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: 15

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 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
- t-SNE + logistic regression vs
- t-SNE + CNN

2.1.1 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 theoretical background and methodology of the project.

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 2.1. 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 2.1, 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 2.1. The chapter concludes with a reflection and perspectives for future work.

Acronyms

CNN Convolutional Neural Network. 3

ML Machine Learning. 4

MNIST Modified National Institute of Standards and Technology. 3

NN Neural Network. 4

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).