Linear versus Non-linear Dimensionality reduction

Feature extraction/engineering as preproccesing of audio/images/statistical match data

Daniel Runge Petersen, Gustav Svante Graversen, Lars Emanuel Hansen, Raymond Kacso, Sebastian Aaholm

Computer Science, cs-22-dat-5-05, 2022-12

Semester Project

Composed and typeset by the authors using the LaTeX Document Preparation System, based on the AAU report template by Daniel Runge Petersen [1].

Electronics and IT
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Linear versus Non-linear Dimensionality Reduction

Abstract:
This is the abstract...

Theme:
Theoretical data analysis and modeling

Project Period:
Fall Semester 2022

Project Group:
cs-22-dat-5-05

Participant(s):
Daniel Runge Petersen
Gustav Svante Graversen
Lars Emanuel Hansen
Raymond Kacso
Sebastian Aaholm

Supervisor(s):
Alexander Leguizamon Robayo

Copies: 1

Page Numbers: 19

Date of Completion:
October 5, 2022

Contents

Preface

This is the preface. You should put your signatures at the end of the preface.

Aalborg University, October 5, 2022

Author 1
<username@student.aau.dk>

Author 2
<username@student.aau.dk>

Author 3
<usernamestudent.aau.dk>

Author 4
<username@student.aau.dk>

Author 5
<username@student.aau.dk>

Author 6
<username@student.aau.dk>

Chapter 1

Introduction

Chapter should probably contain: The initial problem (If we have one), motivation, and the scope or background of the project or theme. Report outline at the end.

In this project we will study some common methods of dimensionality reduction. Inspired by the MNIST dataset for digit recognition, we wish to test similar algorithm on pokemon image data.

Note: Unique problem in pokemon classification based on colors is the handling of SHINY pokemon. How will we handle this?

Keywords: dimensionality reduction, linear methods, nonlinear methods, MNIST, pokemon, pokedex, Computer Vision (CV), Machine Learning, machine intelligence (artificial intelligence).

Use sources [2] and [3] for superficial overview and explanations of umbrella terms.

On theory driven projects

> The overall purpose of the project module is for the student to acquire the ability to analyze and evaluate the application of methods and techniques within database systems and / or machine intelligence to solve a specific problem. **This includes analyzes of the formal properties of the techniques and an assessment of these properties in relation to any requirements for the solution to the specific problem**. [...]
>
> In this project module, the project work is primarily driven by theoretical and analytical considerations about the methods and techniques used. For a specific problem area, a project could, for example, be based on specific performance requirements for the developed software solution, and the project work can thus be guided by the solution's algorithmic time / space complexity as well as formal analyzes and considerations of its theoretical properties and performance guarantees. [4]

1.1 Motivation

High-dimensional data (i.e. data that requires more than three dimensions to be represented) can often be difficult to work with. Not only is it difficult to interpret and visualize but also can require a high use of computational resources. For these (and many more) reasons, it is important to study dimensionality reduction methods. These methods are usually used in exploratory data analysis and for visualization purposes.

The most usual methods of dimensionality reduction are **linear methods**. These methods might assume that the features in the original data are independent and they can produce reduced data by a linear combination of the original data. These assumptions might not apply to all datasets. In fact, there are cases in which linear methods do not capture important features of a dataset. For these cases one can use **nonlinear methods**/ These methods can be used for more general cases while preserving important information from data.

For now this is taken directly from the project proposal. We may rewrite this partially at a later time.

1.2 Report outline

Is this outline in line with the theoretical theme?

The proposed report structure is as follows:

- **Introduction** - This chapter
- **Problem Analysis** - Chapter 2
- **Methodology / Theory and Methods** - Chapter 3
- **Results** - Chapter 4
- **Discussion** - Chapter 5
- **Conclusion** - Chapter 6

The introduction describes the initial problem and the motivation for the project.

The Problem Analysis chapter dives into the initial problem and leads to a final problem statement.

The Methodology chapter describes the methods and theory used to explore the problem statement. It also describes the data used and how it was collected/created.

The Results chapter is an evaluation of the results of the project.

The Discussion chapter is a discussion of the results and the project as a whole.

The Conclusion chapter provides a summary of the project and the results. It also provides perspective and reflection on the project and the process.

Chapter 2

Problem Analysis

This chapter contains the theoretical background for the project leading to the problem statement.

Write about: linear methods, non linear methods, computer vision and machine learning and types of image data. Also, write about the different types of problems that can be solved with machine learning? For example, classification, regression, clustering, etc.?

2.1 ML Pipeline

When working with machine learning, pipelines helps to understand the different steps involved in the process. The following figure 2.1 shows the different steps that are involved in the process of building a machine learning model. ML pipelines are usually divided into three main steps: data, feature extraction, model training and model evaluation. The following sections will describe each of these steps in more detail.

In the figure 2.1 the box with the name "data" is representing the input to the machine learning pipeline. The data can be in different formats, such as images, text, audio, video, etc. The data is usually stored in a database or in a file. This is also the step data cleaning is done, which is the process of removing or replacing missing values in the data. In this step the data is processed into a format that can be used by the machine learning algorithm by for example taking data out a picture.

Feature extraction (in 2.1 FE) is the process of extracting features from the data. This is done by for example dimensionality reduction or labelling correlated data, that is clustered together after the feature extraction. There are different methods that can be used on feature extraction, the ones used in the project will be discussed in another section. After going through the feature extraction the data os now ready to be out into a ML model.

write this section

The box "parameters" in figure 2.1 describes the parameters that are used in the model. The parameters are the values that are used to train the model. The parameters are usually set by the user, but can also be set by the model itself. The parameters uses various metrics such as accuracy, precision, speed and memory usage to detemain if the model is the correct fit. To see if the model is good fit, the

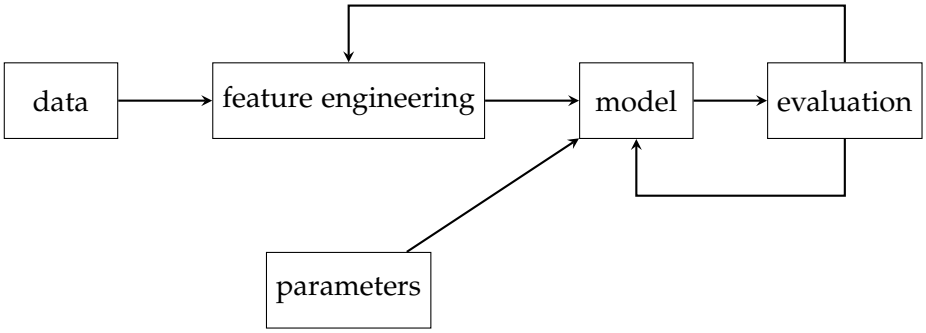

Figure 2.1: ML pipeline

model is evaluated in the evaluation stage.

Model training is the process of training the model with the data. This is done by using a training set and a validation set. The training set is used to train the model to predict the output with highest possible accuracy. The validation set is used to evaluate the model. It is here where the chosen parameters determine the fit of the model. Model evaluation is the process for analysing the models performance. This is done by using a test set. The test set is used to evaluate the model. The test set is not used to train the model. The evaluation is done by comparing the predicted values with the actual values that was to be expected. After the model has been evaluated, new models or feature extraction can be used to try and improve the models performance [5].

2.2 Problem Statement

This project explores the impact of data preproccesing on the performance of machine learning using a logistic regression model versus Convolutional Neural Network (CNN) for the computer vision problem of image classification and recognition. The data preprocessing is done through dimensionality reduction on augmented data from the Modified National Institute of Standards and Technology (MNIST) database, and the machine learning models are trained on the reduced data. The performance metrics used to evaluate the models are accuracy, precision, recall, and F1 score. and of course explainability and speed/size of the models.

Versus!

This project explores the impact of data preprocessing on the performance of a logistic regression machine learning model, for the computer vision problem of image classification and recognition. By data preprocessing is meant dimensionality reduction on augmented data, comparing linear and non-linear dimensionality reduction techniques. The machine learning model is trained on the dimensionality reduced data and the performance is evaluated using accuracy, precision, recall, and F1 score. The performance is further measured against a CNN model, to compare speed, size, and explainability. The data used is the MNIST database.

methods?

- PCA + logistic regression vs

- PCA + CNN vs
- LDA + logistic regression vs
- LDA + CNN vs
- kernel PCA + logistic regression vs
- kernel PCA + CNN

2.2.1 Tools

Data preproccessing, data augmentation and feature engineering. Use Keras to build a Machine Learning (ML) model. Explainability - Neural Network (NN) vs other ML algorithms.

remove this eventually

Notes to self: Humans vs computers in NN. Why are humans good with little training, and computers only accceptable with much more training? Consider perhaps domains (recongnizing epsilon vs. recognizing a 3)

As part of the pipeline, show the images that the models misguessed?

As part of the pipeline, normalize the data in a way that's not dimensionality reduction, but that's still preprocessing? (e.g. subtract mean, divide by standard deviation).

How do we determine recall and precision for logistic regression?

Chapter 3

Methodology

Write about methodology of the project.

3.1 Overview

What we explore is the impact of especially different augmentations of the original data on the application of different dimensionality reduction techniques, and how this affects the performance of the models.

dataset mnist

pre-preprocessing normalization, data augmentation

preprocessing linear and non linear dimensionality reduction (PCA, LDA and Kernel PCA)

models logistic regression and CNN

evaluation accuracy, precision, recall, f1-score, speed/run time, memory usage, and model size

3.1.1 Data collection

The data used is the MNIST database, which is a collection of handwritten digits. The data is split into a training set of 60,000 images and a test set of 10,000 images. The images are 28x28 pixels, and each pixel is represented by a value between 0 and 255, where 0 is black and 255 is white. The images are grayscale, and the values are the intensity of the pixel. The images are labeled with the digit they represent, and the labels are integers between 0 and 9.

3.1.2 Data pre-preprocessing

The data is normalized by dividing each pixel value by 255, so that the values are between 0 and 1. The data is then augmented by rotating the images by 90, 180 and 270 degrees, and flipping the images horizontally and vertically. This results in 10 times as much data as the original MNIST dataset.

This is just placeholder, but something like this is what we want to do.

3.1.3 Data preprocessing

The data is then preprocessed by applying linear and non-linear dimensionality reduction techniques. The linear dimensionality reduction techniques are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The non-linear dimensionality reduction technique is Kernel Principal Component Analysis (kPCA). The data is reduced to 2 dimensions, so that it can be visualized.

An automation pipeline is created based on Figure 3.1.

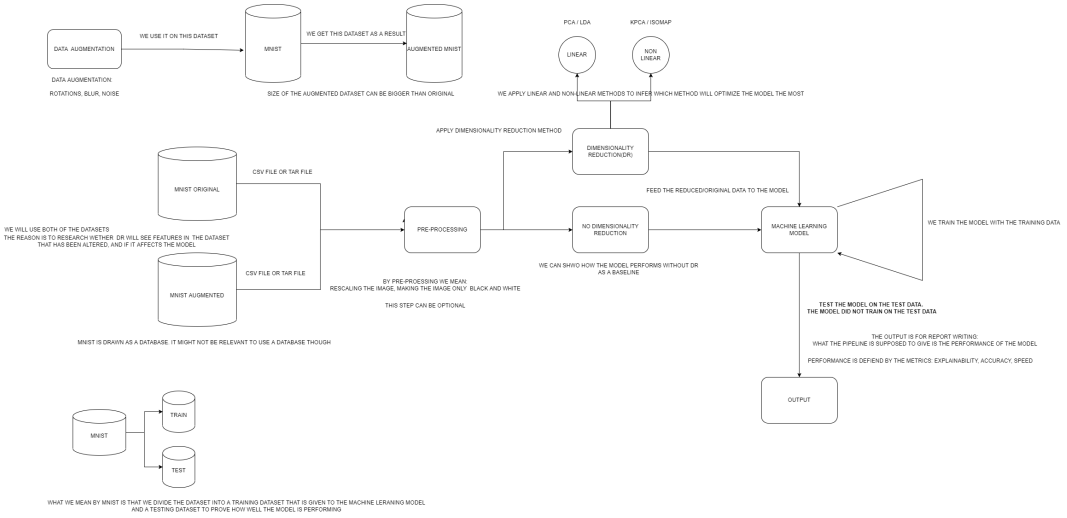

Figure 3.1: Python project pipeline model.

write about the pipeline

3.2 Metrics

In this project the differens dimmensionality reduction methods are evaluated based on how much it improves the metrics of our models. The following paragraphs will go through these chosen metrics. After that the reason for the choices will be explained.

3.2.1 The chosen metric

There are 4 metrics which will be used to evaluate the FE through the model. These metrics are acurracy, precision, recall, F1-score. And there is 2 metric used to evaluate FE directly. Here the speed/computations is used to evaluate the effeciency. The other metric used is how well the dimmensionality reduction clusters the data, this will be calculated for each class as the average distance between class members divided by the average distance between all data points.

3.2.2 Reason for choice of metrics

The four model metrics were chosen because they the metric which generally describes how good the model does in infering the data. These are all essentially just variations on how many does the model get correct devided by diffrent parts of the confusion matrix. The last 2 metrics was chosen to show the tradeoff most models will have in the more time it takes the better results. This makes it possible to show which metric will actually be most useful in most cases since accurracy is often important only to a certain point. The second FE metric also takes into acount how well the data is clustered after aplying the FE which gives insight into how well the FE works in a vacuum and how much performance a actual model whill achiewe from data which has more defined clusters.

Chapter 4

Results

Describe the results of the project.

Chapter 5

Discussion

Discuss the results from chapter 4 and compare them to the problem statement in section 2.2. Also, discuss the methodology and the theoretical background in chapter 3. Finally, discuss the project as a whole and the process of the project.

What went well? What could have been done better? What would we do differently next time? Perhaps include thoughts on UN sustainability goals.

Chapter 6

Conclusion

Based on the discussion in chapter 5, the results from chapter 4 and the problem statement in section 2.2, the following conclusions can be drawn:

This chapter contains the concluding remarks of the project. It is based on the discussion in chapter 5, the results from chapter 4 and the problem statement in section 2.2. The chapter concludes with a reflection and perspectives for future work.

Acronyms

CNN Convolutional Neural Network. 4

kPCA Kernel Principal Component Analysis. 8

LDA Linear Discriminant Analysis. 8

ML Machine Learning. 5

MNIST Modified National Institute of Standards and Technology. 4, 7

NN Neural Network. 5

PCA Principal Component Analysis. 8

Bibliography

[1] Daniel Runge Petersen. *AAU-Dat templates*. URL: https://github.com/AAU-Dat/templates (visited on 08/17/2022).

[2] *Machine Learning*. IBM. URL: https://www.ibm.com/cloud/learn/machine-learning (visited on 09/27/2022).

[3] *What is computer vision?* IBM. URL: https://www.ibm.com/topics/computer-vision (visited on 09/27/2022).

[4] *Theory-driven Data Analysis and Modeling*. Aalborg University. URL: https://moduler.aau.dk/course/2022-2023/DSNDATB521 (visited on 09/27/2022).

[5] *Machine Learning Pipeline*. javapoint. URL: https://www.javatpoint.com/machine-learning-pipeline (visited on 10/04/2022).