Universität Stuttgart

Institut für Visualisierung und Interaktive Systeme

Assignment 6

Information Visualization & Visual Analytics (WS 2019/20)

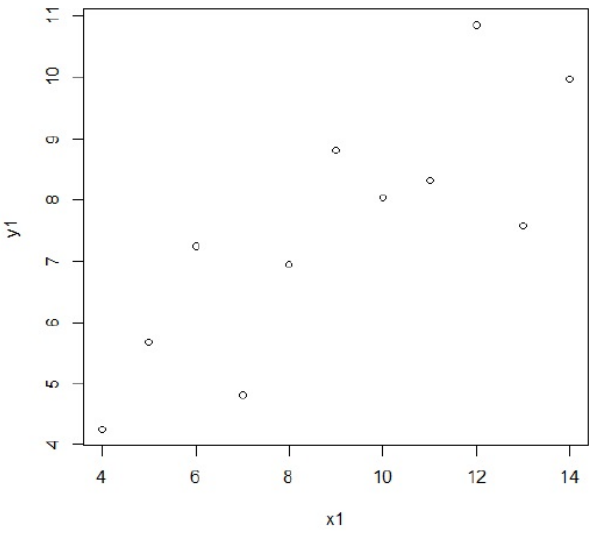
Due: Tuesday, 03.12.2019, 12:00 (theoretical) **Presentation:** Wednesday, 04.12.2019

Please solve the assignment in **groups of up to three (3) students**. Choose one student, who submits your solution in ILIAS as PDF (for theoretical submissions) or ZIP (for practical submissions Impl) as group submission (remember to include all members of your group).

Task 1 Multivariate Data (theoretical) *[Points: 12]*

In this exercise you will visualize bivariate datasets using scatterplots. Having a bivariate dataset as two column table, every data point (row) is represented as point with its corresponding coordinates (columns). The following table has four bivariate datasets with coordinates (x, y). Right next to the table, a scatterplot shows the first dataset (I).

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



(a) *(4 points)* For each of the eight variables (columns) please calculate the mean, the variance, and the correlation coefficient (Pearson). What can you say about the four datasets? Are there correlations between x and y? If so, describe them.

(b) *(4 points)* Please visualize the datasets II, III and IV in a scatterplot (hand-crafted or software of your choice). What can you say about the four datasets? Explain the connection between the visualization and the statistical results.

(c) *(4 points)* We are giving you an additional dataset, where you have $\sim$ 20000 points and $\sim$ 20 dimensions. How could you visualize this scenario in a way that you can still see the patterns? Which interactions could be used to see more details? Sketch a visualization or implement one based on the given data.

Task 2 Dimensionality Reduction (DR) *[Points: 12]*

In this task, you will perform dimensionality reduction based on *principal component analysis (PCA)* and *Isomap* or *t-SNE*.

To this end you may either use:

- **R**[1], a statistical computing software that helps you with the calculations. You can find many tutorials online for the basic use of **R** (e.g.,[2],[3]).
- **Python**[4]. You can find many helpful links in the Jupyter Notebook tutorial from a few weeks ago.

Regardless of the programming language used you may use off-the-shelf libraries that provide fast computation of PCA[5] or the chosen non-linear dimensionality reduction (such as sklearn in Python).

For the results, submit images, explanations, and the script code you wrote to achieve the results!

(a) *(2 points)* In which cases is dimensionality reduction useful?

(b) *(2 points)* Load the provided dataset and create a 3D scatterplot where each point is color coded using the color column. Here is a preview of the dataset in question:

x	y	z	color
0.346381	-0.105122	1.914927	0.091068
2.383992	0.144124	9.767025	-0.097426
2.048126	-1.538079	6.142769	-0.108600
0.140701	-0.456431	0.910520	0.090634
-0.804412	0.184326	1.454936	0.085593

Table 1: Swiss Roll Dataset

(c) *(4 points)* Applying PCA

- Apply PCA on the first three dimensions of the dataset; visualize the result (the first two principal components) and color code the resulting components by the *color* column provided.
- Report or visualize the retained variance of the first two components. How is the retained variance calculated?
- Explain in Layman terms how PCA works. Comment on how the properties retained by PCA might have influenced the results you got. Is the resulting visualization helpful? Does it retain the properties in the data?

(d) *(4 points)* **Bonus Task:** Please choose **one** further dimensionality reduction technique between **tSNE** or **Isomap**[6].

- For your chosen DR, explain in Layman terms how it works. Both these DRs are parametric methods. Explain briefly what these parameters do.
- Visualize the first two dimensions retained by your chosen DR for a range of parameters. Which parameter outputs the best result?
- Comment on your visualizations. What is the advantage of your chosen DR compared to PCA? Did your chosen DR exploit some additional properties in the data?

[1] https://www.r-project.org/
[2] http://www.cyclismo.org/tutorial/R/
[3] http://tryr.codeschool.com/
[4] https://www.anaconda.com/distribution/
[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
[6] https://scikit-learn.org/stable/modules/manifold.html