Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 2, 2021, page 11-19

Algorithms of Optimization in Machine Learning

[1]Ainun Safwani Mahidi @ Mohyedin and [2]Ong Chee Tiong

[1,2]Department of Mathematical Sciences
Faculty of Science, Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia.

e-mail: [1]ainunsafwani@graduate.utm.my, [2]ongutm@gmail.com

Abstract This study focuses on the understanding of how optimization methods can be applied in the field of machine learning. This can be done by surveying the machine learning literature and the optimization methods that exist in aiding the development of machine learning. Gradient descent method was used as the benchmark in this study. The computational results by the data sets are obtained and compared. The comparison is between one with normalization and without normalization. The result is studied to understand the whole idea of optimization in machine learning.

Keywords Optimization method; machine learning; gradient descent.

1 Introduction

Machine learning (ML) is a subset of artificial intelligence (AI) in which it is capable to learn from past experience or historical data in order to provide the best result. Mathematics is the fundamental or the basic core for the development of machine learning specifically in the field of linear algebra, statistics and probability theory, multivariate calculus and optimization. Seagate Technology PLC, an American data storage company reported that International Data Corporation (IDC), a market research company predicts that the Global Datasphere will have a 400% increase by 2025 from 2019 in one of their study titled Data Age 2025 [1]. This clearly shows that the world is slowly moving forward into an era where the existence of data plays an important role in decision making. It is said that the demand for machine learning will increase with the increase in the availability of the data [2].

The implementation of machine learning is widely being used in a lot of industries that can aid human in doing their daily works. For instance, there are speech recognition, image recognition, medical diagnosis and financial services [3]. People used them a lot in their daily life but it is a wonder that most of the times, the results that the algorithms showed are nearly to what it is expected it to be. Most of the time, the users probably shrugged it off thinking it is not something to their concern as a consumer but for companies that use machine learning to attract more customers to buy their products or use their services, it is important for them as it can bring more profits. But despite all of that, the mathematics behind every machine learning is what makes it the way it is now.

There are a few common challenges that engineers and scientists often found in various industries such as aerospace, construction, pharmaceuticals, transportation and energy in which machine learning will be able to solve those challenges. One of it is that it helps to predict the

outcomes more accurate and quickly or in other words, it will be able to accelerate the processing time and at the same time increase the efficiency of the engineers and scientists works. Not only that, in a situation where reliability and safety are paramount, machine learning will be valuable in modelling the probability of different outcomes in a process where it is difficult to do the predictions due to randomness or noise. Furthermore, if there are gaps in a data set which can actually restrain the accuracy of the learning, inference and prediction, machine learning can help to compensate for those missing data. Since, it is known that more relevant data can improve the models that are trained by the machine learning itself.

Machine learning itself has three major categories which are supervised learning, unsupervised learning and reinforcement learning [4,5]. There are also a variety of machine learning algorithms that exist such as Linear Regression, Logistics Regression, Decision Tree, Support Vector Machine (SVM), Naïve Bayes, KNN (K-Nearest Neighbours), K-Means, Random Forest, Dimensionality Reduction Algorithms and Gradient Boosting & AdaBoost [6,7,8]. All of these are important in order to create functional machine learning projects. When something is successfully created from those machine learning projects, it is important to always optimize its usage. Therefore, in this study, some of the algorithms of optimization in machine learning will be studied in order for it to be more efficient and to give the result or prediction that is near to the actual result.

2 Literature Review

2.1 Type of Learning

2.1.1 Supervised Learning

A supervised learning algorithm is basically the first thing that machine learning practitioners should know first. It is known that this type of algorithm is designed to learn by example [9]. Just like how it stands out for its name “supervised”, they learnt from labelled training data to help in order to predict the outcomes for unforeseen data. The training data will consist of inputs paired with their correct outputs. The algorithm will search for some patterns in the data that correlate with the desired outputs. After that, they will be able to take in new unseen inputs and will determine which label the new inputs will be classified as.

2.1.2 Unsupervised Learning

Unsupervised learning is more complicated than supervised learning. If supervised learning is seen as a teacher teaching their students, the unsupervised learning is like without having a teacher to guide the students. Basically, unsupervised learning allows the model to work on its own to discover patterns and information that was previously undetected [10]. Thus, it mainly deals with the unlabelled data. In unsupervised learning, the training data consists of a set of input vectors x without any corresponding target values [11]. Its main goal is to discover groups of similar examples within the data which is called as clustering [13, 14].

Clustering algorithms divide a group of samples into multiple clusters ensuring that the differences between the samples in the same cluster are as small as possible and samples in different clusters are as different as possible. The four most used clustering algorithms are K-means, Fuzzy K-means, Hierarchical clustering and Mixture of Gaussian [11]. Unsupervised learning does have some drawbacks such as we cannot get precise information regarding data sorting and the output as the data used in it is unlabelled and not known [10]. Thus, bringing us some inaccuracy in the results obtained from it.

2.1.3 Semi-Supervised Learning

Semi-supervised learning is something that is between supervised learning and unsupervised learning or it can also be considered as a technique that can combine both supervised and unsupervised learning. During the training process, they incorporate both labelled and unlabelled data. It is also known that the semi-supervised learning has many categories but some of them are Generative Models, Self-Training and Transductive SVM.

One of the oldest semi-supervised learning methods is the Generative Models. It assumes that a structure such as $p(x, y) = p(y)p(x|y)$ where $p(x|y)$ is a mixed distribution. The mixed components can be identifiable within the unlabelled data in which it is enough to confirm the mixture distribution with only one labelled example per component. For Self-Training, a portion of the labelled data is trained with a classifier that is learning by itself. The procedure consisted of the repetition on adding the unlabelled points and the predicted labels together in the training set. Lastly, in Transductive support vector machine or TSVM, both the labelled and unlabelled are being considered. The important part in this method is that they are trying to maximize the margin between the labelled and unlabelled data during the process of labelling the unlabelled data. It is a NP-hard problem when TSVM is used to find the exact solution.

2.1.4 Reinforcement Learning

Reinforcement learning aims to find an optimal strategy function, whose output varies with the environment. Even though reinforcement learning is a more complex and challenging method, it basically deals with learning via interaction and feedback [12]. In other words, it learns to solve a task by trial and error. There are three approaches to implement a Reinforcement Learning algorithm which are Value-Based, Policy-Based and Model-Based [13]. In Value-Based, we are trying to maximize a value function V(s) while in Policy-Based we are trying to come up with such a policy that the action performed in every state helps us to gain maximum reward in the future. Lastly, the Model-Based is where we need to create a virtual model for each environment in which the agent learns to perform in that specific environment.

2.1.5 Multitask Learning

Multitask learning aims to learn the models simultaneously for multiple related tasks [14]. The goal is simple in which it is going to help other learners to have a better performance. Therefore, the process is that the multitask learning algorithms will remember the procedure of how it solved the problem or how it reached a specific conclusion when the algorithms are applied on a task. After that, it will use these steps that it discovered before to find the solution when it encountered other similar problem or task. It can also be identified as inductive transfer mechanism since this type of learning is portraying as helping from one algorithm to another. Thus, instead of learning individually the learners will be able to learn synchronously which is much faster since the learners share their experience with each other.

2.1.6 Ensemble Learning

A form of combining various individual learners as one learner is known as ensemble learning. Naïve Bayes, decision tree and neural network are one of the few examples for the individual learner. The idea of it is that it is almost always better to do a particular job with a collection of learners instead of just using one learner. Boosting and Bagging are one of the popular Ensemble learning techniques.

Boosting is an ensemble learning technique that is used to decrease the value of bias and variance. The idea is that it created one strong learner converted from a collection of weak learners. What differentiate between the classifier of a strong learner and a weak learner is that they are

strongly correlated and barely correlated respectively with the true classification. One of the most popular examples of boosting nowadays is AdaBoost.

Meanwhile, bagging or also known as bootstrap aggregating is an ensemble learning technique in which it is applied when the machine learning algorithm needs to be increased in term of accuracy and stability. Classification and regression can be related to this process specifically. Furthermore, it is also known that bagging can also decreases variance and aids in handling over-fitting.

2.1.7 Neural Network Learning

Neural network learning can also be termed as artificial neural network (ANN) is a type of learning that is derived from the biological concept of neurons. Taking into consideration the cell like structure in a brain which is the neuron, one will understand the flow of the neural network learning more clearly. In a neuron, there are four main parts which are the dendrites, nucleus, soma and axon where they worked together to send electrical impulses around the brain. The neural network learning can be considered to have the same behavior as the neuron. There are three types of neural network learning which are Supervised Neural Network, Unsupervised Neural Network and Reinforced Neural Network.

2.1.8 Instance-Based Learning

The learner learns a specific type of pattern and will try to apply the same exact pattern to the newly fed data. Since it is a type of learner that will only wait for the test data to arrive first before acting on it together with the training data, it is also known as a type of lazy learner too. Therefore, the size of the data affects the complexity of the learning algorithm. The bigger the data, the more complex the learning algorithm is.

One of the popular examples of instance-based learning is *k*-nearest neighbour or also known as KNN. The well-labelled training data will be fed into the learner so that it will be able to compare those data with the test data. Hence, *k* most correlated data will be taken from the training set in which the majority of *k* will serves as the new class for the test data.

2.2 Fundamental Optimization Methods

2.2.1 First-Order Optimization Methods

In first-order optimization method, it revolves around of just using the first derivative in choosing the movement direction inside the search space. One of the most commonly used in the field of machine learning is based on gradient descent. Other examples also include stochastic gradient descent, nesterov accelerated gradient descent, adaptive learning rate method, variance reduction methods, alternating direction method of multipliers and frank-wolfe method.

2.2.2 High-Order Optimization Methods

High-order optimization methods can also be identified as the second-order methods. For this particular method, it is applicable to address the machine learning problem in a situation where the objective function is highly non-linear and ill-conditioned. Besides, by introducing the curvature information they will be able to work effectively. A few examples of algorithms that allow the use of high-order optimization methods in processing large-scale data are conjugate gradient method, quasi-newton methods, stochastic quasi-newton method, hessian-free optimization method, natural gradient and trust region method.

2.2.3 Derivative-Free Optimization Methods

Derivative-free optimization is used to find the solution of the optimal point in some problems in which the derivative of the objective function may not exist or it is considered as not easy to calculate. This discipline of mathematical optimization will be able to find the optimal solution even without acquiring the gradient information like what the first-order and high-order optimization do.

The ideas of derivative-free optimization are mainly surrounded around only on two types of ideas which are the heuristic algorithms and the idea of fitting an appropriate function according to the samples of the objective function. In heuristic algorithms, instead of systematically deriving the solutions, these algorithms choose the methods that have already worked well. Classical simulated annealing arithmetic, genetic algorithms, ant colony algorithms and particle swarm optimization are a few examples of heuristic algorithms. Nevertheless, even though they often obtain approximate global optimal values, the theoretical support is weak. For the second type of idea in derivative-free optimization, some constraints are usually attached to the search space in order to derive the samples. One of the typical derivative-free algorithms which can be extended and is easily applicable to the optimization algorithms in machine learning problems is the coordinate descent method.

3 Methodology

3.1 Linear Regression

In supervised learning, especially for predicting a quantitative response, linear regression models have been widely adopted. The relationship between the independent variables and the dependent variable is the main assumption in which it is representable with a linear function as well as with a reasonable accuracy. The simplicity, extensive range of applications and the easiness of interpretation attract many interests in linear regression models. The ability to explain in a humanly understandable way the role of the inputs in the outcome are what actually machine learning wants to interpret.

The aim of linear regression is to find a linear function f that will be able to express the relationship between an input vector x of dimension p and real-valued output $f(x)$ such as

$$f(x) = \beta_0 + x^T\beta \tag{1}$$

where $\beta_0 \in \mathbb{R}$ is the intercept of the regression line and $\beta \in \mathbb{R}^p$ is the vector of coefficients corresponding to each of the input variables. A training set (X, y) where $X \in \mathbb{R}^{n\times p}$ denotes n training inputs $x_1, \ldots, x_n$ and y denotes n training outputs where each $x_i \in \mathbb{R}^p$ is associated with the real-valued output y_i are needed to estimate the regression parameters β_0 and β.

One of the most commonly used loss function for regression is the least squared estimate in which the regression model will be fitted in order to minimize the residual sum of squares (RSS) such as

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 . \tag{2}$$

It is known that this type of loss function has the smallest variance among all the other linear unbiased estimates and it is said to have a closed form solution.

3.2 Gradient Descent Method

The gradient descent alternates the following two steps until it converges:
Derive $L(\theta)$ for θ_j to get the gradient corresponding to each θ_j:

$$\frac{\partial L(\theta)}{\partial \theta_j} = -\frac{1}{N}\sum_{i=1}^{N}(y^i - f_\theta(x^i)x_j^i\,. \tag{3}$$

Update each θ_j in the negative gradient direction to minimize the risk function:

$$\theta_j' = \theta_j + n\,.\frac{1}{N}\sum_{i=1}^{N}(y^i - f_\theta(x^i)x_j^i\,. \tag{4}$$

3.3 Feature Normalization

Feature normalization is a trick that is frequently used in machine learning. In this case, it will be extremely helpful for the gradient descent to reach convergence by normalizing the data. Equation (5) shows the formula for normalization

$$x_i = \frac{x_i - \max(x)}{\max(x) - \min(x)}, \tag{5}$$

where x_i is the i^{th} training example, $\max(x)$ denotes the biggest value in the data set and $\min(x)$ denotes the smallest value in the data set.

4 Results

4.1 Gradient Descent Method with Normalization

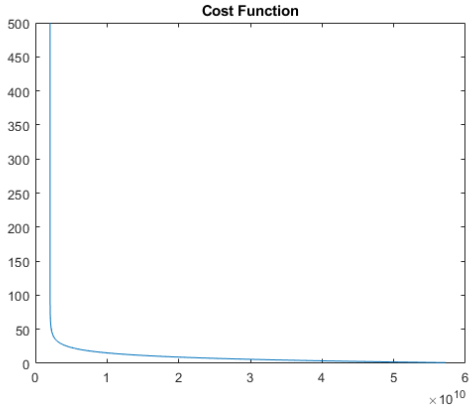

Figure 1: Cost Function for Dataset 1

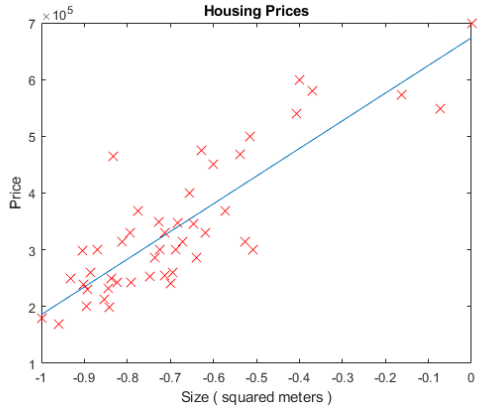

Figure 2: Graph of Hypothesis Line for Dataset 1

From Figure 1, Dataset 1 obtained the optimal solution with the learning rate of 1.3 and 500 iterations. Figure 2 shows the hypothesis line of Dataset 1 based on the chosen learning rate and iterations.

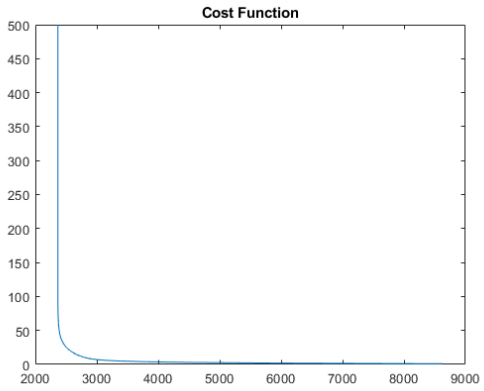

Figure 3: Cost Function for Dataset 2

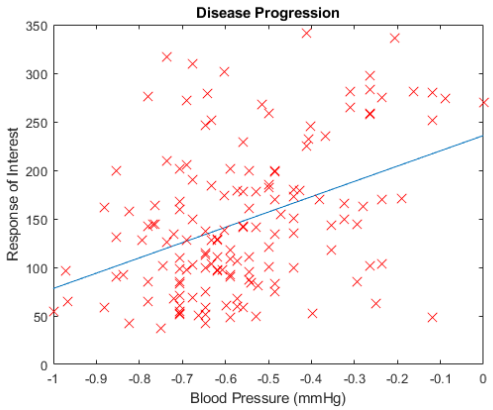

Figure 4: Graph of Hypothesis Line for Dataset 2

From Figure 3, Dataset 2 obtained the optimal solution with the learning rate of 1.3 and 500 iterations. Figure 4 shows the hypothesis line of Dataset 2 based on the chosen learning rate and iterations.

4.2 Gradient Descent Method without Normalization

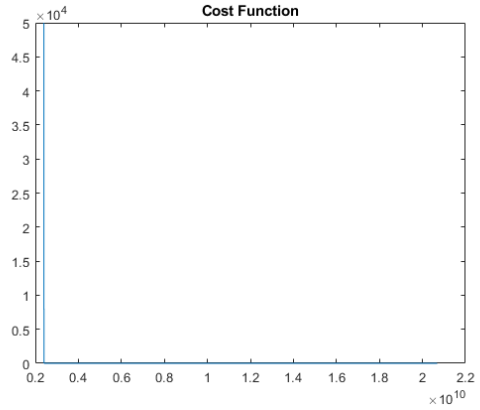

Figure 5: Cost Function for Dataset 1

Figure 6: Graph of Hypothesis Line for Dataset 1

From Figure 5, Dataset 1 obtained the optimal solution with the learning rate of 0.0000001 and 50000 iterations. Figure 6 shows the hypothesis line of Dataset 1 based on the chosen learning rate and iterations.

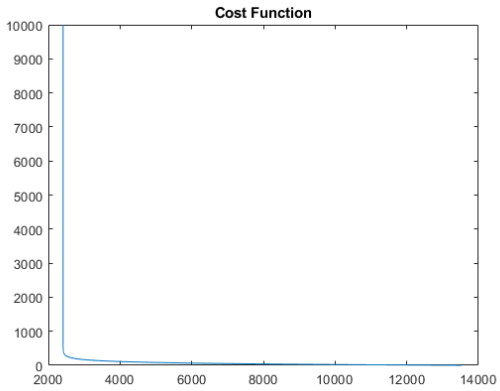

Figure 7: Cost Function for Dataset 2

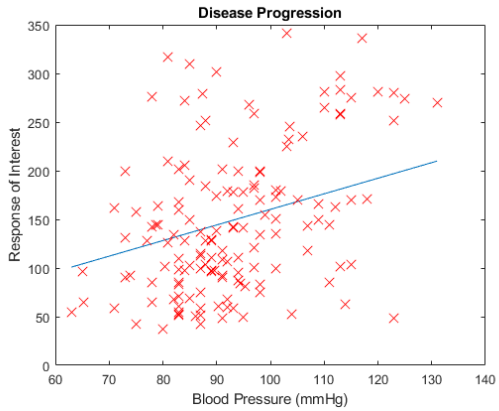

Figure 8: Graph of Hypothesis Line for Dataset 2

From Figure 7, Dataset 2 obtained the optimal solution with the learning rate of 0.000001 and 10000 iterations. Figure 8 shows the hypothesis line of Dataset 2 based on the chosen learning rate and iterations.

5 Conclusion

The results show that, it is more efficient to implement gradient descent with normalization rather than without one. This is due to the application of normalization will lead to less iterations needed for convergence. By doing this, the optimal solution will be obtained faster and it is more effective and efficient.

References

[1] "DataAge 2025 - The Digitization of the World: Seagate US." Seagate.com, www.seagate.com/as/en/our-story/data-age-2025/.

[2] Vanshika Rastogi, Sugandha Satija, Pankaj Kumar Sharma and Sanika Singh, Machine Learning Algorithms: Overview, International Journal of Advanced Research in Engineering and Technology, 11(9), 2020, pp. 512-517. http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=9

[3] Kaur, A., Wadhwa, A., Richards, A., NT, A., Jeevan, A., Rogers-Nelson, A., & H, A. (2020, January 21). Top 10 real-life examples of Machine Learning. Retrieved November 16, 2020, from https://bigdata-madesimple.com/top-10-real-life-examples-of-machine-learning/

[4] Heidenreich, H. (2018, December 04). What are the types of machine learning? Retrieved November 17, 2020, from https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f

[5] The Three Types of Machine Learning Algorithms - Pioneer Labs: Technology Strategists & Delivery Experts. (n.d.). Retrieved November 17, 2020, from https://pioneerlabs.io/insights/the-three-types-of-machine-learning-algorithms/

[6] Tavasoli, S. (2020, November 08). Top 10 Machine Learning Algorithms You Need to Know in 2020. Retrieved November 17, 2020, from https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article

[7] Ray, S. (2020, October 18). Commonly Used Machine Learning Algorithms: Data Science. Retrieved November 17, 2020, from https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
[8] Yildirim, S. (2020, July 27). 11 Most Common Machine Learning Algorithms Explained in a Nutshell. Retrieved November 17, 2020, from https://towardsdatascience.com/11-most-common-machine-learning-algorithms-explained-in-a-nutshell-cc6e98df93be
[9] Wilson, A. (2019, October 01). A Brief Introduction to Supervised Learning. Retrieved November 17, 2020, from https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590
[10] 99, G. (2020). Unsupervised Machine Learning: What is, Algorithms, Example. Retrieved November 17, 2020, from https://www.guru99.com/unsupervised-machine-learning.html
[11] Mishra, S. (2017, May 21). Unsupervised Learning and Data Clustering. Retrieved November 17, 2020, from https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a
[12] AI, S. (2019, February 18). Reinforcement Learning algorithms - an intuitive overview. Retrieved November 17, 2020, from https://medium.com/@SmartLabAI/reinforcement-learning-algorithms-an-intuitive-overview-904e2dff5bbc
[13] 99, G. (2020). Reinforcement Learning: What is, Algorithms, Applications, Example. Retrieved November 17, 2020, from https://www.guru99.com/reinforcement-learning-tutorial.html
[14] Smith, V., Chiang, C.-K., Sanjabi, M., & Talwalkar, A. (2017). Federated multi-task learning. In NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems (Vol. 30, pp. 4427–4437).