Data Science Techniques and Applications 2018-19

Coursework II: Multi-dimensional analysis and reduction

Amended text: see the changelog for the details

The goal of this coursework is to build on your analysis of a Kaggle[1] dataset to develop a Data analytics, applying the concepts seen in class. A light coding and a technical annex describing the method and the results on your dataset of choice will suffice.

The technical annex can assume all the concepts analysis described in Coursework I (CW I), which is available to the markers and will be read back-to-back with Coursework II.

If inclusion from CW I is needed, it won't count for plagiarism. However, such inclusions should be made as quotations (see example of quotation in the text of CW I), also to avoid Turnitin detection.

Your work, and the writing of the technical annex shall be organized in phases as follows.

Phase 1.

Reconsider the Kaggle dataset analysed in CW I, in particular the analysis of the most important dimensions.

Select three dimensions of interest, and project their values with appropriate matplotlib scatterplots.

Note: advanced python modules such as Bokeh and Dash are allowed but not required.

The three dimensions to be considered should be, in the domain of the dataset, a candidate predictor and an important dimension to predict, respectively.

For instance, in the Iris dataset we could consider Sepal length, Petal length and Sepal width. Classification is an extra dimension that can be represented, e.g., by colour of the points in the scatterplot.

Phase 2.

Write a simple Python program that loads the reduced dataset and performs Principal Component Analys on the 3-dimensional dataset using a Scikit-Learn[2].

As an illustration of how to invoke PCA, consider this example by Gaël Varoquaux.

Should the projected dataset contain an high number of datapoints, to the extent that PCA computation is unfeasible on your computer, you can reduce size by randomly selecting a fraction of the datapoints. Please explain your choices in the essay if you do so.

[1] https://en.wikipedia.org/wiki/Kaggle

[2] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Phase 3.
Describe the results of the analysis and comment them, possibly with a graphical display of the results (see again Varoquaux's example).

In view of possible ambiguities in the original text, any interpretation of the coursework will be considered valid; it won't affect marking.

Important dates: Please refer to dates and times on Moodle.

Submission: **an essay, with code shown inset and commented plus code ready to be run in a separate file.**

Please use your judgement on the right amount of data and length of presentation for a simple technical description.

Please edit your essay using mainstream formats (e..g., HTML, PDF, ODT, TEX, MD) that support your presentation style. Please also be advised that heavy Microsoft Office formats (e.g., DOCX, PPTX) could cause anomalies when displayed by Libre Office[3] so please refrain from using them. As a courtesy, please use an easy-to-read style similar to that of this document (Times New Roman font or similar, size 12 or bigger, 1.15 line spacing or higher, justified alignment).

Plagiarism: please be advised that Moodle deploys a state-of-the-art plagiarism detection software[4] to evaluate coursework submissions against both Web sources and other submissions, past and present. Each submission will be scored for originality; submissions with low originality might be discarded or penalized.
It is however possible to insert quotations by using appropriate typographic style and providing the reference:

> *This phrase is an example of a typographic style for citation and reference: it will be discounted by Turnitin analysis.*

Please make use of a formalized citation system and report articles and books you refer to, e.g. [Narayanan et al., 2011].

References

[Narayanan et al., 2011]
A. Narayanan, E. Shi, B. I. P. Rubinstein, 2011.

[3] https://www.libreoffice.org/
[4] https://turnitin.com/gateway/index.html

Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge
Proc. of the 2011 Int'l Joint Conference on Neural Networks.
https://arxiv.org/abs/1102.4374

Changelog

This text has been amended as follows.

1.
Phrase:

Take this example by Gaël Varoquaux as a model encoding.

Has been substituted with

As an illustration of how to invoke PCA, consider this example by Gaël Varoquaux.

2.
Phrase

in the example used for CW I we could consider age of the viewer and name of the favourite TV series.

Has been substituted with

in the Iris dataset we could consider Sepal length, petal length and sepal width. Classification is an extra dimension that can be represented, e.g., by colour of the points in the scatterplot.

3.
Phrase

In view of possible ambiguities in the original text, any interpretation of the coursework will be considered valid; it won't affect marking.

Has been added.