Assignment 3: Mastering:
SVM, RandomForest, XGBoost
PCA, LLE, t-SNE

W. Kowalczyk

30/11/2021

Introduction

The purpose of this assignment is to master three most popular and powerful algorithms for classification tasks: SVM, RandomForest and XGBoost. Additionally, you should master three algorithms for dimensionality reduction and data visualization: PCA, LLE and t-SNE.
By "mastering algorithms" we mean here the ability of:

- installing these algorithms on your computers,
- downloading and preprocessing (when needed) relevant datasets,
- applying the algorithms to the data,
- demonstrating results.

You should use just one data set for the classification task and one (possibly the same) data set for the data visualization task. You are free to search the internet to find some interesting datasets, examples, blogs, etc. However, in your final submission, you should provide references to these sources. And, needless to say, your work should be substantially different than "copy & paste & get-it-working"!
As a starting point we recommend reading the following papers and visiting the sites:

Random Forests: Chapter 15 of the ESLII book:

https://web.stanford.edu/~hastie/Papers/ESLII.pdf

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

XGBoost:

https://xgboost.readthedocs.io/en/latest/tutorials/model.html

https://arxiv.org/pdf/1603.02754.pdf

https://github.com/dmlc/xgboost/blob/master/demo/guide-python/sklearn_examples.py

t-SNE and others:

https://www.youtube.com/watch?v=RJVL80Gg3lA&list=UUtXKDgv1AVoG88PLl8nGXmw

https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf

Possible sources of data:

https://www.kaggle.com/datasets

https://www.kdnuggets.com/datasets/index.html

https://archive.ics.uci.edu/ml/index.php

What to deliver?

Two self-contained Jupyter notebooks that cover your experiments for both tasks: one with SVM, RandomForest, XGBoost, and another one with the results of 3 visualization methods.

The notebook for the classification task should contain the following sections:

- The problem and the data description
- Experimental setup:
  - Data preprocessing (optionally),
  - Exploratory data analysis (optionally),
  - Parameter tuning.
- Results
- Conclusions

The notebook for the visualization task should contain data description (unless you use the same data set as for the classification task) and commented results of applying visualization algorithms to your data. Obviously, the code used for generating these visualizations should also be included in your notebook.

Evaluation Criteria:

- Quality of the code
- Quality of the reporting/descriptions (embedded in your notebooks!)
- Reproducibility (the code should work on other computers as well!)

DEADLINE: will be announced soon…