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 MacQueen[**?**]. 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

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

3.2.2 Hierarchical clustering

Hierarchical clustering is a clustering method which is less suceptible to outliers compared to K-means. The method produces clusters in hierarchies by separating the data into clusters from which the process continues recursively. The two primary strategies for forming clusters are agglomerative and divisive. Agglomerative clustering initializes one cluster for each data point and combines them in a hierarchy according to the linkage criterion until all clusters are part of the hierarchy. Divisive strategies process in counter to the agglomerative strategies by initializing one universal cluster for all data points and divide the points into separate clusters according to the linkage criterion. The method proceeds until all data points are separated to their own cluster within the unifying hierarchy. The project described in this thesis uses the agglomerative strategy. All strategies depend on the distance measure and linkage criterion. A usual choise for the distance metric is the euclidian distance, calculated by equation 3.8.

$$d(u, v) = \sqrt{\Sigma_i (u_i - v_i)^2} \tag{3.8}$$

The euclidean measure is use in the project. The linkage criteria used in this project is the complete-linkage method (maximum linkage). The linkage criterion determines where clusters will be merged in the clustering method.

EXPAND THIS SECTION, ADD SKLEARN REFERENCES! [1]

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 randomly 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 sufficiently 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. Within this project it is also vital to examine and determine if the samples of one class are different with respect to samples of other classes i.e. The sample-classes are heterogeneous. This criterion is important, since there must be a sufficient difference between the different classes to separate them appropriately. This also requires that the within-class samples are sufficiently similar i.e. All tumors belonging to the same class are homogeneous. Should this criterion not be satisfied the model is expected to overfit to the training data and fail to generalize to the test data. It is also necessary to prepare for eventual worst case scenarios. A possible worst case scenario is that all tumors are heterogeneous, which would mean that any model would be unable to satisfyingly distinguish between the classes. This means that we must evaluate the model on a patient by patient basis as the model would learn to recognize patterns in individual patients. This becomes relevant when selecting which samples to use for training the model and testing it.

4.1 Data Representation

The data of the glioma 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; this 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 width and height of the sample respectively. 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. Each element inside these arrays is a real number without clear bounds. The largest absolute element found within is 79427.0625, some values are negative which is confirmed by the providers to have significance for the projects purpose. The project aims to predict the methylation-types by feeding in one of these samples i.e. one vector of shape $(1, 1738)$

4.2 Data preparation

The preparation of the data is vital to form non-biased models, provided that the data can be restructured and rescaled to sufficiently represent the majority of samples. Should this be impossible the resulting predictive model will fail to grasp the necessary features in the data for forming predictions. The model would in this respect overfit to the data used to train it. To avoid this we preform a qualitative analysis on the data to determine the predictability. Each sample i cathegorized according to their methylation-types, there are six distinct classes of methylation in this data, the labels are denoted by: LGm1, LGm2, LGm3, LGm4, LGm5 and LGm 6.

4.2.1 Balancing

The data used to train a model must in some way be balanced. Should one label-class be over-represented with respect to the others, the model will as a consequence of it's learning-algorithm become biased towards certain predictions **PROVE THIS!**. The data suffers heavily from this problem, a table of the number of spectra in each class may be seen in Table 4.1.

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# of samples	5	16	7	13	15	5
# of spectra	37319	210586	39636	50660	62256	20176
percentage	9%	50%	9%	12%	15%	5%

Table 4.1: Table showing the distribution of data in the initial dataset

Table 4.1 shows the per class separation in the data, the first row shows the labels of each class. The second row shows the number of samples belonging to each class, these are the tumors which will be analysed. The third row display the total number of spectra across each class, these must be considered when balancing is performed. The last row of table 4.1 shows the percentage each class makes of the entire dataset. Initially LGm2 is the majority class while LGm 6 is the minority, consisting of only 5% of the entire dataset. Some samples prove to be problematic as their size make them too big to load into memory. Some even included what appears to be erroneous readings. For simplicity they are removed from the analysis. Their plots are shown in appendix **FILL IN APPENDIX!**6. The data is now represented in Table 4.2

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# of samples	5	11	4	13	15	5
# of spectra	37319	71846	14896	50660	62256	20176
percentage	15%	28%	6%	20%	24%	8%

Table 4.2: The data distribution after removing problematic samples

Before the data is balanced, the testing data is selected and separated from the rest. This is done by separating at least one patient and all their samples from the rest of the data. This way it will be possible to test if the model is overfit to the patients in training and if the patient samples are heterogeneous with respect to the other samples of the same class. With the test-patients separated, balancing the classes which contain less elements by a factor larger than or equal to two than the majority class (LGm2) is done by repeating the spectrum in each sample by that factor. Following this method the majority class will stay the majority which is be crucial provided the sample pattern is correct. The resulting dataset is gained by doubling the samples in LGm1, quadrupling the samples in Lgm3 and tripling the samples in LGm6. The distributions of the final training dataset is shown in Table 4.3.

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# of spectra	22374	51698	11296	34276	43892	12976
Factor	2	1	4	1	1	3
# of spectra	44748	51698	45184	34276	43892	38928
percentage	17%	20%	17%	13%	17%	15%

Table 4.3: Distribution of the training data

Table 4.3 shows the number of spectra before multiplying on row two and the

number of spectra after on row four. The precentages on row five shows the distribution on the modified dataset. The percentages are now closer in scope, meaning each class is now in relative balance to the other classes.

4.2.2 Clustering

To ensure each sample include relevant information requires intuitive analysis. During extraction it is possible droplets of certain fluids were present, it is also possible the laser hit the tissue at a thin point which would scan the plastic underneath. This data must be removed to avoid model dependency on erroneous spectra. To find and remove the erroneous spectra, a clustering method is used. A suggested method is the k-means clustering algorithm by McQueen **SOURCE HERE!** performed on every spectra on a sample by sample analysis. The resulting clustering may be displayed as a 2-dimensional image, some of which are shown in **APPENDIX!!**. K-means do however suffer heavily from outlier influences, as these outliers may be present at this stage, it is necessary to consider other methods. Hierarchical clustering is a clustering method which avoids outlier dependency by separating each individual spectra into its own cluster and then reducing the number of clusters depending on the linkage type. Complete linkage is used to best capture the features,**Fill in information on how the method works, from sklearn**

4.2.3 Feature selection

Each number in the spectra is a frequency at which the scattered light is gathered. This light is expected to be sufficient for predicting the methylation-type of the tumor-tissue. Due to the number of spectra inside each sample memory quickly becomes an issue, it is therefore relevant to examine which frequencies would have the most impact on classification. For this reason the best features are extracted with SelectKBestfeatures **SOURCE FROM SKLEARN**. The 70 best features were extracted from each spectra which is shown to be appropriate to recreate the original tumorshape by performing clustering

Chapter 5

Results

Chapter 6

Conclusion

[2]

Bibliography

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

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

Appendix

Appendix A

INSERT TABLE

Appendix B

INSERT TABLE