


Submission deadline:

The submission deadline for this assignment is 5pm on Friday **6th March** 2020 (week 8).

Format

Your report should be in the form of an academic paper. If you use latex please use the latex template given below. If you use other word processing software please ensure that the format of your report matches the [latex format](#) .

Your report should be no more than 15 pages in length, A4, 12pt characters with single spacing. The 15 page limit includes any figures and bibliography.

Assignment details:

Your task is to investigate a data set using Principal Components Analysis (PCA), and other techniques from the IDA course.

You can choose any of the data set(s) from the list below but finding your own dataset to investigate is much preferable and there are bonus marks for using your own dataset. For example, in addition to numerical data sets there are many potential text data sets (Twitter, online reviews, etc) and these can be converted to numerical vector format using methods from the course.

If you decide to go the easier route and use some of the data sets provided, un-tar the relevant file and go to the corresponding folder. The folder contains the data set, as well as additional information about the data. Read the available information, especially the description of the features (data dimensions).

Depending on the software that you use you may need to pre-process the data, for example so that it contains only numerical features (dimensions) and the features are space-separated (not comma-separated). You may also need to worry about the range of values in each of the data dimensions.

Labelling scheme

To make the plots informative, you need to devise a labelling scheme for data points.

If the data can be classified into several classes (you can find this out from the data and feature description), use that information as the basis for your labelling scheme. In that case exclude the class information from the data dimensions that you analyse. Alternatively, you can make labels out of any dimension, e.g. by **quantising it into several intervals**. For example, if the data dimension represents age of a person, you can quantise it into 5 labels (classes) [child, teenager, young adult, middle age, old]. Associate the data labels with different markers and use the markers to show what kind of data points get projected to different regions of the visualisation plot.

You need to justify your choice of classes in the labelling scheme. What do you expect PCA to tell you about the relationship between these classes and the other dimensions in the data set? Are these relationships evident from projecting the data onto existing pairs of dimensions? What additional insights do you get from applying PCA?

You should:

- Learn as much as you can about your chosen data set(s) using the visualisation and clustering methods developed in the module.
- Use different data labelling schemes.
- In the case of visualisation, compare projections onto the principal components from PCA with straightforward projections onto the existing coordinates/dimensions.

Example

Before starting to work on the assignment, please carefully study the [example](#) that Professor Tino prepared using the boston database. To do this, first un-tar the file [boston.ex.tar.gz](#) and go to the folder "BOSTON.EX". The sub-folder "FIGURES" contains all the relevant figures as eps or gif files. Please consult the file "boston.read.me" in BOSTON.EX.

The report should describe experiments with your chosen data set along the lines of the "boston" example. In the labelling scheme, concentrate on more than one coordinate (dimension), e.g. in the "boston" example don't just consider the 'price' feature, but run separate experiments with 'per capita crime rate in the town', or 'pupil-teacher ratio in the town' instead of the 'price' coordinate.

Marking scheme

Your report should address the following factors. The figures in [square] brackets indicate the weight of each of these factors in the marking scheme:

- Introduction [15%]: Briefly describe the data. What are the potential classes in the data set? Demonstrate an understanding of the data by making hypotheses about what you expect your analysis to reveal about the data. Which software tools will you use? Briefly explain PCA and any other data analysis method that you use.
- Pre-processing [10%]: How did you pre-process the data? Why did you do it in this way?
- Labelling [15%]: What features (coordinates/dimensions) did you use for labelling the projected points with different markers? Why did you choose these classes?
- Projection onto existing dimensions [15%]: What interesting aspects of the data did you detect based on projection of the data onto pairs of existing dimensions?
- PCA [25%]: What interesting aspects of the data did you detect based on eigenvector and eigenvalue analysis of the data covariance matrix? What is the relationship between the principal components and the original data dimensions?
- Other analysis methods [20%]: What interesting aspects of the data did you detect based on other types of analysis of the data?
- Own data set [10%]: Bonus points for finding and using your own data set.

- **Total [110%]**

You should demonstrate that you understand the visualisation techniques used and that you are able to extract useful information about otherwise impractically high-dimensional data using dimensionality-reducing visualisation techniques.

The overall mark (in the range 0–110%) will be linearly scaled to the range 0–22% by multiplication by 0.2.