Take-home Assessment

1 Description

This final practical is based on the material covered in the lectures and previous practicals. You should write a practical report that should consist of a description and evaluation of the work done of not more than 2500 words excluding tables, graphs and images. The final practical will contribute 80% of the final mark. The deadline for handing in completed reports to student administration is **Friday 30th November 2018, 5pm**.

Additionally, you will need to submit your code (Jupyter notebook(s) or python script(s)) to the Moodle webpage. The assessors may run your code, but you will not be assessed on the quality of code writing. The assessment will be based on the report and the clarity of description of the work done and evaluation performed.

2 Dataset

The data set is a subset of CENSUS-INCOME (KDD) DATA SET.[1] This data set contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau.

You will be working with the `census/census-income.csv` file. The following modifications have been done to the original data:

- to speed up preprocessing, only about 10% of the original data is used in the practical;
- the data was extracted in such a way as to provide enough training instances for all classes;
- the target values are converted into binary $[0, 1]$ representation.

The explanation of the feature names and values in the original dataset can be found on the UCI Machine Learning Repository webpage, or in the accompanying `census/census-income.names` file.

3 Your task

Your task is to build a machine learning pipeline to predict the target value based on the other demographic and employment related variables in the dataset. The target value represents the level of income: below $\$50K$ (value 0) or above $\$50K$ (value 1).

Your implementation and report should include the following steps:

- Data exploration: note that the dataset contains a combination of categorical and numerical-valued features. It also contains a number of missing values. Explore the different features in

[1] `https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29`

the dataset (you might want to remove the features with mostly missing values), gain insights from the data and report your findings.

- Machine learning algorithms implementation: apply machine learning algorithms that you used in the previous practicals. Find out which algorithm works best and report your results.
- Evaluation: look into different ways and measures for evaluating the algorithms. You may consider looking into: accuracy, precision, recall, F_1, trade-offs, ROC curves and AUC. Report your results and present your findings for the best-performing ML algorithm.
- Visualisation and dimensionality reduction: look into dimensionality reduction. For instance, you may consider using PCA on a selected set of features, plotting a scatter plot of the components and colour-coding the points by a selected categorical feature.