Dimensionality Reduction and high-dimensional data visualization

R. Benítez

The following exercises can be implemented using MATLAB, Python or R. Use the support examples and codes in the digital campus ATENEA. At the end of the document we provide a list of useful web references for further information on how to use each method in each platform.

Exercise 1: PCA with synthetic data

Task 1: Generate two random vectors with 10^4 observations normally distributed with different means and variances, i.e. $x_1 = N(m_1, \sigma_1)$ and $x_2 = N(m_2, \sigma_2)$.

Task 2: Generate two new variables x_3, x_4 as a linear and nonlinear combination of the previous variables (For instance $x_3 = 0.3x_1 + 1.2x_2$ and $x_4 = \sqrt{x_1} + x_2$).

Task 3: Construct a data matrix of size $m \times n$ with $m = 10^4$ observations and $n = 4$ attributes $\{x_1, x_2, x_3, x_4\}$.

Task 4: Plot a scatter correlation plot matrix of the data and visually inspect correlations between attributes. Discuss your observations with your lab mates.

Task 5: Perform a PCA decomposition of the data. Plot the resulting eigenvalues in decreasing order and select how many of them are needed in order to represent a 95% of the variance in the data.

Task 6: Project and represent the data to the reduced dimensionality PCA space.

Exercise 2: Dimensionality Reduction with PCA

Task 1: Select and download a database with *numerical* features from the UCI repository `http://archive.ics.uci.edu/ml/`.

Task 2: If the database is aimed for classification, make sure to keep only the features and remove the class labels. The resulting matrix should be of size $m \times n$ with m observations and n attributes/variables/features.

Task 3: Repeat tasks 4-6 from the previous exercise using the UCI dataset.

Exercise 3: High-dimensional data visualization with MDS

Task 1: Use Multidimensional scaling in order to visualize the data of the previous exercises in 2D.

Task 2: Select three observations in the data and check whether the euclidean distance between them is similar in both the original and the 2D MDS spaces.

Useful references

Principal Component Analysis (PCA):

- MATLAB: `https://es.mathworks.com/help/stats/pca.html`
- Python: `http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`
- R: `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/princomp.html`