CT4101

Machine Learning

Name: Andrew Hayes
E-mail: `a.hayes18@universityofgalway.ie`
Student ID: 21321503

2024–09–14

Contents

1 Introduction **1**

1.1 Lecturer Contact Details 1

1.2 Grading 1

1.3 Module Overview 1

1.3.1 Learning Objectives 1

2 What is Machine Learning? **1**

2.1 Data Mining 3

2.2 Big Data 3

1 Introduction

1.1 Lecturer Contact Details

- Dr. Frank Glavin.
- `frank.glavin@universityofgalway.ie`

1.2 Grading

- Continuous Assessment: 30% (2 assignments, worth 15% each).
- Written Exam: 70% (Last 2 year's exam papers most relevant).

1.3 Module Overview

Machine Learning (ML) allows computer programs to improve their performance with experience (i.e., data). This module is targeted at learners with no prior ML experience, but with university experience of mathematics & statistics and **strong** programming skills. The focus of this module is on practical applications of commonly used ML algorithms, including deep learning applied to computer vision. Students will learn to use modern ML frameworks (e.g., scikit-learn, Tensorflow / Keras) to train & evaluate models for common categories of ML task including classification, clustering, & regression.

1.3.1 Learning Objectives

On successful completion, a student should be able to:

1. Explain the details of commonly used Machine Learning algorithms.
2. Apply modern frameworks to develop models for common categories of Machine Learning task, including classification, clustering, & regression.
3. Understand how Deep Learning can be applied to computer vision tasks.
4. Pre-process datasets for Machine Learning tasks using techniques such as normalisation & feature selection.
5. Select appropriate algorithms & evaluation metrics for a given dataset & task.
6. Choose appropriate hyperparameters for a range of Machine Learning algorithms.
7. Evaluate & interpret the results produced by Machine Learning models.
8. Diagnose & address commonly encountered problems with Machine Learning models.
9. Discuss ethical issues & emerging trends in Machine Learning.

2 What is Machine Learning?

There are many possible definitions for "machine learning":

- Samuel, 1959: "Field of study that gives computers the ability to learn without being explicitly programmed".
- Witten & Frank, 1999: "Learning is changing behaviour in a way that makes *performance* better in the future".
- Mitchelll, 1997: "Improvement with experience at some task". A well-defined ML problem will improve over task T with regards to **performance** measure P, based on experience E.
- Artificial Intelligence $\neq$ Machine Learning $\neq$ Deep Learning.
- Artificial Intelligence $\not\supseteq$ Machine Learning $\not\supseteq$ Deep Learning.

Machine Learning techniques include:

- Supervised learning.
- Unsupervised learning.
- Semi-Supervised learning.
- Reinforcement learning.

Major types of ML task include:

1. Classification.
2. Regression.
3. Clustering.
4. Co-Training.
5. Relationship discovery.
6. Reinforcement learning.

Techniques for these tasks include:

1. **Supervised learning:**
    - **Classification:** decision trees, SVMs.
    - **Regression:** linear regression, neural nets, k-NN (good for classification too).
2. **Unsupervised learning:**
    - **Clustering:** k-Means, EM-clustering.
    - **Relationship discovery:** association rules, bayesian nets.
3. **Semi-Supervised learning:**
    - **Learning from part-labelled data:** co-training, transductive learning (combines ideas from clustering & classification).
4. **Reward-Based:**
    - **Reinforcement learning:** Q-learning, SARSA.

In all cases, the machine searches for a **hypothesis** that best describes the data presented to it. Choices to be made include:

- How is the hypothesis expressed? e.g., mathematical equation, logic rules, diagrammatic form, table, parameters of a model (e.g. weights of an ANN), etc.
- How is search carried out? e.g., systematic (breadth-first or depth-first) or heuristic (most promising first).
- How do we measure the quality of a hypothesis?
- What is an appropriate format for the data?
- How much data is required?

To apply ML, we need to know:

- How to formulate a problem.

- How to prepare the data.
- How to select an appropriate algorithm.
- How to interpret the results.

To evaluate results & compare methods, we need to know:

- The separation between training, testing, & validation.
- Performance measures such as simple metrics, statistical tests, & graphical methods.
- How to improve performance.
- Ensemble methods.
- Theoretical bounds on performance.

2.1 Data Mining

Data Mining is the process of extracting interesting knowledge from large, unstructured datasets. This knowledge is typically non-obvious, comprehensible, meaningful, & useful.

The storage "law" states that storage capacity doubles every year, faster than Moore's "law", which may results in write-only "data tombs". Therefore, developments in ML may be essential to be able to process & exploit this lost data.

2.2 Big Data

Big Data consists of datasets of scale & complexity such that they can be difficult to process using current standard methods. The data scale dimensions are affected by one or more of the "3 Vs":

- **Volume:** terabytes & up.
- **Velocity:** from batch to streaming data.
- **Variety:** numeric, video, sensor, unstructured text, etc.

It is also fashionable to add more "Vs" that are not key:

- **Veracity:** quality & uncertainty associated with items.
- **Variability:** change / inconsistency over time.
- **Value:** for the organisation.

Key techniques for handling big data include: sampling, inductive learning, clustering, associations, & distributed programming methods.