Evaluation of unsupervised classification algorithms on the MNIST dataset

1 Description

Classification of handwritten characters and digits is a difficult problem that has help to advance research on machine learning algorithms. The MNIST database [1], which is available from `http://yann.lecun.com/exdb/mnist/`, contains images of 70000 handwritten digits (60000 in the train set and 10000 in the test set. This dataset is extensively used to evaluate the accuracy of supervised classification methods. It can also be used to investigate the behavior of clustering algorithms and methods for feature selection.

2 Objectives

The goal of the project is to compare of different clustering algorithms on the MNIST dataset. This means, to apply the unsupervised learning methods without using any knowledge about the labels of the digits and then evaluate and visualize whether the clusters learned capture some pattern and information about the labels.

The student should: 1) Design any preprocessing or feature selection step to deal with the MNIST dataset 2) Define and apply at least three different clustering algorithms to the data 3) Design a validation method to evaluate to what extent the clusters capture any commonality between the images related to the digit they represent. 4) Use some sort of visualization to illustrate the differences between the clustering algorithms or the characteristics of the clusters learned.

As in other projects, a report should describe the characteristics of the design, implementation, and results. A Jupyter notebook should include calls to the implemented function that illustrate the way it works.

3 Suggestions

- The MNIST datase is very large. You may select only a subset of the data points (at least 1000 images).
- You may use the `keras` or `tensorflow` Python packages that have auxiliary function to download the MNIST dataset, or just download the data set from the link provided in this document.
- Implementations can use any other Python library.

References

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.