Interim Final Year Project Report

Lorcan Williamson

October 7, 2019

2

Contents

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Introduction and Project Description | | 5 |
|  | 1.1 | Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 5 |
|  | 1.2 | Project Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . | 5 |
| 2 | Literature Survey | | 7 |
|  | 2.1 | Problem Categories . . . . . . . . . . . . . . . . . . . . . . . . . | 7 |
|  | 2.2 | Classification Problems . . . . . . . . . . . . . . . . . . . . . . . | 7 |
|  |  | 2.2.1 Machine Vision . . . . . . . . . . . . . . . . . . . . . . . | 8 |
|  | 2.3 | Regression Problems . . . . . . . . . . . . . . . . . . . . . . . . | 8 |
| 3 | Theory | | 9 |
|  | 3.1 | . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 9 |
| 4 | Outline Design | | 11 |
|  | 4.1 | Classification Problem 1: Student Performance . . . . . . . . . . | 11 |
|  | 4.2 | Classification Problem 2: Detecting Pneumonia . . . . . . . . . . | 12 |
|  | 4.3 | Regression Problem: Bitcoin Prices . . . . . . . . . . . . . . . . | 12 |
| 5 | Action plan | | 13 |
| 6 | Requirements | | 15 |

3

4 CONTENTS

1. Introduction and Project Description

1.1 Introduction

Machine learning has advanced hugely through deep learning in recent years. This project is to explore the landscape using Python machine learning frameworks such as Scikit-Learn, Tensorflow and Keras. This will be done by looking at common machine learning problems and the steps to develop solutions using these frameworks.

1.2 Project Outline

This project will examine how code is developed to solve problems in the two main use cases of machine learning; classification, and regression. The problems will be chosen to examine the most model architectures and development prac-tices in as few problems as possible. To do this three problems have been chosen, two classification and one regression. The classification problems involve one image recognition problem1 and one feature based problem2. Two classification problems were chosen to offer a use case for CNNs3, which are used almost ex-clusively for machine vision problems. Another classification was also chosen

* Detecting Pneumonia in chest X-rays
* Determining student performance based on a number of factors
* Convolutional Neural Network

5

* CHAPTER 1. INTRODUCTION AND PROJECT DESCRIPTION

so that other architectures could be examined, as it is uncommon to use CNNs on anything but image based problems. The regression problem4 was chosen to look at time-series problems, which are very often regression problems. This al-lows a look at architectures such as RNNs5 and LSTM6 networks, a special type of RNN. The datasets were chosen to try represent realistic problems machine learning would be used to solve instead of using one of the more common bench-marking datasets such as CIFAR-10[1] or Iris, to create a more realistic example of the code development process, especially data preparation, which is often al-ready done for common benchmarking datasets.

* Predicting Bitcoin prices based on historic data
* Recurrent Neural Networks

6Long-Short Term Memory

2. Literature Survey

2.1 Problem Categories

This project focuses on code development for machine learning problems using python. Only supervised learning problems are being examined in this, due to supervised learning for many non-linear classification problems as discussed in Bradley C Love’s paper comparing supervised and unsupervised learning for clas-sification problems[2]. From Soren’s article on PracticalAI.io[3] supervised ma-chine learning problems fall into two main categories; classification, and regres-sion.

2.2 Classification Problems

Classification problems involve the class of a presented input from a discrete number of predefined classes (labels), such as determining whether an image is of a dog or a cat, or determining the species of a plant given measurements taken from it. S.B Kotsianstis[4] discusses many of the common machine learn-ing algorithms for use on classification problems, such as decision trees/random forests (DTs/RFs), support vector machines (SVMs), and artificial neural net-works (ANNs), as well as many others.

7

8 CHAPTER 2. LITERATURE SURVEY

2.2.1 Machine Vision

Machine Vision covers a large number of problems, such as image-recognition. This is an inherently classification type problem, rather than regression, and is one of the most common uses of machine learning in machine vision. In their 2012 pa-per ImageNet Classification with Deep Convolutional Neural Networks[5], Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton showed that CNNs could achieve excellent results on image recognition tasks when compared to other ma-chine learning or deep learning algorithms. This is what led to the choice to have two classification problems, an image-recognition problem and one other. With only an image based problem there would have been no need to look at any other algorithms, as CNNs outperform them on these types of tasks, and with a non machine vision task then CNNs would not have been very applicable.

2.3 Regression Problems

Regression problems are problems of predicting values rather than classes, such as stock prices. Some regression problems can be converted to classification prob-lems by quantising the output values into classes. An empirical comparison of machine learning models for time series forecasting (2010)[6] compares a large number of models for forecasting values based on historic time-series data. This should be very useful for the regression problem examined in this project.

3. Theory

3.1

9

10 CHAPTER 3. THEORY

4. Outline Design

This project will look at three different problems, two classification and one re-gression, and how one may go about solving them using Python machine learning frameworks. This will cover code development from examining and preparing the training data, through selecting and optimising learning algorithms to test-ing the chosen solution. So far the datasets for the problem have been selected, and examination of them has begun. The frameworks used for this project will be Scikit-Learn1 and Keras2, a high-level API that can be built on top of, and is available in, Tensorflow.

4.1 Classification Problem 1: Student Performance

The first problem examined is a classification problem predicting students per-formance in final year[7]. This dataset has quite high dimensionality3, so will probably benefit from some dimensionality reduction technique such as principal PCA4 or some similar technique[8]. It should be noted that one of the features of the dataset, feature 31, the first/second period grade might be ignored, due to it trivialising the problem given how strongly it correlates to the final grade. It may

* https://scikit-learn.org/stable/index.html
* https://www.tensorflow.org/guide/keras

3The dataset contains 31 features

* Principal Component Analysis

11

12 CHAPTER 4. OUTLINE DESIGN

be examine however, as it will allow for very easy visualisation of the data and the way in which various classification algorithms separate it.

4.2 Classification Problem 2: Detecting Pneumonia

The second classification problem involves detecting pneumonia in patients from images of chest x-rays[9]. The images are binary classified, into Normal ans Pneu-monia. The Pneumonia images include cases of both viral and bacterial, but no distinction is made. There are 5216 images in the training data, 1341 examples of a normal x-ray, and 3875 examples of an x-ray with pneumonia. This means some class balancing technique such as those discussed in Mateusz Buda1’s, Atsuto Maki’s, and Maciej A. Mazurowski’s 2018 paper on class imbalance problems in CNNs[10] may be required.

4.3 Regression Problem: Bitcoin Prices

The last problem examined is a regression problem, predicting future trading prices of Bitcoin based on historical data[11]5. The data included spans from January 2012 to August 2019, in one minute intervals, though some gaps exist in the data. There are also missing values for some entries. This may not be an issue, but if it is then some of the techniques discussed in A Review of Missing Values Handling Methods on Time-Series Data[12] may be useful.

* Available under Creative Commons license (https://creativecommons.org/licenses/by-sa/4.0/)

5. Action plan

13

14 CHAPTER 5. ACTION PLAN

6. Requirements

All software required for this project is free and open-source. There are no real

hardware requirements, though an Amazon web-services EC2 instance will be

used to speed up training times.

15

16 CHAPTER 6. REQUIREMENTS

Bibliography

1. A. Krizhevsky, “Learning multiple layers of features from tiny images,” 2009”.
2. B. C. Love, “Comparing supervised and unsupervised category learning,” Psychonomic bulletin & review, vol. 9, no. 4, pp. 829–835, 2002.
3. S. D. (2017), [Online]. Available: https://www.practicalai.io/ categorizing-machine-learning-problems/.
4. S. Kotsiantis, “Supervised machine learning: A review of classification techniques,” 2007.
5. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” pp. 1097–1105, 2012.
6. N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, “An empir-ical comparison of machine learning models for time series forecasting,” Econometric Reviews, vol. 29, no. 5-6, pp. 594–621, 2010.
7. P. Cortez and A. Silva, “Using data mining to predict secondary school student performance,” 2008.
8. L. Van Der Maaten, E. Postma, and J. Van den Herik, “Dimensionality re-duction: A comparative,” J Mach Learn Res, vol. 10, no. 66-71, p. 13, 2009.

17

18 BIBLIOGRAPHY

1. D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., “Identifying medi-cal diagnoses and treatable diseases by image-based deep learning,” Cell, vol. 172, no. 5, pp. 1122–1131, 2018.
2. M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” Neural Networks, vol. 106, pp. 249–259, 2018.
3. Zielak. (2019), [Online]. Available: https : / / www . kaggle . com / mczielinski/bitcoin-historical-data.
4. I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, “A review of missing values handling methods on time-series data,” in 2016 Inter-national Conference on Information Technology Systems and Innovation (ICITSI), IEEE, 2016, pp. 1–6.