Machine Learning for Investing Behaviours on Cryptocurrencies

Chengkai Lu

Submitted for the Degree of Master of Science in

Data Science and Analytics with a Year in Industry

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

May 26, 2018

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count:

Student Name:

Date of Submission:

Signature:

Abstract

Your abstract goes here.

Contents

1 Introduction

1.1 Motivation

1.2 Aims And Objectives

1.3 Project Structure

1.3.1 Technologies

1.3.2 Programme Structure

1.3.3 Report Structure

2 Background Research

2.1 Correlation Analysis

2.2 Dimensionality Reduction

It is always challenging to analyse a dataset with high-dimensional data points. Due to the curse of dimensionality, which was discovered by Richard Ernest Bellman in 1961, higher dimensional space increases the difficulties of analysing and organising data exponentially[1]. Especially in machine learning, given a certain number of samples, the accuracy of predictions on these samples will increase followed by the rising dimensions to a peak but then gradually drop. This is known as Hughes phenomenon[4].

In order to reduce the dimensionality, there are two approaches can be implemented:

- Feature Selection:
- Feature Extraction:

The reasons and benefits of executing dimensionality reduction can be summarised as follow:

1. Computational efficiency
2. Statistical generalisation
3. Better explanation

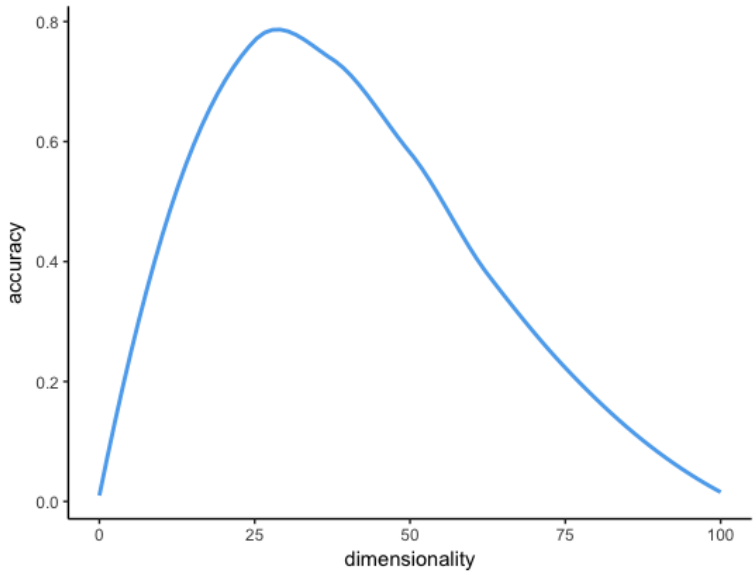

Figure 1: Hughes phenomenon

2.2.1 Principal Component Analysis

2.3 Clustering

Clustering, cluster analysis or data segmentation is a non-parametric algorithm in the subtree of unsupervised learning. It is used to separate data into different groups using their natural dissimilarities. Unlike supervised learning, this type of learning algorithms does not have any indicator for assessing the quality of results, and this means that it does not have any meaning or objective itself. Instead, it discovers the distribution of data and uses the definition given by people who have the specific domain knowledge. By giving the rules for partitioning data self-defined meanings, useful information can be obtained and utilised in different domains[3].

Generally,

2.3.1 K-means Clustering

An example of a reference: [2].

References

[1] Wikipedia contributors. Curse of dimensionality — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Curse_of_

dimensionality/. [Online; accessed 26-May-2018].

[2] Adam Gibson and Josh Patterson. *Deep Learning - A Practitioner's Approach*. O'Reilly Media, first edition, 2017.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition, 2009.

[4] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, January 1968.