A breathtaking title

author: Joel Sjöberg 38686
Masters thesis in Computer Science
Supervisor: Luigia Petre
The Faculty Of Science And Engineering
Åbo Akademi University
2020

Contents

Foreword

Backword

Insert joke about consuming coffee here

Abstract

But not abstract enough.

Chapter 1

Introduction

Please!! [**?**]

Chapter 2

Project Description

For this thesis we will consider a dataset consisting Raman Spectrum gathered from 24 different tumor samples, each belonging to a unique patient. Several samples were collected from a select few of the patients with a variable size.

To determine the validity of this experiment, k-means clustering is performed on each sample pair for

Chapter 3

Theoretical Background

Within this chapter can be found the essential mathematical theory on which this thesis is based. The concepts are considered fundamental for understanding the methods applied in our project. It begins by covering necessary mathematical theory required to understand data representations and handling within Machine Learning (ML). It then proceeds by defining common concepts in machine learning and the subject of supervised and unsupervised learning.

3.1 Mathematical Foundations

Machine learning is founded on statistical and linear algebraic theory. Moreover, certain methods within ML is heavily dependent on calculus to compute the change required to reach optimal solutions. The relevant theory for these concepts, accompanied by necessary examples will be covered in this section.

3.1.1 Set Theory

Set theory is a useful schematic originally presented by Georg Cantor in **[YEAR] [SOURCE HERE]**. Set theory concerns the theory of sets within the universe and allows for formalization of collections of elements e.g. the set of all natural numbers $\mathbb{N}$ is the set containing all whole numbers greater than or equal to 0. The cardinality of the set is ∞ as there is an infinite number of natural numbers expressed formally as 3.1.

$$|\mathbb{N}| = \infty \tag{3.1}$$

The elements bellonging to a certain set is denoted by $2 \in \mathbb{N}$. Collections of sets bound by the operators $\cup$ (Union) and $\cap$ (intersection) produce sets of their own. Thus the following statements are theorems of set theory:

$$\mathbb{N} \cup \mathbb{Z} \equiv \mathbb{Z} \tag{3.2}$$

$$\mathbb{N} \cap \mathbb{Z} \equiv \mathbb{N} \tag{3.3}$$

Formally, the sets are collections of elements not limited to numbers, sets are primarily collections whose elements are devoid of order. Therefore we may have sets of items, , datatypes, people etc. The set devoid of elements is the empty set $\emptyset$. This set is in contrast to the universal set U containing all elements in the universe. The complement of a set is the set containing all elements in the universe excluding the complemented set. Form this reasoning follows the following theorems of set theory.

$$\emptyset^c \equiv U \tag{3.4}$$

$$U^c \equiv \emptyset \tag{3.5}$$

3.1.2 Statistics

With ML we have the capability to analyze and develop models for systems or phenomena within them without rigorous definition of said systems. This may be achieved by gathering data which in some way describes the system in question. A collection of data is called a dataset, datasets consists of examples which may range from single valued numbers to multi-dimensional tensors. Formally we say a dataset X is a subset of an unknown Population X' which includes every possible example x.

Creating a model for predicting the systems behavior for the entire population is the goal of ML. Empirically it is rare to access data from the entire population E. Instead a subset of examples $x' \sim E_X$ is drawn from the population distribution E_X for use within the ML-model.

3.1.3 Linear Algebra

Linear algebra is founded on the theory of tensors. ML uses this theory immensely as the models produced in it are collections of numbers stored within matrices or multidimensional tensors. It is therefore vital to understand the preliminary concepts and terms within linear algebra.

A vector...

3.1.4 Calculus

3.2 Machine Learning

Machine learning is the practice of computing models for relationships between sets of data. The field has garnered significant interest within akademia and industry alike due to the promising result in a number of applications. Within the field there are mainly two paradigms for learning: Supervised learning (using labeled data to approximate models) and unsupervised learning (finding patterns within the data itself).

Models are used to great length within many scientific domains. Though each domain has defined this term differently, the definitions in the context of machine learning shall be used. In this context, a model is a collection of vector transformations which may be performed on any input vector x to produce a prediction y'

Definition: A model is an approximation of a desired function f which produces relevant results based on human definitions.

Mathematically a model may be represented as a collection of numbers M which may in turn be used to compute f for any given example. In the context of machine learning a set of parameters may be tuned during a learning process (or training process). These parameters are combined with samples of data through some mathematical procedure to effectively model a distribution from which the data was extracted. The equation below is an example of a n-dimensional object.

$$x_0 \theta_0 + x_1 \theta_1 + ... + x_n \theta_n$$

3.2.1 K-means Clustering

Clustering is an unsupervised learning method whose primary use is in grouping sets of data. In this thesis we consider a fundamental version of such a clustering algorithm called *K-means clustering*. The following is the formal definition of *K - means clustering* as defined by McQueen

Given a set of N-dimensional points E_N and a desired amount of partitions k in said population, partition the elements of E_N into a partitioned set $S = \{S_1, S_2, ...S_k\}$. The partitioning of E_N is performed by initializing k N-dimensional points as randomly selected points within E_N. We define the set V with elements v where v_i is the i:th cluster center where $i \in [0, k]$. The partitioning of the elements $x \in E_N$ into their

respective partition S_i is performed by computing the closest cluster center $\forall_{x \in E_N}$. Let T_i be the set of $x \in E_N$ such that the distance from the element to the relevant cluster is minimal, T_i is defined by formula 3.6.

$$T_i = \{x : x \in E_N | (\forall_{j \in [0,k]/i} |x - v_i| \leq |x - v_j|)\} \tag{3.6}$$

For centers who share equal distance to any give x the cluster with the smallest index is chosen as the containing set. The partitions $S_i \in S$ are defined by formula 3.7

$$S_i = T_i \cap \bigcap_{j=0}^{(i-1)} S_j^c \tag{3.7}$$

3.3 Deep Learning

The field of Artificial Intelligence is founded on the notion of designing algorithms for solving problems. The field encountered tremendous progress in **[FIND YEAR, AI FOUNDATIONS]** referred to by **[NAME]** as the "look ma, no hands" era of Artificial Intelligence. One such method which have proven useful for these tasks is the practice of approximating models through Artificial Neural Networks.

3.3.1 Artificial Neural Networks

Artificial Neural Networks ("ANNs") have been used to great success during the 20th century [**Source Here**]. With the use of ANNs several fields including Natural Language Processing, Encoding and Image classification have undergone revolutionary leaps in performance regarding optimization due to the predictive power of these networks [**Source Here**]. At the same time they are heavily criticized for their complexity, yielding a structure much more akin to a so called *"black box"* than a reliable and deterministic method for prediction[**Source Here**]. This complexity is due to numerous different structural typologies available at present and an awesome number of tuned parameters which are modified with the goal of minimizing the predictive error [**Source Here**].

A consequence of this is hard skepticism in regards to the correctness of their functionality within practical use. While these models have shown great promise when compared to their human counterparts, the question remain whether or not perfect performance can be yielded from the constructed models.

Definition 1. Training an ANN is allowing minuscule changes through the ran-

domly initialized structure in order to approximate a collection of nested functions

$$f_n(f_{n-1}(...f_1(X)))$$

3.3.2 Computational Representations

The initial purpose of ANNs was to create a computational model of the human cortex which took the form of the McCulloch, Pitts neuron. The multilayer perceptron (MLP) introduced in **[year here]** formed the basic structure which would become ANNs.

Chapter 4

Data Exploration

In this section the data is presented in greater detail. Due to the low number of samples available, it is necessary to examine each sample in detail to determine it's predictability. The data must be satisfyingly diverse between the given classes and similar within those classes for the predictive model to work appropriately. Should this not be the case, the model will struggle to reach desired performance by either failing to capture basic features of the data or by overfitting to it.

The data of the glioblastoma patients is provided by **PROVIDERS HERE** and consists of the Raman-spectrum of 45 tumors from separate patients. Multiple samples of tissue was extracted from the same tumor in some cases, yielding several samples for the respective tumors. The samples are sorted by the patient to whom they belong to maintain separation between the patient samples. Moreover, there is large variation with regard to the sample shape within the data. Each sample is a 3-dimensional array of size $(w, h, 1738)$ where w and h are the respective width and height of the sample. This formalization is necessary, as width and height have non-zero variance among different samples. the number 1738 is constant through all samples and represent the length of the Raman-spectrum.

Chapter 5

Results

Chapter 6

Conclusion

[1]

Bibliography

[1] A. Einstein, "Zur Elektrodynamik bewegter Körper. (German) [On the electro-dynamics of moving bodies]," *Annalen der Physik*, vol. 322, no. 10, pp. 891–921, 1905.