2019-01-01

Earthquake Magnitude Prediction Using Support Vector Machine and Convolutional Neural Network

Esther Amfo
University of Texas at El Paso, amfoesther3@gmail.com

EARTHQUAKE MAGNITUDE PREDICTION USING SUPPORT VECTOR MACHINE AND CONVOLUTIONAL NEURAL NETWORK

ESTHER KESEWA AMFO

Master's Program in Mathematical Science

APPROVED:

Maria C. Mariani, Ph.D.,Chair

Joe Guthrie, Ph.D

Thompson Sarkodie-Gyan, Ph.D

Stephen Crites, Ph.D.
Dean of the Graduate School

EARTHQUAKE MAGNITUDE PREDICTION USING SUPPORT VECTOR MACHINE AND CONVOLUTIONAL NEURAL NETWORK

by

ESTHER KESEWA AMFO

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2019

Acknowledgements

I would like to express my deep-felt gratitude to my advisor, Dr. M . Mariani of the Mathematics Department at the University of Texas at El Paso, for her advice, encouragement, enduring patience and constant support. She was never ceasing in her belief in me I also wish to thank the other members of my committee, Dr. Joe Guthrie of the Mathematics Department and Dr. Sarkodie-Gyan Thompson of the Engineering Department, both at The University of Texas at El Paso. Their suggestions, comments and additional guidance were invaluable to the completion of this work.

Abstract

A deep learning-based method Convolutional Neural Network,(CNN) and Suport Vector Machine(SVM) for earthquake prediction is proposed. Large-magnitude earthquakes triggered by earthquakes can kill thousands of people and cause millions of dollars worth of economic losses. The accurate prediction of large-magnitude earthquakes is a worldwide problem.

In recent years, deep learning technology that can automatically extract features from mass data has been applied in image recognition, natural language processing, object recognition, etc., with great success. We apply deep learning technology to earthquake prediction ,we propose a deep learning method for continuous earthquake prediction using historical seismic events.

In this study we apply the CNN and SVM algorithm to predict earthquake data. The modeling is a machine-learning-based method. It involves a training phase with associated input and a predicting phase with target output decision values. In recent years, these two method has become increasingly popular for prediction of earthquake magnitudes. Taking New Mexico as an example, we train our deep leaning network model, using the images of the dataset.

Finally, we make earthquake predictions, using the trained network model. The result shows that we can get the best result, when we predict earthquakes. The proposed method performs well without manually designing feature vectors, as in the traditional neural network method. This method can be applied to earthquake prediction in other seismic zones.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Tectonic earthquakes occur anywhere in the earth where there is sufficient stored elastic strain energy to drive fracture propagation along a fault plane. The sides of a fault move past each other smoothly and a seismically only if there are no irregularities or asperities along the fault surface that increase the frictional resistance. Most fault surfaces do have such asperities and this leads to a form of stick-slip behavior. Once the fault has locked, continued relative motion between the plates leads to increasing stress and therefore, stored strain energy in the volume around the fault surface. This continues until the stress has risen sufficiently to break through the asperity, suddenly allowing sliding over the locked portion of the fault, releasing the stored energy. This energy is released as a combination of radiated elastic strain seismic waves, frictional heating of the fault surface, and cracking of the rock, thus causing an earthquake.

An earthquake is a highly destructive natural disaster. If earthquakes can be accurately predicted, many lives can be saved and economic loss can be avoided. Mexico is a country with high seismic activity currently. A famous instrument for earthquake monitoring in ancient times was Di Dong Yi in the Han Dynasty. Earthquakes are vibrations of Earth caused by large releases of energy that accompany volcanic eruptions, explosions, and movements of Earth's crust along fault lines. The earthquake vibrations are waves of energy that radiate through Earth away from the focus. These waves of energy can be recorded on a seismograph, which produces a recording called a seismogram. Seismographs record the Primary waves (P-waves) and Secondary waves (S-waves). They also detect Surface waves called Love waves (L-waves) and Rayleigh waves (R-waves). Travel-time curves are graphs that indicate how long it takes each type of seismic wave to travel a distance measured

on Earth's surface. The difference between the S wave arrival time and the P-wave arrival time corresponds to the distance of the seismograph station from the earthquake focus. This time difference can be converted easily into distance using the travel-time curves

1.1 Background of the Study

1.1.1 Puebla Region

Mexico is one of the world's most seismically active regions, sitting atop several intersecting tectonic plates. The border between the Cocos Plate and North American Plate, along the Pacific Coast of Mexico, creates a subduction zone that generates large seismic events.Activity along the edges of the Rivera and Caribbean plates also generate seismic events. All together, these seismic forces cause an average of 40 earthquakes a day in Mexico. Mexico City is built on a dry lakebed with soft soil made up of sand and clay, which amplifies the destruction that major earthquakes cause. Loose sediments near the surface slow the shockwaves' speed from 1.5 miles (2,414m) per second (8690.4 km/h) to roughly 150 feet (45,72m) per second (164.592 km/h).This increases the shockwaves' amplitude, which causes more violent shaking. Deeper and denser soil layers increase amplified shockwaves' destructive duration. Less than two weeks before the Puebla earthquake, Mexico had been struck by an earthquake in Chiapas on 7 September, which killed almost 100 people. Despite its close timing, the Puebla earthquake was not an aftershock of the Chiapas event, as the epicenters were 650 km (400 mi) apart. The possibility of a link between the earthquakes was being investigated in the days after the second one. Big earthquakes can increase the long-term risk of seismic activity by transferring "static stress" to adjacent faults, but only at a distance of up to four times the length of the original rupture. In 19 September earthquake, static stress transfer was considered unlikely due to the distance between the earthquakes, in excess of the expected 400 km maximum. "Dynamic triggering", with seismic waves propagating from one quake affecting other faults, may operate

at much longer distances, but usually happens within hours or a few days of the triggering quake; a 12-day gap is hard to explain. 19 September is designated as a day of remembrance for the 1985 Mexico City earthquake, which killed approximately 10,000 civilians. Every year at 11 a.m., a national earthquake drill is conducted by the government through the use of public loudspeakers located throughout Mexico City. The 2017 drill took place as scheduled, at 11 a.m., around two hours before the central Mexico earthquake.

1.1.2 Oaxaca Region

Oaxaca lies on the destructive plate boundary where the Cocos Plate is being subdued beneath the North American Plate. In the region of this earthquake, the Cocos Plate moves approximately northeastward at a rate of 60 mm/yr. Historically, several significant earthquakes have occurred along the southern coast of Mexico. In 1932, a M 8.4 mega thrust earthquake struck in the region of Jalisco, several hundred kilometers to the northwest of the Oaxaca event. On October 9, 1995, a M 8.0 earthquake struck in the Colima-Jalisco region, resulting in at least 49 fatalities and leaving 1,000 people homeless. The deadliest nearby earthquake occurred on September 19, 1985, in the Michiogan region 500 km to the northwest of the February 16th event. This M 8.0 earthquake resulted in at least 9,500 fatalities, injured about 30,000 people, and left 100,000 people homeless. In 2003, a M 7.6 earthquake in Colima, Mexico, resulted in 29 fatalities, destroyed more than 2,000 homes and left more than 10,000 people homeless. In March 2012, a M 7.4 earthquake 60 km to the northwest of the February 16, 2018 event killed 2 and injured 11 in the Oaxaca region. The hypo center of the M 8.2 earthquake off the shore of Chiapas in September 2017 was located 440 km southwest of this earthquake. The Chiapas event caused at least 78 fatalities and 250 injuries in Oaxaca, and a further 16 deaths in Chiapas. Eleven days later, a M 7.1 earthquake struck closer to Mexico city, 230 km northeast of today's earthquake, resulting in over300 fatalities and significant damage in Mexico city and the surrounding region.

A second quake, registering 5.9, struck Oaxaca on February 19, 2018, around 12:57 AM local time. The quake, believed to be an aftershock, had an epicenter about 69 miles

southwest of Oaxaca City; its impact was registered in Mexico City. No deaths were reported from this quake.

Figure 1.1: Distance from Puebla to Oaxaca

1.1.3 Problem Statement

Earthquakes are becoming very prevalent in Mexico. For the Puebla and Oaxaca regions which happens to be a little closer experienced powerful earthquakes and a building only competes a few moths ago would have been built using modern construction techniques designed to sustain even more severe seismic shocks but it also collapsed even though it has been for a year. "How is it possible" but some buildings next to had no damages ,no scars,no nothing."How can it pass"?, the reason is still not solved.

Following another tremor in Mexico on the same day in 1985, the government issued tough regulations to help make new buildings quake-proof, including making the foundations with a stronger mix of cement, and making stronger walls and column. But there are suspicions some unscrupulous builders may have ducked on these rules with the help of

corrupt officials.

The earthquake on Tuesday 2012 damaged over 3,000 buildings in Mexico City, the government has said; many have only superficial damage but dozens collapsed completely, and many more have structural cracks and need to be demolished. Thousands of families are staying with friends or in shelters, scared to go back in case their homes tumble down on them in the night. Most of the buildings that fell appear to have been built before the 1985 earthquake and the new rules. But others were recent.

Puebla, church steeples had toppled in the city of Cholula, and a church on the slopes of PopocatÃ©petl in Atzitzihuacan collapsed during mass, killing 15 people. A second church, which was built in the 17th century, fell in Atzala during a baptism, killing 11 people including the baby. At least 44 buildings collapsed in Mexico City due to the earthquake, trapping people inside, creating large plumes of dust, and starting fires.At least 50 to 60 people were rescued by emergency workers and citizens.Several buildings caught fire.

1.1.4 Thesis Organization

This report is organized in five chapters. Chapter 1 is the introductory section to the study. It takes a general look at the the background of earthquake as well as the study area, problem statement Chapter 2 reviews related Literature reviews based on the objective of the study. Chapter three describes the theory to be used. Chapter 4 is dedicated to the data collection, analysis and results. Chapter 5 concludes the entire study on the major findings.

Chapter 2

Literature Review

This chapter describes the literature available on Earthquake in general. Earthquakes have been known to man since ancient times. They represent one of natural hazards human society has to face, often without any kind of warning. It involves a severe, shaking of the earth below our feet affecting all systems and structures standing on it. Generally, it lasts for a fraction of a minute but will often causes great loss of life and property.

Structural engineers usually consider two aspects of earthquake engineering in every seismic design. These can be described as demand and capacity. The demand is the level of seismic loading that might be applied to the building while the capacity is the resistance of the building to resist the demand. The conventional seismic design attempts to make the buildings that do not collapse under strong earthquake shaking, but may sustain damage to non-structural elements and to some structural members in the building. This may render the building non-functional after the earthquake, which may be problematic in some structures like hospitals, which need to remain functional in the aftermath of the earthquake. The need to minimize earthquake damage is critical and important.

The base isolation works by decoupling the building or structure from the horizontal components of the earthquake ground motion by interposing a layer with low horizontal stiffness between the structure and the foundation.

In 1909, a medical Doctor Calantarients in England applied for a British patent on an earthquake-resistant design approach. Frank(1921) was the first person to implement the idea of base isolation, He applied the base isolation idea to the foundation design for the

Imperial Hotel in Tokyo in 1921, under the site was an eight feet layer of fairly good soil and below that a layer of soft mud. Accordingly, the idea of floating the building came into the picture for the resistance of earthquake shock.

The flexible first-storey concept was first proposed by Martel (1929) and further studied by Green (1935), Jacobsen (1938). In this approach the lateral stiffness of the columns of the first-storey would be designed to be much lower than that of the columns above, and under earthquake loading the deformations would be concentrated in these first-storey columns.

In the search for a mechanism that can overcome the difficulty of a flexible first-storey, Ryuiti (1941, 1951, 1952) and Caspe (1970, 1984) proposed many types of roller bearing system and several have been patented and tested. However, as the earthquake movement can be in any direction, these types of roller bearing system did not become popular. As a result, it made necessary to use spherical bearings or two crossed layers of rollers. Lee and Medland (1979) examined the effectiveness in respect of EI Centro earthquake excitation of a multi-storey shear type structure isolated by the lead rubber bearing. Tadjbakhsh and Ma (1982) and Pan and Kelly (1983, 1984) studied the seismic response of base isolated buildings by modeling the superstructure as a rigid block supported on an isolation system. The hysteretic force deformation behavior of the lead rubber bearing is modeled as bilinear. Tadjbakhsh (1983, 1985, 1985a) studied the response of a shear type building supported on the laminated rubber bearing system under random ground motion.

Kelly and Tsai (1985) studied the seismic response of light internal equipment in base isolated multi degree shear type structures. They had shown that the use of base isolation can not only attenuate the response of the primary structural system but also reduce the response of the secondary systems. Mostaghel and Khodaverdian (1987) proposed the resilient-friction base isolation (R-FBI) system. Paul and Novak (1989) studied the response of base isolated building to wind loading by modeling the superstructure as a rigid block supported on an isolation system. Tasi and Kelly (1989) demonstrated effect of

superstructure flexibility using a discrete multiple degrees of freedom system having only horizontal degree of freedom at each floor.

Ghobarah and Ali (1989) proposed a simple design procedure for highway bridges, which aims at optimum balance between the shear forces transmitted to the supports and tolerable deck displacements for isolated highway bridges using the inelastic response spectra approach. Simplified charts are presented which provide a design aid for new bridges as well as the retrofitting and upgrading of existing ones. The method is shown to be simple and reasonably accurate. It takes into account the flexibility of the pier and is suitable for a code-type approach.

Briseghella et al. (1989) presented a design approach for applying base isolation technologies to typical medium-span continuous concrete deck bridges. A method for constructing non-linear response spectra for rigid-plastic systems is explained in which a direct strength-displacement relationship is obtained without depending on the elastic period.

Kelly (1990) illustrated through a linearised theory of base isolation the effect of superstructure flexibility using two degrees of freedom system.Fan and Ahmadi (1990) observed that use of base-isolation system eliminates the resonance peak of the floor spectra, which occurs at the natural frequency of the fixed base system for earthquake' ground excitation. In their study they had modeled the superstructure as a shear type building.

Koh and Kelly (1990) presented a fraction Kelvin model to define the force-deformation relation of elastomeric bearings: An efficient numerical integration scheme is presented for the solution of the equation of motion for a base isolated system. The numerical examples reveal a good performance of the algorithm developed. In addition, a shaking table test indicated that the fractional derivative model agrees well with the experimental model.

Su et al. (1991) proposed the design of the sliding resilient-friction (S-RF) base isolator. This isolator combines the desirable features of the EDF and the R-FBI systems. It was suggested to replace the elastomeric bearings of the EDF base isolation by the R-FBI units.

Constantinou et al. (1991) proposed an isolation system consisting of multi-directional sliding Teflon bearings and displacement control devices. The displacement control devices provide re-centering capability and displacement control during earthquakes and rigidity under service loads.

Mayes et al. (1992) presented an overview of the basic concepts and design principles of seismic isolation and discussed the objectives and philosophy of the provisions of American Association of State Highway and Transport Officials (AASHTO, 1991) and concluded with a procedure to compare the performance of isolation systems with different damping values.

Fan and Ahmadi (1992) studied the seismic response of secondary systems in base isolated shear type structures. They had shown that the use of base isolation provides considerable protection for structural contents. Gueirreiro and Azevedo (1992) studied the ductility demand of base-isolated structure with non-linear behavior. They considered superstructure as elasto-plastic and isolation system as bilinear. They constructed design diagrams for behavior coefficients to be used for a given structure and base isolation .

2.0.1 Health Impact

The human impact of earthquakes includes mainly mortality and injury. Earthquakes cause a considerably high number of primary deaths.The injury to mortality ratio of most earthquakes has been observed to be about 3 injured to 1 death. This makes the injuries from earthquakes a special concern.

Mortality

Earthquakes cause high impact deaths. These are primarily attributed to the serious injuries sustained and difficulties of rescue or immediate relief. This is also the reason why mortality once rescued is low. However, if the serious injured are rescued immediately the primary mortality may decrease but secondary mortality due to delayed deaths may increase. Thus this has to be interpreted in view of effectiveness of the rescue and immediate relief. All in all most report that secondary deaths were far lower in frequency.

Injury and Infection

Injury and infection Falling debris and entrapment pose the greatest risks for injuries (morbidity). A large number of injuries also occur due to being trapped in between objects or being hit by furniture. (PeekAsa et al. 1998) Entrapment after an earthquake poses serious risks including the lack of oxygen, asphyxia, body compression, hypothermia, smoke, and water penetration. (Redmond 2005) A detail understanding of the mechanisms of injury and the building components causing injuries can help direct preventive efforts (Shoaf et al. 1998).

Injury Severity

Injury severity Most injuries sustained after the earthquake were simple contusions, lacerations and cuts. These are usually treated on an outpatient basis and seldom documented. Over 50% of the studies reporting injury morbidity reported data on single hospital in-patients. Majority of them found in-patient data more complete than out patient data.

Often injury reporting was among the survivors and no concrete study reporting injuries post-mortem. This may be viewed as a bias. Those sustaining severe injuries particularly serious injuries for chest, head, and abdomen die either due to delayed rescue or due to inadequate first aid and response. Second most important is that this was the injury distribution among the survivors. One study that did not complete the autopsies on all deceased but observed that commonest injuries seen in those dead was head (48.5%), thoracic (42.4%) followed by abdominal and lower extremity injuries (Peek-Asa et al. 1998). Another study reported similar findings and stated that although the total percentage of head, abdominal and thoracic injuries constituted less than 7.5% of the total injured, the mortality in this group was the highest (Tanaka et al. 1999).

Crush Injury and Crush Syndrome

Crush injury and crush syndrome Most serious injuries from earthquakes come from being trapped under or between heavy objects."Crush Syndrome is the systemic manifestation of rhabdomyolysis caused by prolonged continuous pressure on muscle tissue. It is characterized by hypovolaemic shock, hyperglycemia, acute renal failure and muscle necrosis. The first cases of Crush Syndrome were reported during the Sicilian earthquake in Messina in 1909.(Donmez et al.2001)

Crush syndrome is clearly related to the building design (Tanaka et al. 1999). It is more common in concrete multi-story building collapse. This was seen in the earthquakes in Japan and Armenia. In the study of morbidity after the Hanshin-Awaji Earthquake of 1995, and Armenian earthquake of 1988 reported higher number of Crush syndrome cases

(Tanaka et al. 1999). Far lower numbers were reported in Nicaragua, Guatemala and Iran due to the mud and adobe construction or probably due to a poor rescue operation. Thus, the number of patients developing crush syndrome seems to also be related to the rescue time.

The early clinical signs of crush syndrome if recognized and treated urgently with appropriate medication, fluid resuscitation and dialysis can help prevent death. The incidence of acute renal failure was reduced mainly by administration of intravenous fluids at rescue (Dhar et al. 2007). Unfortunately first responders, rescue workers, paramedics and even untrained nephrologists are unfamiliar with recognizing early symptoms. This is further compounded by the lack of infrastructure for dialysis to cope with the mass casualties (Vanholder et al. 2007). This problem however arises only when the search and rescue level is effective enough to retrieve these patients alive in large numbers. This was demonstrated in the Gujarat earthquake where less than two percent cases of crush syndrome were reported a documented case of poor rescue performance (Cooper 2006).

Crush injury (CI) occurs when a body part is subjected to a high degree of force or pressure that leads to bleeding, bruising, increased pressure in the compartment, fracture and lacerations (Medical Encyclopedia 2007). Crush injury and crush syndrome both are more common in children with a higher risk to develop acute renal failure (DÃ¶nmez et al. 2001;Iskit et al. 2001). This is further compounded by the difficulty in diagnosis both clinically and diagnostically (Iskit et al. 2001). Close monitoring of children is therefore solicited to detect early signs. Deliverables Annex 2 - D1.1.2 HWG Literature Review 16.02.2009 11 The other aspect with crush injuries is secondary complications due to heightened risk of secondary infections, gangrene and amputations of the affected limb(s) thereof (Personal observation during the work in Gujarat).

Chapter 3

Support Vector Machine

The support vector machine (SVM) is a supervised learning method that generates input-output mapping functions from a set of labeled training data. Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory.

Support vector machine was initially popular with the NIPS community and now is an active part of the machine learning research around the world. SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task . It is also used for hand writing analysis, face analysis and so forth, especially for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by Vapnik and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SRM minimizes an upper bound on the expected risk, where as ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were developed to solve the

classification problem, but recently they have been extended to solve regression problems

3.0.1 Statistical Learning Theory

The statistical learning theory provides a framework for studying the problem of gaining knowledge, making predictions, making decisions from a set of data. In simple terms, it enables the choosing of the hyper plane space such a way that it closely represents the underlying function in the target space .

In statistical learning theory the problem of supervised learning is formulated as follows. We are given a set of training data $(x_1, y_1) \ldots (x_1, y_1)$ $\Re^n$ x $\Re$ sampled according to unknown probability distribution P(x,y), and a loss function $V(y, f(x))$ that measures the error, for a given x, $f(x)$ is "predicted" instead of the actual value y. The problem consists in finding a function f that minimizes the expectation of the error on new data that is, finding a function f that minimizes the expected error. In statistical modeling we would choose a model from the hypothesis space, which is closes to the underlying function in the target space. More on statistical learning theory can be found on introduction to statistical learning theory.

3.0.2 Mathematical Appraoch

$$Y_i = +1; wx_i + b \geq 1 \tag{3.1}$$

$$Y_i = -1; wx_i + b \leq 1 \tag{3.2}$$

For all i,

$$Y_i = yi(wxi + b) \geq 1 \tag{3.3}$$

In this equation x is a vector point and w is weight and is also a vector. So to separate the data $(wx_i + b)$ should always be greater than zero. Among all possible hyper planes, SVM selects the one where the distance of hyper plane is as large as possible. If the training

data is good, every test vector is located in radius r from training vector. Now if the chosen hyper plane is located at the farthest possible from the data this desired hyper plane which maximizes the margin also bisects the lines between closest points on convex hull of the two datasets. The above three equations are represented as given in Figure 3.1

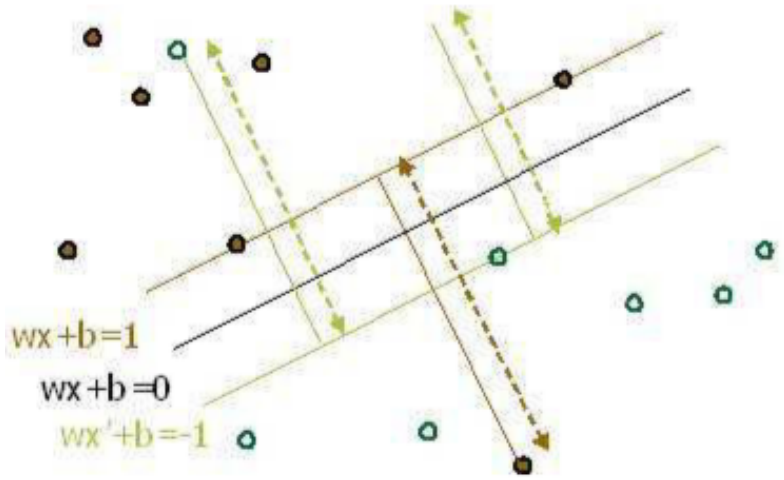


Figure 3.1: Representation of hyper planes

Distance of closest point on hyperplane to origin can be found by maximizing the x as x is on the hyper plane. Similarly for the other side points we have a similar scenario. Thus solving and subtracting the two distances we get the summed distance from the separating hyper plane to nearest points.

Maximum Margin =

$$M = 2/||w|| \tag{3.4}$$

Now maximizing the margin is same as minimum. Now we have a quadratic optimization problem and we need to solve for w and b. To solve this, we need to optimize the quadratic function with linear constraints. The solution involves constructing a dual problem and where a Langlier's multiplier αi is associated. We need to find w and b such that

$$f(w) = \frac{1}{2}|w'||w| \tag{3.5}$$

is minimized;

for all

$$(x_i, y_i) : y_i(w * xi + b) \geq 1 \tag{3.6}$$

Then we get that

$$w = \sum \alpha i * xi; b = y_k - w * x_k \tag{3.7}$$

for any x_k such that $\alpha k \neq 0$ Now the classifying function will have the following form:

$$f(x) = \sum \alpha i y_i x_i * x + b \tag{3.8}$$

3.0.3 SVM Representation

In this we present the QP formulation for SVM classification. This is a simple representation only. SVM classification

$$\min ||f||_k^2 + c \sum_{i-1}^{1} i \tag{3.9}$$

$y_i f(x_i) \geq 1 - \zeta_i, \qquad$ for all i$\zeta_i \geq 0$

3.0.4 SVM classification, Dual formulation

$$\min_{\alpha_i} \sum_{i-1}^{i} \alpha_i - \frac{1}{2} \sum_{i-1}^{1} \sum_{i-1}^{1} \alpha_i \alpha_j y_i y_j K(x_i, x_j) 0 \qquad 0 \leq \alpha_i \leq C \tag{3.10}$$

for all i

$$\sum_{i-1}^{1} \alpha_i y_i = 0 \tag{3.11}$$

Variables ζ_i are called slack variables and they measure the error made at point (x_i, y_i). Training SVM becomes quite challenging when the number of training points is large . A number of methods for fast SVM training have been proposed.

3.0.5 Soft Margin Classifier

In real world problem it is not likely to get an exactly separate line dividing the data within the space. And we might have a curved decision boundary. We might have a hyper plane which might exactly separate the data but this may not be desirable if the data has noise in it. It is better for the smooth boundary to ignore few data points than be curved or go in loops, around the outliers. This is handled in a different way; here we hear the term slack variables being introduced. Now we have,

$$y_i(w'x + b) \geq 1 - S_k \tag{3.12}$$

This allows a point to be a small distance S_k on the wrong side of the hyper plane without violating the constraint. Now we might end up having huge slack variables which

allow any line to separate the data, thus in such scenarios we have the Lagrangian variable introduced which penalizes the large slacks.

$$\min L = \frac{1}{2}w'w - \sum \alpha_i(y_k(w'x_k + b) + s_k - 1) + \alpha \sum s_k \tag{3.13}$$

where reducing g allows more data to lie on the wrong side of hyper plane and would be treated as outliers which give smoother decision boundary.

3.0.6 Learning and Generalization

Early machine learning algorithms aimed to learn representations of simple functions. Hence, the goal of learning was to output a hypothesis that performed the correct classification of the training data and early learning algorithms were designed to find such an accurate fit to the data . The ability of a hypothesis to correctly classify data not in the training set is known as its generalization. SVM performs better in term of not over generalization when the neural networks might end up over generalizing easily. Another thing to observe is to find where to make the best trade-off in trading complexity with the number of epochs; the illustration brings to light more information about this.

Figure 3.2: Number of Epochs vs Complexity

SVM modeling is a method based on nonlinear transformations of covariates into a

higher dimensional feature space (Vapnik, 1995) and not a strict theory based basis, but also a strong prediction capacity, which can better solve the practical problems of small samples, nonlinearity, higher dimension and local minimum point. Wang et al (2005, 2006) predicted strong earthquakes in Chinese mainland and studied the non-linear relations between the time series of strong seismicity in China using the Support vector machine which he obtained meaningful results. At present, SVM are widely used in text classification, handwriting recognition,image classification and in a lot of other classification problems , and now has been extended for the prediction of time series models.

There are two principal ideas that underlie SVM modeling for discriminant-type statistical problems. The first is an optimum linear separating hyperplane that separates data patterns. The second is the use of kernel functions to convert the l nonlinear data patterns into a linearly separable pattern in a high-dimensional feature space (Yao et al., 2008).The goal of SVM is to search an n-dimensional hyperplane differentiating the two classes by their maximum

3.0.7 Kernel Trick

Kernel: If data is linear, a separating hyper plane may be used to divide the data. However it is often the case that the data is far from linear and the datasets are inseparable. To allow for this kernels are used to non-linearly map the input data to a high-dimensional space. The new mapping is then linearly separable . A very simple illustration of this is shown below in figure 3.2.

Feature Space: Transforming the data into feature space makes it possible to define a similarity measure on the basis of the dot product. If the feature space is chosen carefully and pattern recognition can be easy as given in equation (4.14)

$$\langle x_1, x_2 \rangle \leftarrow k(x_1, x_2) = \langle (x_1) \cdot (x_2) \rangle \tag{3.14}$$

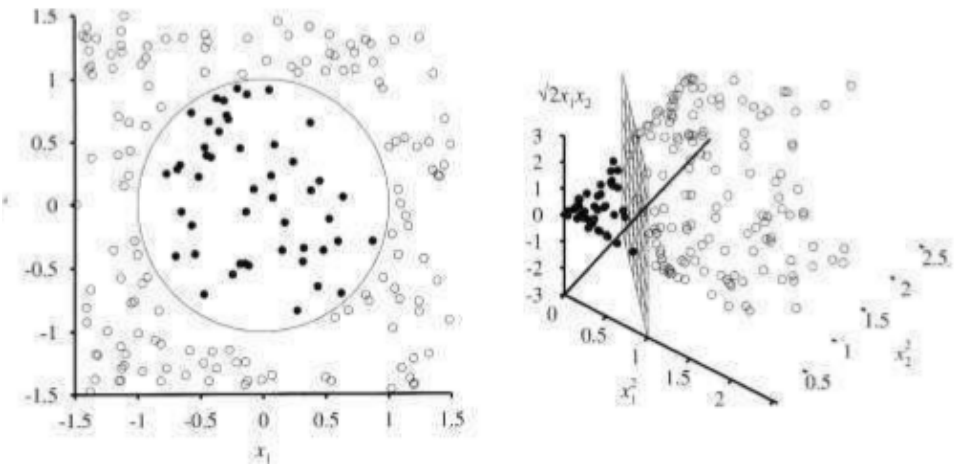


Figure 3.3: Feature Space representation

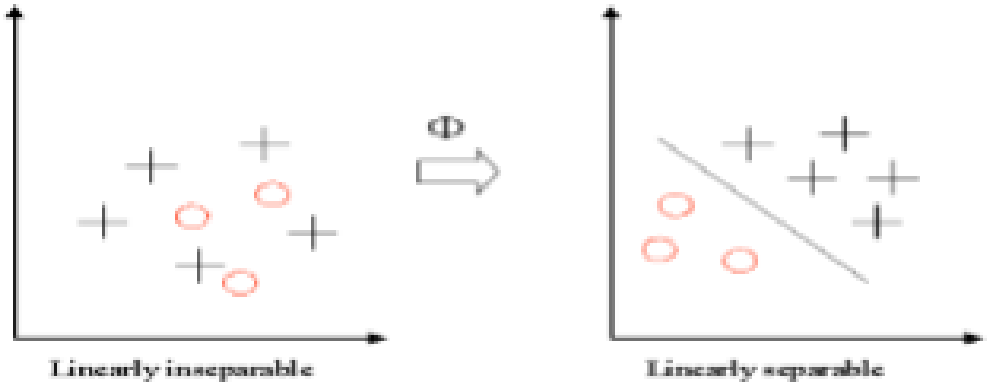


Figure 3.4: Kernel

3.0.8 Kernel Functions

The idea of the kernel function is to enable operations to be performed in the input space rather than the potentially high dimensional feature space. Hence the inner product does not need to be evaluated in the feature space. We want the function to perform mapping of the attributes of the input space to the feature space. The kernel function plays a critical role in SVM and its performance. It is based upon reproducing Kernel Hilbert Spaces.

$$K(x, x') = \langle(\phi(x), \phi(x')\rangle \tag{3.15}$$

If K is a symmetric positive definite function, which satisfies Mercer's Conditions,

$$K(x, x') = \sum_{m}^{\infty} a_m \phi_m(x)\phi_m(x') \qquad a_m \geq 0 \tag{3.16}$$

$$\int\int K(x, x')g(x)g(x')dx\dot{x}' \qquad g\epsilon L_2 \tag{3.17}$$

Then the kernel represents a legitimate inner product in feature space. The training set is not linearly separable in an input space. The training set is linearly separable in the feature space. This is called the "Kernel trick"

The different kernel functions are listed below. The below mentioned ones are extracted from there and just for mentioning purposes are listed below.

- Polynomial: A polynomial mapping is a popular method for non-linear modeling. The second kernel is usually preferable as it avoids problems with the hessian becoming Zero.

$$K(x, x') = (\langle x, x\rangle^d \tag{3.18}$$

$$K(x, x') = \langle x, x\rangle + 1)^d \tag{3.19}$$

- Gaussian Radial Basis Function: Radial basis functions most commonly with a Gaussian form

$$K(x, x') = \exp \frac{-||x - x'||^2}{2\sigma^2} \tag{3.20}$$

- Exponential Radial Basis Function: A radial basis function produces a piecewise

linear solution which can be attractive when discontinuities are acceptable.

$$K(x, x') = \exp \frac{-||x - x'||}{2\sigma^2} \tag{3.21}$$

- Multi-Layer Perceptron: The long established MLP, with a single hidden layer, also has a valid kernel representation.

$$K(x, x') = \tanh(\rho\langle x, x'\rangle + e) \tag{3.22}$$

In this paper, SVM is used for a synthetic earthquake prediction in Peubla and Oaxaca areas in Mexico seismicity parameters and observed precursory data.

3.1 Support Vector Machine - Regression

Suppose we are given training data $(x_1, y_1), \ldots, (x_n, y_n)$ X R, where X denotes the space of the input patterns (e.g. x = Rd). These might be, for instance, exchange rates for some currency measured at subsequent days together with corresponding econometric indicators. In ϵ-SV regression [Vapnik, 1995], our goal is to find a function f(x) that has at most ϵ deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible. In other words, we do not care about errors as long as they are less than ϵ, but will not accept any deviation larger than this.This may be important if you want to be sure not to lose more than ϵ money when dealing with exchange rates. The linear discriminant function is

$$f(x) = w \cdot x + b \tag{3.23}$$

where $\langle ., . \rangle$ denotes the dot product in X. Flatness in the case of (3.23) means that one seeks a small w. One way to ensure this is to minimize the norm,3 $||w||^2 = \langle w \rangle$ We can write this problem as a convex optimization problem

minimize $1 = 2||w||^2$

$$\text{subject to} = \begin{cases} y_i \langle w, x_i \rangle - b \leq \epsilon. \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases} \tag{3.24}$$

The tacit assumption in the equations (3.24) was that such a function f actually exists that approximates all pairs (x_i, y_i) with ϵ precision, or in other words, that the convex optimization problem is feasible. Sometimes, however, this may not be the case, or we also may want to allow for some errors. Analogously to the "soft margin" loss function [Bennett

and Mangasarian, 1992] which was adapted to SV machines by Cortes and Vapnik [1995], one can introduce slack variables ζ_i, $\zeta_i^\star$ to cope with otherwise infeasible constraints of the optimization problem (3.24) .

Hence we arrive at the formulation stated in [Vapnik, 1995]. Supposing that all the training data can be fitted as a linear function without error when ϵ is introduced then that is,

minimize $1 = 2||w||^2 + C\sum_{i=1}^{i}(\zeta_i + \zeta_i^\star)$

$$\text{subject to} = \begin{cases} y_i\langle w, x_i\rangle - b \leq \epsilon + \zeta_i. \\ \langle w, x_i\rangle + b - y_i \leq \epsilon + \zeta_i^*. \\ i \geq 0 \end{cases} \tag{3.25}$$

The constant $C < 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ϵ are tolerated. This corresponds to dealing with a so called ϵ insensitive loss function $|\eta|_\epsilon$ described by

$$|\zeta|_\epsilon = \begin{cases} 0 & \text{if } |\zeta| \leq \epsilon \\ \zeta - \epsilon & \text{otherwise} \end{cases} \tag{3.26}$$

3.1.1 Dual Problem and Quadratic Programms

The key idea is to construct a Lagrange function from the objective function (it will be called the primal objective function in the rest of this article) and the corresponding constraints,by introducing a dual set of variables. It can be shown that this function has a saddle point with respect to the primal and dual variables at the solution. For details see e.g. [Mangasarian, 1969, McCormick, 1983, Vanderbei, 1997]

$$y - (w \cdot x + b) \leq \epsilon \qquad \text{where i} = 1, 2, ..n \tag{3.27}$$

The goal of SVM is to search an n-dimensional hyperplane differentiating the two classes by their maximum gap , which can be expressed as minimizing

$$\frac{1}{2}w \cdot w + C\sum_{i=1}^{\ell}(\xi + \xi_i^\star) - \sum_{i=1}^{\ell}(\eta\zeta_i + \eta_i^\star\zeta_i^\star). \tag{3.28}$$

$$-\sum_{i=1}^{k}\alpha_i(\epsilon + \zeta_i - y_i + \langle w, x_i\rangle + b) \tag{3.29}$$

$$-\sum_{i=1}^{\ell}\alpha_i^\star(\epsilon\zeta_i^\star + y_i - \langle w, x_i\rangle - b) \tag{3.30}$$

where L is the lagrangian and $\zeta, \zeta^\star\alpha, \alpha^\star$ are Lagrange multipliers. Hence the dual variables in (3.28) are not less than zero. note that by $\alpha^(\star)_i$ we refer to α_i and $\alpha_i^\star$ it follows from the saddle point condition that the partial derivative of L with respect to the primal variables $(b, w, \zeta_i, \zeta_i^\star)$ have to vanish for optimality.

$$\delta_b L = \sum_{i=1}^{\ell}(\alpha_i^\star - \alpha_i) = 0 \tag{3.31}$$

$$\delta_w L = w - \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^\star)x_i = 0 \tag{3.32}$$

$$\delta_\zeta^\star L = C - \alpha^{(\star)} - \eta_i^{(\star)} = 0 \tag{3.33}$$

Substituting the above equations yields the dual optimization problem.

The dual function of the Lagrangian function is;

$$\text{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_j - \alpha_i{}^\star)(\alpha_j - \alpha_j^\star)(x_i \cdot x_j^\star). \\ -\epsilon\sum_{i=1}^{\ell} y_i(\alpha_i - \alpha_i{}^\star) + \sum_{i=1}^{\ell}(\alpha_j - \alpha_i^\star)\epsilon. \end{cases} \tag{3.34}$$

subject to $\sum_{i=1}^{k}(\alpha_i - \alpha_i{}^\star) = 0 \qquad \text{and}\alpha_i, \alpha_i^\star$

In deriving (3.34) we already eliminated the dual variables $\zeta_i, \zeta_i^\star$ through condition (3.33) which can be reformulated as $\zeta_i^\star$ = C- $\alpha^\star$ i . Eq. (3.32) can be rewritten as follows

$$w = \sum_{i=1}^{k}(\alpha_i - \alpha_i{}^\star); f(x) = \sum_{i=1}^{k}(\alpha_i - \alpha_i{}^\star)\langle x, x_i\rangle + b. \tag{3.35}$$

This is the called Support Vector expansion, i.e. w can be completely described as a linear combination of the training patterns xi. In a sense, the complexity of a function's representation by SVs is independent of the dimensionality of the inputs pace X, and depends only on the number of SVs. Moreover, note that the complete algorithm can be described in terms of dot products between the data. Even when evaluating f(x) we need not compute w explicitly.

3.2 Neural Networks and Deep Learning

An (artificial) neural network comprises a set of interconnected processing units [Bis95, p. 80-81]. Given input values $w_0, x_1, \ldots, x_D$, where w_0 represents an external input and $x_1, \ldots, x_D$ are inputs gotten from other processing units within the network, a processing unit computes its output as y = $f(z)$. Here, $f$is called activation function and z is obtained by applying a propagation rule which maps all the inputs to the actual input z.

This model of a single processing unit includes the definition of a neuron where instead of a propagation rule an adder is used to compute z as the weighted sum of all inputs.Neural networks can be visualized in the means of a directed graph3 called network graph [Bis95, p. 117- 120]. Each unit is represented by a node labeled according to its output and the units are interconnected by directed edges. For a single processing unit this is illustrated in figure 1 where the external input w_0 is only added for illustration purposes and is usually omitted [Bis95, p. 116-120]. Now,we distinguish input units and output units. An input unit computes the output y := x where x is the single input value of the unit whilst Output

units might accept an arbitrary number of input values. The network represents a function y(x) which dimensions are fixed by the number of input and output units.

3.2.1 Multilayer Perceptrons

A (L+1)-layer perceptron, illustrated in figure 2, consists of D input units, C output units, and several so called hidden units. The units are arranged in layers, that is a multilayer perceptron comprises an input layer, an output layer and L hidden layers4 [Bis95, p. 117-120]. The i^{th} unit within layer l computes the output

$$y_i^{(\ell)} = f(z_i^{(\ell)}) \quad \text{with} \quad z_i^{(\ell)} = \sum_k^{m^{(\ell-1)}} w_{i,k}^{(\ell)} y_k^{(\ell-1)} + w_{i,0}^{(\ell)} \tag{3.36}$$

where $w_{i,k}^{(\ell)}$ denotes the weighted connection from the k^{th} unit in layer$(\ell - 1)$ to the i^{th} unit in layer ℓ, and $w_{i,0}^{(\ell)}$ can be regarded as external input to the unit and is referred to as bias. Here, $m^{(\ell)}$ denotes the number of units in layer ℓ, such that $D = m^{(0)}$ and C = m(L+1). For simplicity, the bias can be regarded as weight when introducing a dummy unit $y_0^{(\ell)} := 1$ in each layer:

$$z_i^{(\ell)} = \sum_{m^{(\ell-1)}}^{k=0} w_{i,k}^{(\ell)} y_k^{(\ell-1)} \quad \text{or} \quad z^{(\ell)} = w^{(\ell)} y^{(\ell-1)} \tag{3.37}$$

where $z^{(\ell)}$, $w^{(\ell)}$and $y^{(\ell-1)}$ denote the corresponding vector and matrix representations of the actual inputs $z_i^{(\ell)}$, the weights $w_{i,k}^{(\ell)}$ and the outputs $y_k^{(\ell-1)}$, respectively.

3.2.2 Activation Functions

There are three types of activation functions and these are discussed: threshold functions, piecewise-linear functions and sigmoid functions. A common threshold function is given by the Heaviside function:

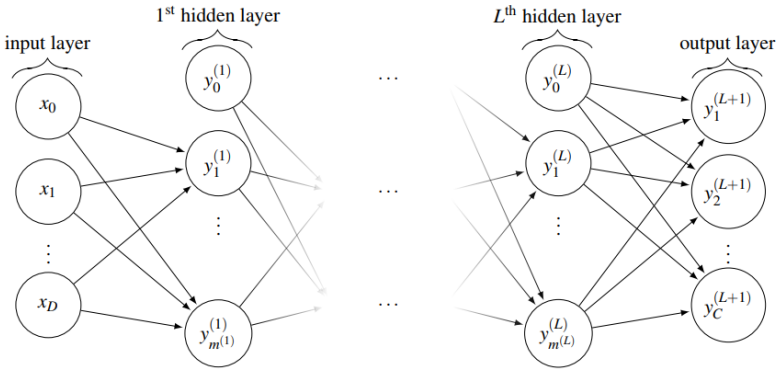


Figure 3.5: Representation on Hyperplanes

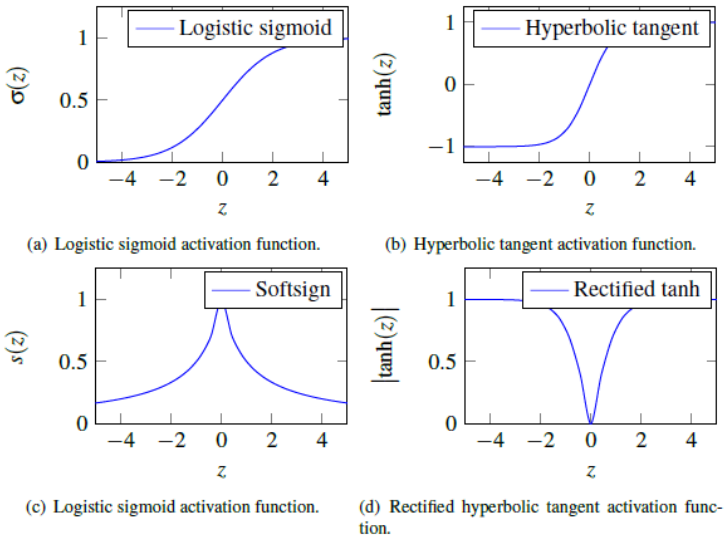
$$h(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{3.38}$$

However, both threshold functions as well as piecewise-linear functions have some drawbacks. First, for network training we may need the activation function to be differentiable. Second, nonlinear activation functions are preferable due to the additional computational power they induce [DHS01, HSW89]. The most commonly used type of activation functions are sigmoid functions. As example, the logistic sigmoid is given by

$$\sigma(z) = \frac{1}{de1 + \exp(-z)} \tag{3.39}$$

Its graph is s-shaped and it is differentiable as well as monotonic. The hyperbolic tangent tanh(z) can be regarded as linear transformation of the logistic sigmoid onto the interval $\big[-1, 1\big]$. When using neural networks for classification, the softmax activation function for output units is used to interpret the output values as posterior probabilities. Then the output of the i^th unit in the output layer is given by

$$\sigma(z^{(L+1)}, i) = \frac{exp(z^{(L+1)i})}{\sum_{k=1}^{C} exp(z_k^{(L+1)})} \tag{3.40}$$



(a) Logistic sigmoid activation function.

(b) Hyperbolic tangent activation function.

(c) Logistic sigmoid activation function.

(d) Rectified hyperbolic tangent activation function.

Figure 3.6: Activation Function

3.3 Convolutional Neural Network

Convolutional neural networks (CNNs) are a special type of Neural networks (NNs) well poised for image processing . The convolutional in the name owes to separate square patches of pixels in a image being processed through filters. As a result, the model can mathematically capture key visual cues such as textures and edges that help discerning classes.

The convolutional layer and subsampling (also called pooling) layer are the key layers of a convolutional neural network (CNN). The traditional multi-layer perception neural networks have too many weights and often over fitting. CNNs use three techniques to solve above problems ,local receptive elds, shared weights and sub-sampling. Each unit of a layer receives input from a set of units located in a small neighborhood in the previous layer. The technique is called local receptive elds.

The convolutional layer is composed of feature maps. All units in a feature map share the same weights. The shared weights technique will reduce many weight parameters and form more useful feature maps. Subsampling layer is also called polling layer. The subsampling layer will reduce the resolution of the feature maps and reduce the sensitivity of the output to shifts and distortions.

We then assume a grayscale image to be by defined by a function.

$$I : 1, \ldots, n_1 * 1, \ldots, n_2 \to W \subseteq \Re, (i, j) \mapsto I_{i,j}$$

Given the filter K $\in \Re^{2h_1+1 \star 2h_2+1}$,the discrete convolution of the image I with filter K is given by

$$(I * K)_{r,s} =: \sum_{u=-hv_i}^{h_1} \sum_{v=-h_i}^{h_2} K_{u,v} I_{r+u.s+v} \tag{3.41}$$

where r,s are output positions and the filter K is given by

$$\begin{vmatrix} K_{-h_1,-h_2} & \dots & K_{-h_1,h_2} \\ \vdots & K_{0,0} & \vdots \\ K_{h_1,-h_2} & \dots & K_{h_1,h_2} \end{vmatrix}$$

A commonly used filter for smoothing is the discrete Gaussian filter K_G and given by

$$\left(K_{G(\sigma)}\right)_{r,s} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{r^2+s^2}{2\sigma^2}\right) \tag{3.42}$$

where σ is the standard deviation of the Gaussian distribution

3.3.1 Activation Functions

In neural networks, the activation function of a node defines the output of the node given an input or set of inputs If the summation of the weight inputs was used as the output of a node, the entire neural network, no matter how many layers or how many nodes used, could only model a linear function. Passing the weighted sum into a nonlinear activation function allows MLPs and other multilayer ANNs to model nonlinear functions. Nearly any nonlinear function can be used as an activation function, but rectified linear units (ReLU, leaky ReLU , Sigmoid and Hyperbolic tangent) are most popular. We used Rectifier Linear Unit (RELU) as the activation function in this work

3.3.2 Reducing Overfitting

A common problem during training is the network may overfit to the training dataset and perform worse when deployed than training suggests. This is because the network begins to learn noise rather than features in the training dataset that is not representative of the actual data of interest.

There are several tools commonly used to prevent over fitting; namely early-stopping and dropout. In addition to the training dataset, a completely separate small dataset is used for validation. It is important that the validation dataset is unique from the training dataset, but is still representative of the same types of data the network was trained on. After a set number of training iterations, the network tests its accuracy by predicting the outputs on the validation dataset while all weights are frozen. There are many methods to reduce over-fitting, such as label-preserving. This method will set the output of each hidden neuron to zero with an adjustable probability. Through the method, more useful robust feature can be learned .

3.3.3 Convolutional Layer

Let layer ℓ be a convolutional layer. Then the input layer ℓ comprises $m_1^{(1-\ell)}$ features maps from the previous layer, each of size $m_2^{(1-\ell)}m_3^{(1-\ell)}$. In the case of layer ℓ =1 , the input is a single image I consisting of one or more channels. This way, a convolutional neural network directly accepts raw images as inputs. The output of layer ℓ consists of $m_1^{(\ell)}$ features maps of size $m_2^{(\ell)}m_3^{(\ell)}$. The i^{ith} feature map in layer ℓ, denoted Y_i^{ℓ}.

$$Y_i^{\ell} = \beta_i^{(\ell)} + \sum_{j=1}^{m_1^{(1-\ell)}} K_{i,j}^{(\ell)} * Y_i^{(1-\ell)} \tag{3.43}$$

where $\beta_i^{(\ell)}$ is a biaz matrix and $K_{i,j}^{(\ell)}$ is the filter of size $2h_1^{(\ell)} + Ix2h_2^{(\ell)}$ +1 connecting the j^th feature map in layer $(\ell - 1)$ with the i^{th} feature mapping layer ℓ. As mentioned above , $m_2^{(1-\ell)}$ and $m_3^{(1-\ell)}$ are influenced by border effects. When applying the discrete convolutional only in the so called valid region of the input feature maps, that is only for pixels where the sum of the equation is defined properly, the output feature have the size.

$$m_2^{(\ell)} = m_2^{(1-\ell)} - 2h_1^{(\ell)} \quad \text{and} \quad m_3^{(\ell)} = m_3^{(1-\ell)} - 2h_2^{(\ell)} \tag{3.44}$$

Often the features used in computing a feature map $Y_i^{(\ell)}$ are the same as $K_{i,j}^{(\ell)} = K_{i,k}^{(\ell)}$ for $j \neq k$. In addition , the sum in equation (13) may also run over a subset of input maps. To relate the Convolutional layer and its operations to the multi-layer perceptron, we rewrite the above equation .

$$\left(Y_i^{(\ell)}\right)_{r,s} = \left(\beta_i^{(\ell)}\right)_{r,s} + \sum_{j=1}^{m_1^{(1-\ell)}} \left(K_{i,j}^{(\ell)} * Y_i^{(1-\ell)}\right)_{r,s} \tag{3.45}$$

$$= \left(\beta_i^{(\ell)}\right)_{r,s} + \sum_{j=1}^{m_1^{(1-\ell)}} \sum_{u=-h_1^{(\ell)}}^{h_1^{(\ell)}} \sum_{v=-h_2^{(\ell)}}^{max} \left(K_{i,j}^{(\ell)}\right)_{u,v} * \left(Y_i^{(1-\ell)}\right)_{r+u,s+v} \tag{3.46}$$

Chapter 4

Background of Data

The study area has two regions located in Mexico; Puebla and Oaxaca. The 2017 earthquake in puebla according to the National Seismological Service of Mexico, the epicenter was located 12km southeast of Axochiapan, Morelos and 120km from the Mexico city. The earthquake measured a magnitude of 7.1 occurring at 13;14;40 central Daylight Time at the depth of 51km.

The 2018 Oaxaca earthquake occurred on February 16,2018 at 17:39 local time in the Sierra Madre del Sur in Oaxaca States in Southern Mexico. It had a magnitude of 7.2 on the moment magnitude scale and a maximum felt intensity of VII on the Mercalli intensity scale then the US Geological Survey reported he magnitude of the earthquake that hit southern southern Mexico on Thursday at 8.1 making the largest in Mexico in 100 years. It was larger than the one in 1985 when thousands were killed in four Mexican states.

4.0.1 Analysis of the Data

We analyst the data for both regions. The maximum magnitude for region A is 7.1 which occurred 11.6 days after the maximum magnitude earthquake in region B of magnitude 8.1. The negative values indicates days after the earthquake and the positive indicates days before the earthquake.

First, we plot the data set of longitude,magnitude,depth,month and day to investigate the potential pattern in the Peubla data and its observed that the month,latitude and day are linear in the years except for the depth which increases and decreases at certain points

Figure 4.1: Distance from Puebla to Oaxaca regions

which is because the Earthquakes occur at different depths from near the Earth's surface so they tend to bend from the figure 4.5.

Also, from the Oaxaca data all the variables remained constant in the years excluding the depth which clustered at certain points and gives a concave shape, which might imply that the occurrence of the earthquake in the earth crust was nearly at the same depth as sen in figure 4.6.Thus we examine the relationship between the depth of the Puebla and the Oaxaca respectively.

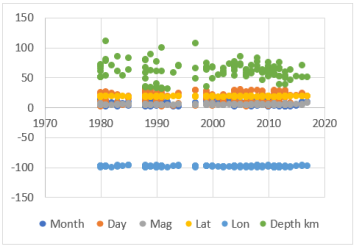


Figure 4.2: Puebla

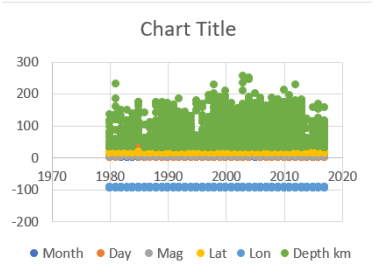


Figure 4.3: Oaxaca

4.1 Exploratory Data Analysis

4.1.1 Depth of the data

After analyzing the whole data set and breaking it to individual components, its observed that for each data set the fluctuations depends on the years but the data set for the depth for both regions changes significantly as the other factors remain constant.

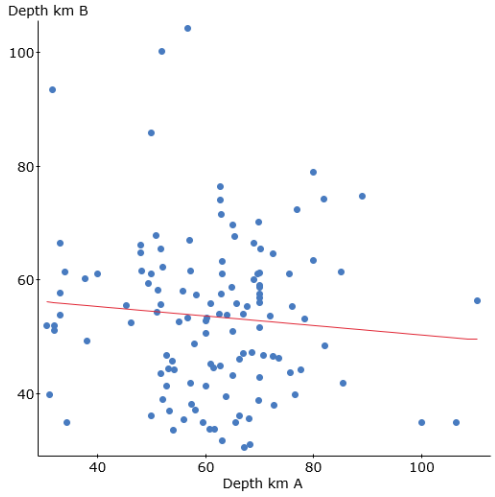


Figure 4.4: Comparison of the Depth from the two Regions

By observing 4.8 a negative line pattern suggests a weak association between the two data. Even though the data from Region B is quite larger than that of region A we are certain that there is no linear relationship between the two and thus with a correlation of -0.0088. We use the linear regression to fit the depth data which yields;

$$D_B = 58.761 - 0.0835 D_A + \epsilon \tag{4.1}$$

We consider that the existing termϵ is for the following reasons:

- The Data given from Region A and B are of their size. So we assume that the either the constant or the intercept might be a result of inconsistency of the data.
- Many factors in the test environment can affect the result of the data collection. When people in industry collect data, the different tests cannot be done at the same time. If the experimenter takes the data at different days as indicated in both data it may cause the nature of the earthquake change.

Table 4.1: Statistical Analysis for Depth

Parameter	Estimate	Standard Deviation	P-value
intercept	58.7611	5.3841	<0.0001
Slope	-0.0835	0.0854	<0.3299

4.1.2 Association between the features of the two different regions

The correlation is found between the features of the two regions. From our deductions its clear that there is little correlation between them which can be observed from table .

Table 4.2: Correlation of the two data set

	MagA	MagB	LatA	LatB	LonA	LonB	DepthKmA	DepthKmB
MagA	1	0.03	0.315	-0.108	0.131	0.049	0.043	-0.102
MagB	0.03	1	0.079	-0.062	0.131	0.056	0.032	0.073
LatA	0.315	0.079	1	-0.071	-0.166	0.169	-0.068	-0.107
LatB	-0.108	-0.062	-0.071	1	-0.033	-0.652	0.186	-0.126
LonA	0.131	0.085	-0.166	-0.033	1	0.015	0.318	0.134
LonB	0.049	0.056	0.015	-0.652	0.169	1	-0.652	0.329
DepthKmA	0.043	0.032	-0.068	-0.126	0.318	0.186	1	-0.088
DepthKmB	-0.102	0.073	-0.107	0.214	0.134	0.329	-0.088	1

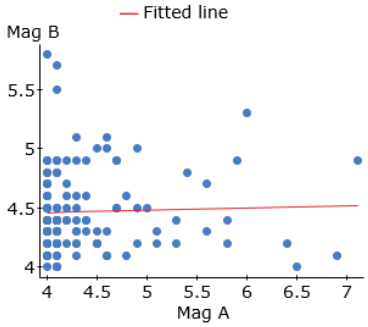


Figure 4.5: Magnitude

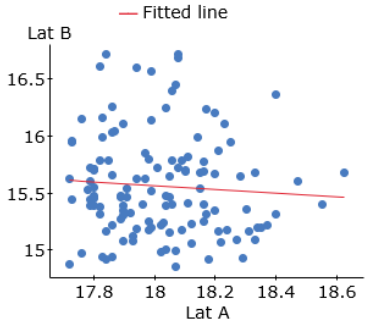


Figure 4.6: Latitude

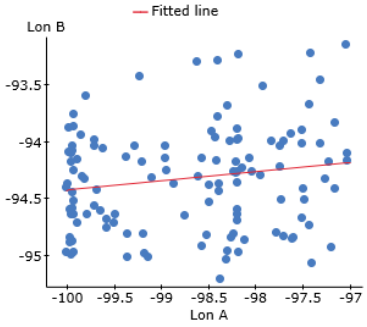


Figure 4.7: Longitude

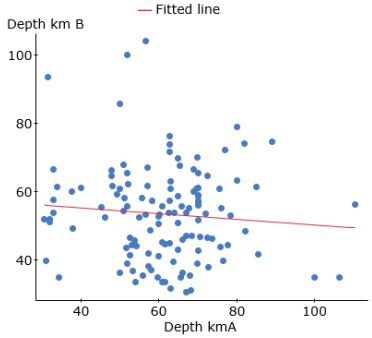


Figure 4.8: Depth

In the above figures the correlation coefficient between each feature are very weak correlated and approaches to zero except for the latitude which gives a weak correlation of 0.2. Thus there is little correlation found in both regions.

An earthquake's size is typically reported simply by its magnitude, which is a measure of the size of the earthquake's source, where the ground began shaking.

Now our focus is on the magnitude of the quakes in each region.

In magnitude figure above , we hypothesize that it looks like a function of the form $M_B = -a - M_A + f$ for some a, $f > 0$, however there are many types of models. Thus using the linear regression model to fit the data;

$$M_B = 4.386 + 0.011 M_B$$

In the Table below shows how well the statistical model fits in the data set. its observed that 90 percent of the total variation can be explained by the linear relationship. In addition the correlation coefficient is 0.033 which indicates a a very weak correlation since it approaching zero. The error of standard deviation is small and the also since the p - value for the slope is greater than its true that there is no significant relationship between them. we assume our model is accurate.

Table 4.3: Statistical Analysis for Magnitude

Parameter	Estimate	Standard Deviation	P-value
intercept	4.386	0.2197	<0.0001
Slope	0.001809	0.04935	<0.7147

4.2 Results and Discussion

4.2.1 Analysis SVM

The SVM requires both positive and negative data to train the model. The data was set into two groups: 70% of the data was a training set and 30% was for the test set.As mentioned before, negative training data was also needed. The Negative values only means time

(days) before this earthquakes and positive values means time (days) after the earthquake. In SVM modelling, the input of controlling factors should be as a vector of real numbers. The performance of the SVM model is depended on the choice of kernel functions and their parameters especially the penalty factor C and γ terms.

In this study, we compared sigmoid, polynomial and radial kernel. Pairs of (C, γ) were generated through a grid search with C = $10^1, 10^0, 10^1$ ... 2, 3, 4 and γ = 0.5,1,2,3 .... For region A the best value of C and γ for sigmoid were 1 and 1with the RMSE 0.680 . In the case of polynomial, the best C andγ were 1 and 1 respectively, with the RMSE 0.680 while radial used 1 and 1 as the best C and γ with RMSE of 0.626 for radial. This indicates that radial performs better with a minimum root mean square error ; w =-5.132 and b = -0.081 for region A. Where w and b are the weight and bias respectively.

The best value of C and γ for sigmoid were 1 and 0.1 with the RMSE 1.1168 . In the case of polynomial, the best C and γ were 1 and 0.1 respectively, with the RMSE 0.087 while radial used 1 and 0.1 as the best C and $\gamma$with RMSE 0.085 for region B. This indicates that radial performs better with a minimum root mean square error ; w = -0.09701021and b = 0.2538338

4.2.2 Analysis for CNN

The model type we are using is sequential which is an easiest way of building a model in the package Keras We used momentum and weight delay technology in our model. In the formulas i is the iteration index, $\triangle$ is the weight variable β is the momentum $\triangle$ ω is the weight decay , L is the cost function ,α is the learning rate and and $\left(\frac{\delta L}{\delta w}|_{wi}\right)_{Di}$ is the average over the ith batch $D_i$of the derivative of the objective with respect to w evaluated at w_i

$$\triangle w_{i+1} = \beta * \triangle w_i - \lambda * \alpha * \left(\frac{\delta L}{\delta, w}|_{wi}\right)_{Di}, \tag{4.2}$$

$$\triangle w_{i+1} = w_i + \beta * \triangle w_i \tag{4.3}$$

In the selection of learning rate, we found that higher (0.1) and lower (0.0001) base learning rates would decrease the accuracy(only about 55%). The momentum were used to accelerated model convergence. In this paper, we used 0.9 as the momentum value. Weight decay is one kind of regularization and will decrease useless weight to zero . We used a default for the weight decay and the learning rate . The dropout rate of the two dropout layers was 0.5, meaning the relative hidden neuron weight would be set to zero with the probability 0.9 to avoid over fitting. The loss function during the training of the models that achieved the best results reported in this paper was the standard Softmax loss function, that is, the log-likelihood error function. For better normalization, we used the rectified linear unit activation function(ReLU), not the sigmoid activation function.

Model

Layer (type)	Output Shape	Param #
dense_16 (Dense)	(None, 8)	64
dense_17 (Dense)	(None, 4)	36
dense_18 (Dense)	(None, 1)	5

Total params: 105
Trainable params: 105
Non-trainable params: 0

Figure 4.9: Model Configuration

4.2.3 Loss and Accuracy

In the case of neural networks, the loss is usually negative log-likelihood and residual sum of squares for classification and regression respectively. Then naturally, the main objective

in a learning model is to reduce (minimize) the loss function's value with respect to the model's parameters by changing the weight vector values through different optimization methods, such as back propagation in neural networks.
Loss value implies how well or poorly a certain model behaves after each iteration of optimization. Ideally, one would expect the reduction of loss after each, or several, iteration(s). From our plot at first the loss decrease fast - the highest decrease is always during the first 1 to 5 epochs. Afterwards it slowly continues converging until it reaches some minima, which occurs around epoch 100. After this minima it heavily decreases then it finds converges a bit until epoch at 500 for region A . For region B the loss (train+test) decreases fast till it c reaches a minimal point around 0 epoch where it decreases a little then converges from the 100 to 500 epoch. This results indicates that our model is accurate in training our datasets.

The accuracy of a model is usually determined after the model parameters are learned and fixed and no learning is taking place. Then the test samples are fed to the model and the number of mistakes (zero-one loss) the model makes are recorded, after comparison to the true targets.
From the plot of accuracy we can see that the model could probably be trained a little more on region A dataset as the trend for accuracy on it is still rising for the last few epochs but the region B dataset remains constant from the 100th to the last epoch. We can also see that the model has not yet over-learned the training dataset for the data for region A .

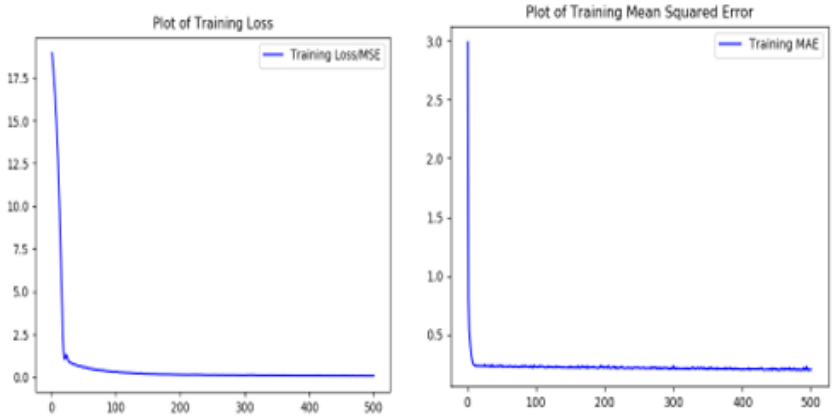


Figure 4.10: Training Loss for Region A and B

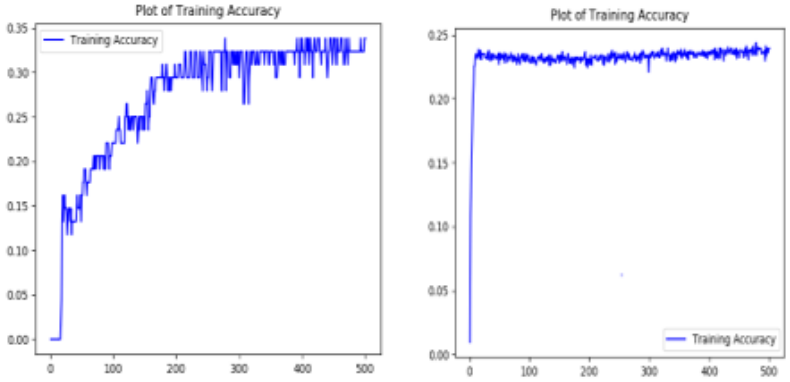


Figure 4.11: Training Accuracy for Region A and B

4.2.4 Comparison of SVM and CNN

Table 4.4: Comparing the prediction model for SVM and CNN for region A

Actual	Predicted SVM	Predicted CNN
4.9	4.194	4.408
4.0	4.182	4.234
4.3	4.177	4.116
4.0	4.329	4.024
4.1	3.907	3.993

Table 4.5: Comparing the prediction model for SVM and CNN for region B

Actual	Predicted SVM	Predicted CNN
4.1	4.434	4.463
4.8	4.406	4.368
4.2	4.312	4.465
4.7	4.394	4.503
5.1	4.394	4.567

Table 4.6: Root Mean Square Errors

	SVM	CNN
Region A	0.64	0.906
Region B	0.32	0.3625

4.3 Summary of Results

From the tables the predicted values for both models coincides with the actual values of the magnitude data . Also, it was observed that region B covered more after shots. There is a correlation between the two data sets. Latitude A and B are correlated.

Cross validation was an efficient tool for parameters optimization,this method avoided the subjectivity in parameter selection for the SVM model.These are parameters to the learning algorithm (and hyperparameters) whose main purpose is to identify learning parameters that generalises well across the population samples we learn from in each fold. Thus hyperparameters in this model are the kernels; cost and gamma.

The predicted magnitude for the maximum magnitude in region A is 4.4 and the predicted value for the maximum magnitude in region B is 5.15. For CNN the predicted magnitude for in region A is 4.9 and the predicted value for region B is 5.63. Furthermore, it could be observed that the models are good on the data in region B but needs a little training on region A for both SVM and CNN. Moreover, In comparing the root mean square error of both models for the two different regions, svm has RMSE of 0.64 and 0.32 for regions A and B respectively while CNN has 0.906 and 0.3625 respectively.

4.4 Error Analysis

In error analysis it is important to understand how to express a data, how to analyze and draw meaningful conclusions from it. Thus the descriptive statistics is the method to use on the data sets

4.4.1 Error Analysis for Magnitude in Region A

- From the real and the predicted data, we observed that the error is very small which might occur as a result of variations in the measurements. The experimenters have little or no control for variations in the measurements, or the error is caused by the way the experiment was conducted.

- From the Figure 4.12, all the values seems to come closer to zero and clusters are

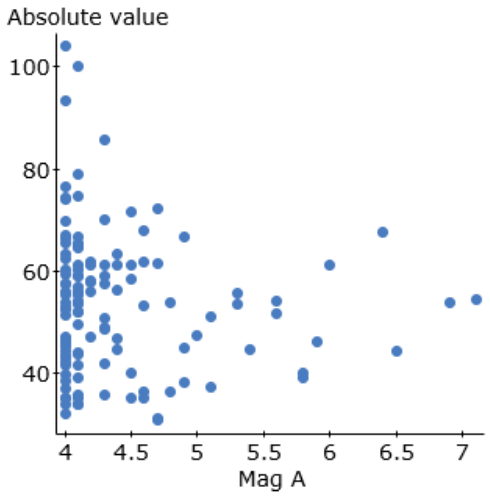


Figure 4.12: Absolute error for Magnitude A

from 4th to 4.5th . The amount of outliers is very small, thus the model has a high performance rate.

- The standard deviation of the absolute error in Table 4.7 is an indication of the dispersion of the residuals, which can be observed on how they are dispersed. standard deviations are small indicating , the predicted values seems reliable.

- The 95%confidence level indicates that we are 95% confident to use the model to predict the value within the range of $[M_A - 755; M_A + 755]$.

4.4.2 Error Analysis for Magnitude in Region B

- From the real and the predicted data, we observed that the error is very small. indicating a high performance .

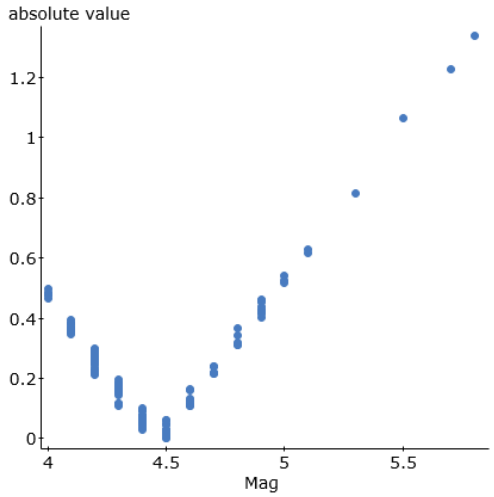


Figure 4.13: Absolute error for Magnitude B

- From the Figure 4.13, most of the errors approach the 4.5th then increases.
- The standard deviation is small in absolute sense, the dispersion of the error is small enough and thus the model is reliable.

Table 4.7: Comparing the Absolute Errors of Both Regions

	Region A	Region B
	Absolute error	Absolute error
mean	0.411	0.259
standard.d	0.429	0.22
variance	0.184	0.049
confidence In.	755.03	0.0392

Chapter 5

5.1 Concluding Remarks

In this work we used two machine learning models namely; Support Vector Machine and Convolutional Neural Network for earthquake magnitude prediction. Now our discussion follows. We used three kernel functions which includes the polynomial , RBF and sigmoid function in this work. SVM is a good topic in the fields of statistical theory learning and machine learning, which is in a boom period where as deep learning is a set of powerful machine learning algorithms and concepts that have seen groundbreaking success now.

In the current study of earthquake prediction, the physical processes of earthquakes are not known clearly so a lot of researchers have applied mathematical statistical methods to earthquake prediction. we introduced two models in predicting the maximum magnitude from different cities of the same state. Conclusively, Support Vector Machine is the perfect model for this data sets.

5.2 Future Work/ Recommendations

- We may also determine if the speed of the earthquake enhanced its cascading effects, by promoting coastal and submarine landslides.
- The data for region A should be large for better predictions.
- The data should be grouped in years so that during predictions the years will also be included for a diverse range of prediction.

References

[1] M.C. Mariani, H. Gonzalez, Huizar, M.A.M. Bhuiyan, O.K. Tweneboah. *Using Dynamic Fourier Analysis to Discriminate Between Seismic Signals from Natural Earthquakes and Mining Explosions.* (2017), AIMS Geosciences, 3(3), 438-449.

[2] M.C. Mariani, M.A.M. Bhuiyan, O.K. Tweneboah, H. Gonzalez-Huizar, I. Florescu (2018) *Volatility model Applied to geophysics and high frequency financial market data, Physica A: Statistical Mechanics and its Applications, 503, 304-321.* Statistical Mechanics and its Applications, 503, 304-321.

[3] Suhua Zhou and Ligang Fang *Support vector machine modeling of earthquake induced landslides susceptibility*

[4] S Carrara A, Guzzetti F, Cardinali M, Reichenbach P (1999) *Use of GIS technology in the prediction and monitoring of landslide hazard*

[5] Jiang C, Chen H R, Tian S and Wang J G (2000) *Matter-element models for comprehensive earthquake prediction and their applications.* Acta Seismologica Sinica 13(4): 448453.

[6] Chun Jiang , Xueli Wei Xiaofeng Cui and Dexiang You , *Application of support vector machine to synthetic earthquake prediction* Kluwer Academic Publishers, Norwell, MA, 1996.

[7] Thibaut Perola, Micha et Gharbib,Marine A. Denollec *Convolutional Neural Network for Earthquake Detection and Location*

[8] Kong, R. M. Allen, L. Schreier, Y.-W. Kwon, Myshake. *A smartphone 375 seismic network for earthquake early warning and beyond.* Science ad376vances 2 (2016) e1501055.

[9] Icon Valley, San Jose, *I*nterpreting Deep Convolutional Neural Networks for Handwritten Digit/Letter Recognition.

[10] Purdy Ho, Bernd Heisele, Tomaso Poggio *Face Recgonition with Support Vector Machine.*

[11] https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regressi

[12] https://www.hindawi.com/journals/cmmm/2019/6509357/

[13] Jun HE,Yue LIU,Shuai LI,Jin-ming SHEN *An Analysis of Convolutional Neural Networks for Image Recognition*

[14] http://svms.org/introduction.html

[15] https://www.researchgate.net/publication/245095741 Predicting Time Series with a Local Support Vector Regression Machine

[16] Sarker Md, Tanzim Sadia, Yeasmin Muhammad, Abrar HussainT. M. Rezoan, Tamal-Rashidul, HasanTanjir Rahman *Analysis of Spatial Data and Time Series for Predicting Magnitude of Seismic Zones in Bangladesh*

Curriculum Vitae

Esther Kesewa Amfo was born on June 11, 1992. She graduated from Kwame Nkrumah University,Kumasi, Ghana, in 2016.

In the fall of 2017, she entered the Graduate School of The University of Texas at El Paso. While pursuing a master's degree in Mathematical Science, she worked as a Teaching Assistant. she was a member of SAMSI, the Chapter of Black Engineers and a Treasurer for African Students Organization.

Permanent address: 240 Porfirio Diaz Street
El Paso, Texas 79912-4927