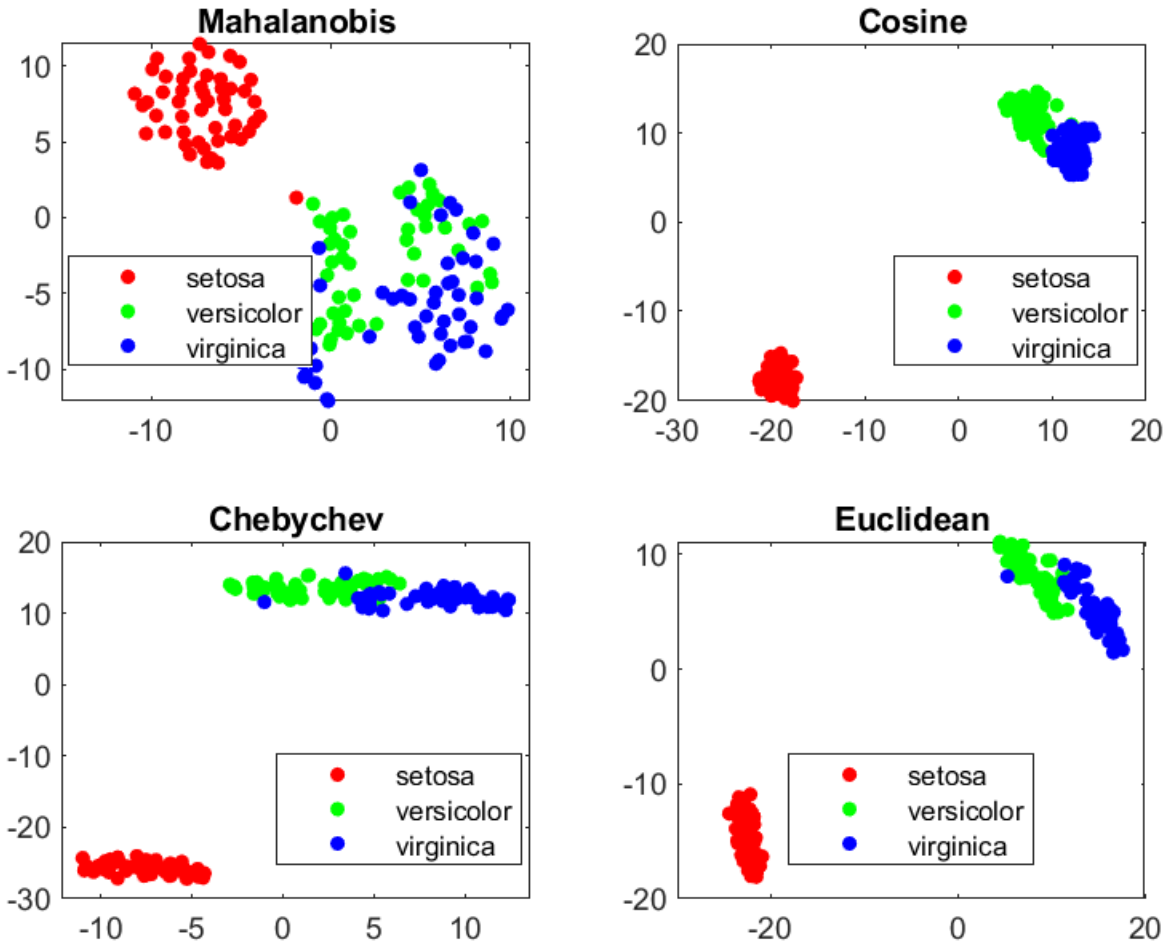
t-SNE – MATLAB

- t-SNE algorithm:
  - The `'exact'` algorithm optimizes the Kullback-Leibler divergence of distributions between the original space and the embedded space.
  - The `'barneshut'` algorithm performs an approximate optimization that is faster and uses less memory when the number of data rows is large.
- Examples of distance metrics you can use in MATLAB implementation of t-SNE are:
  - `'euclidean'` — Euclidean distance.
  - `'chebychev'` — Chebychev distance, which is the maximum coordinate difference.
  - `'mahalanobis'` — Mahalanobis distance, computed using the positive definite covariance matrix `nancov(X)`.
  - `'cosine'` — 1 minus the cosine of the included angle between observations (treated as vectors).

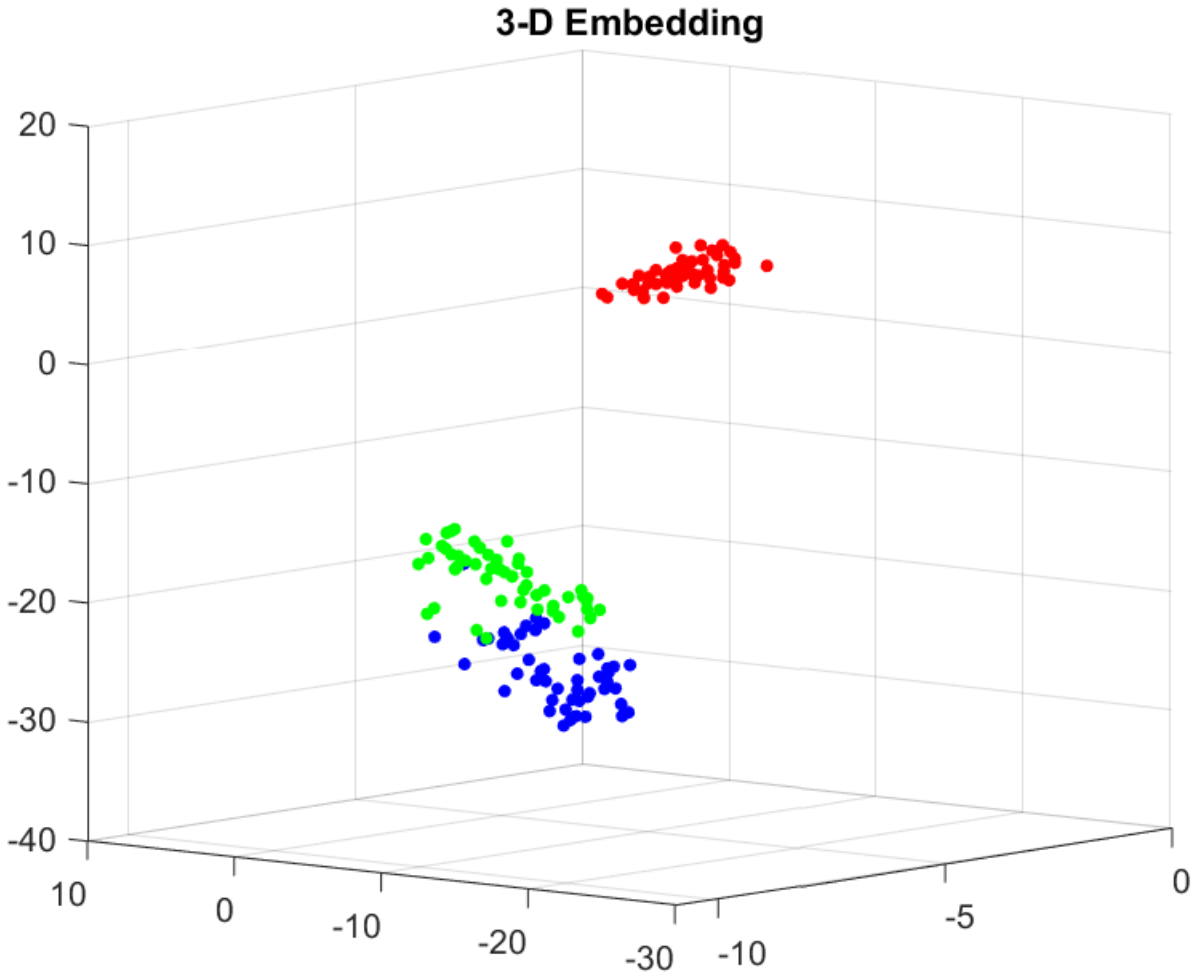
t-SNE in MATLAB



t-SNE in MATLAB



t-SNE in MATLAB



References

- MATLAB Documentation
  - PCA.
    URL: https://www.mathworks.com/help/stats/pca.html
  - t-SNE.
    URL: https://www.mathworks.com/help/stats/tsne.html
- Wikipedia.
  - Dimensionality Reduction.
    URL: https://en.wikipedia.org/wiki/Dimensionality_reduction
  - Curse of Dimensionality.
    URL: https://en.wikipedia.org/wiki/Curse_of_dimensionality