CSCI433/CSCI933: Machine Learning - Algorithms and Applications
Assignment Problem Set #2

Lecturer: Prof. Philip O. Ogunbona(philipo@uow.edu.au)
School of Computing and Information Technology
University of Wollongong

Due date: Saturday May 2, 6:00 p.m.

Introduction

Often the number of features collected in a machine learning problem is very large and can be represented as data in a large dimensional vector space. For instance one may need to solve a classification problem and the number of features collected may number in several hundreds or thousands. Hence the feature vector will be of a high dimension. Dealing with such data could be problematic because of the so-called curse of dimensionality. It may be the case that the information required to characterize the classification problem can be represented with a feature vector of much smaller dimension. In this situation the information characterising the problem lies in a low-dimensional manifold of the original vector space. The problem of dimensionality reduction is how to find this low-dimensional manifold.

In this assignment, you will study some of the non-linear dimensionality reduction methods (van der Maaten, Postma, & van den Herik, 2008) used in machine learning. You are to read, study, understand and replicate aspects of the paper by van der Maaten et al. (2008). The assignment gives you opportunity to generate and visualize artificial data and to work with both artificial and natural dataset. You will use the Python programming language and the libraries available for machine learning (scikit-learn), plotting and visualization (e.g. matplotlib, seaborn, etc.) to explore some of the methods of dimensionality reduction. You will be aiming to replicate the results obtained by the authors of the paper cited as (van der Maaten et al., 2008). There is also an extended version of the paper that describes how the artificial data was generated (van der Maaten, Postma, & van den Herik, 2009). This should help you when implementing code to generate the data. The two papers are included in the specification pack provided for this assignment.

What needs to be done

1. Read, study and understand the two papers. You are replicating the short paper (van der Maaten et al., 2008). The longer paper describes how to generate the artificial datasets and includes more details about the techniques.

2. Generate and plot (visualize) the artificial datasets `Swiss roll`, `Broken Swiss` and `Helix`. See for example Fig. 4 in van der Maaten et al. (2008). You will include the plot you generated in your report and write about it.

3. Download and prepare to use the natural datasets: `MNIST` and `Olivetti faces`. You can use the `scikit-learn` module in Python to download the `MNIST` and `Olivetti faces` datasets as shown in this code snippet (or read the scikit-learn documentation).

```
.
.
import sklearn
from sklearn import datasets
from sklearn.datasets import fetch_openml

mnist_data = fetch_openml('mnist_784', version=1, return_X_y=True)
olivetti_faces = sklearn.datasets.fectch_olivetti_faces
.
.
```

Ensure that you really understand the organisation of the datasets. This is absolutely important - check the size, shape, etc.

4. Using Python programming language, implement the dimensionality reduction methods: PCA, Kernel PCA, Autoencoders, LLE (see Table 2 in van der Maaten et al. (2008)) as described in the paper. Use the parameter settings provided in the paper. As a hint, these techniques are implemented in the scikit-learn Python machine learning library.

5. Using generalization errors of 1-Nearest Neighbour classifier trained on the datasets, compare the performance of the dimensionality reduction methods mentioned in item (4) above. Your results will be presented as in Table. 4 of the paper for the datasets listed in items (2 and 3) above.

6. Your report will be presented in a conference paper format (see accompanying template) and should detail your understanding of theory of the techniques and experiments in the assigned paper. You will describe the techniques in your own words with appropriate equations. When you write an equation, the meaning of the symbols must be explained as well as the intuition behind the equation itself. Your report MUST not be more than nine (9) pages in the format specified by the template.

7. Please cite any other paper or book you have read in gaining deeper understanding of the concepts and methods.

What needs to be submitted

- You will prepare a "zip" or "rar" file containing your report (9-page PDF file) and Python code (named : "dim_reduc.py") file.

- Your code must run from command line as:

```
python3 dim_reduc.py
```

  and write your results to standard output (stdout).

- Submit the "zip" or "rar" via Moodle dropbox provided on or before the deadline.

References

van der Maaten, L. J. P., Postma, E. O., & van den Herik, H. J. (2008). *Dimensionality reduction : A comparative review.* online. Retrieved March 2020, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.5472&rep=rep1&type=pdf

van der Maaten, L. J. P., Postma, E. O., & van den Herik, H. J. (2009). *Dimensionality reduction : A comparative review.* online. Retrieved March 2020, from https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf