SURVEY OF OPTMIZATION METHODS FOR MACHINE LEARNING

Deepak Kumar Malviya
M.Tech.
Department of IT,UIT BU Bhopal MP
malviya.deepak@gmail.com

Umesh Kumar Lilhore
Associate Professor
Department of AI ,SAGE University Indore MP
umeshlilhore@gmail.com

ABSTRACT

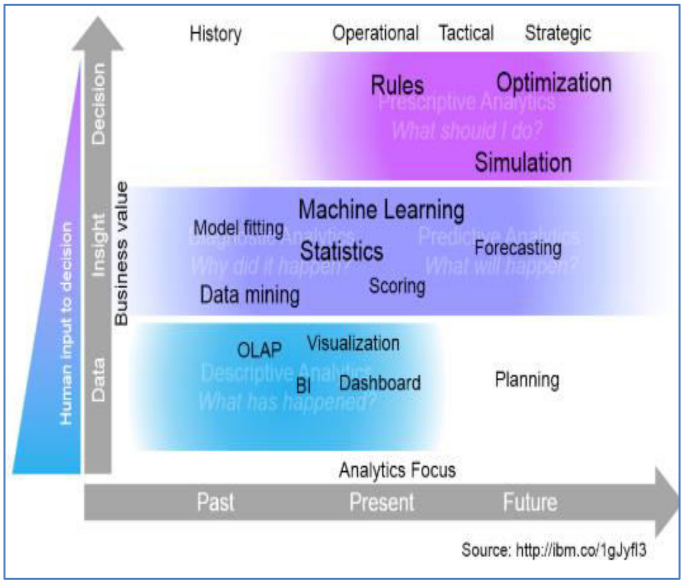
Machine learning is a paradigm that may refer to learning from past experience (which in this case is previous data) to improve future performance. The sole focus of this field is automatic learning methods. Learning refers to modification or improvement of algorithm based on past “experiences” automatically without any external assistance from human. ML is a sub branch of AI. In machine learning we trained a machine or method by using a training algorithm with some give set of input data. The Machine Learning is used to address a specific problem. However, the optimization of these systems is particularly difficult to apply due to the dynamic, complex and multidisciplinary nature.A ML method Extract knowledge from data and make predictions. Optimization is going through a period of growth and revitalization, driven largely by new applications in many areas.Numerical optimization has played an important role in the evolution of machine learning, touching almost every aspect of the discipline. Stochastic approximation has evolved and expanded as one of the main streams of research in mathematical optimization. In this paper we are presenting review of various optimization methods used in machine learning.

Keywords- *Optimization techniques, Machine learning, Artificial Intelligence*

1. INTRODUCTION

Machine learning (ML) is a branch of Artificial Intelligence that pushes forward the idea that, by giving access to the right data, machines can learn by themselves how to solve a specific problem. Machine-Learning (ML), can help discovering patterns and to perform certain tasks through the generalization of cases and the use of data [2]. As the basis of these decisions are the learning and knowledge systems. These systems are enriched with information in the form of structured or unstructured data to better search, match and get the best forecasts and analysis of the problem in question.

This issue raises fundamental philosophical questions about what constitutes "learning" in general, typically defined as: gain knowledge or skills, to study or experience; commit to memory; be warned, be informed; becoming aware; the behavior modification through interaction with the environment reasoning premises to conclusions. We can define information as data plus meaning (events) with significance, as knowledge plus experience can be considered wisdom in understanding the information [3]. Optimization lies at the heart of machine learning. By quantitatively formulating the objective of modeling, it allows machine learning methods to flexibly incorporate domain knowledge in applications such as computer vision and natural language processing [11].



Source: http://ibm.co/1gJyfl3

Figure 1.1 Machine learning

To effectively and efficiently solve the optimizations problem, large bodies of techniques have been developed recently. In this review paper we are presenting study and review of various optimization methods for machine learning. This paper covers introduction of machine learning, optimization methods used in machine learning, their importance and applications [5].

2. OPTIMIZATION OF MACHINE LEARNING

ML is focused on developing systems that learn from the data. This involves a training phase where the system learns to complete certain tasks (predictive or classification) using a given data set containing information representative of the problem. After the training phase, the system is able to analyze new data having the same set of parameters and suggest a prediction.

2.1 OPTIMZATION METHODS- Optimization algorithms helps us to minimize (or maximize) an Objectivefunction (another name for Error function) E(x) which is simply a mathematical function dependent on the Model's internal learnableparameters which are used in computing the target values(Y) from the set of predictors(X) used in the model. For example, we call the Weights (W) and the Bias (b) values of the neural network as its internal learnable parameterswhich are used in computing the output values and are learned and updated in the direction of optimal solution i.e minimizing the Loss by the network's training process and also play a major role in the training process of the Neural Network Model [7, 9].

The internal parameters of a Model play a very important role in efficiently and effectively training a Model and produce accurate results. This is why we use various Optimization strategies and algorithms to update and calculate appropriate and optimum values of such model's parameters which influence our Model's learning process and the output of a Model.

2.2 TYPES OF OPTIMIZATION ALGORITHMS- Optimization Algorithm falls in 2 major categories[10]–

2.2.1 First Order Optimization Algorithms-these algorithms minimize or maximize a Loss function **E(x)** using its **Gradient** values with respect to the parameters. Most widely used First order optimization algorithm is **Gradient Descent.**The First order derivative tells us whether the function is decreasing or increasing at a particular point. First order Derivative basically gives us a **line** which is **Tangential** to a point on its Error Surface.

2.2.2 Second Order Optimization Algorithms- Second-order methods use the **second order derivative** which is also called **Hessian** to minimize or maximize the **Loss** function.The **Hessian** is a Matrix of **Second Order Partial Derivatives**. **Since the second derivative is costly to compute, the second order is not used much** .The second order derivative tells us whether the **first derivative** is increasing or decreasing which hints at the function's curvature.Second Order Derivative provides us with a **quadratic** surface which touches the curvature of the **Error Surface**.

2.3 NEW MACHINE LEARNING MODELS USING EXISTING OPTIMIZATION METHODS - The special topic papers include novel machine learning models based on existing primarily convex programs such as linear, second order cone, and semi-definite programming. The reader unfamiliar with the basic convex programs can see their definitions in the Appendix. In these papers, the authors develop novel modeling approaches to uncertainty, hypothesis selection, incorporation of domain constraints, and graph clustering, and they use off-the-shelf optimization packages to solve the models.

2.3.1 Dealing with Uncertainty Using Second Order Cone Programming–Paper [11] "Second Order Cone Programming Approaches for Handling Missing and Uncertain Data" [3] presents an extension to SVM that deals with situations where the observations are not complete or present uncertainty. The SVM Quadratic Program (QP) problem is cast into a more convenient Second Order Cone Program (SOCP) and uncertainty is represented as probabilistic constraints (SVM slack variables turn out to be random variables). They also come up with an interesting geometrical interpretation of their method as every data point being the center of an ellipsoid and the points within this ellipsoid being assigned to the class of the center. The study is extended to multiclass classification and regression.

2.3.2 Convex Models for Hypothesis Selection- Author [11, 13] addressed the hypothesis selection. Looks at pruning an ensemble of classifiers constructed from a pool of already trained classifiers. The goal is to make the performance of the smaller group equivalent to that of the whole pool, thus saving storage and computational resources. Traditionally, this selection process has been carried out using heuristics or by using greedy search. In [21], the goal is to identify a small subset of hypotheses that exclude the true targets with a given error probability.

2.3.3 SDP Methods for Graph Clustering- The paper “Fast SDP Relaxations of Graph Cut Clustering, Transduction, and Other Combinatorial Problems” [7] proposes an SDP relaxation to the normalized cut problem. The normalized cut problem arises when one wishes to partition a data set where similarity relationships among instances are defined. The mathematical formulation of this problem leads to an intractable combinatorial optimization problem.

Spectral relaxation has been used to avoid this intractability. In spectral relaxation, the combinatorial optimization is cast onto a simplerEigen decomposition problem that gives the subsets of data. The new approach in [12] consists of an SDP relaxation of the combinatorial problem that turns out to be tighter than the spectral one, although at the expenses of a larger computational burden. Moreover, they also present a scheme to develop a cascade of SDP relaxations that allows control of the trade-off between computational cost and accuracy. This study is extended to applications in semi-supervised learning [15].

2.3.4 Refining the Classics: Improvements in Algorithms for Widely used Models Widely used methods such as SVM and Bayesian networks have well-accepted core optimization problems and algorithms. The demand for the ability to learn with massive amounts of data is increasing. The immediate answer to this demand from the optimization and machine learning communities is to try to come up with more efficient implementations of these solid and reliable optimization methods.

3. LITERATURE SURVEY OF OPTIMZATION METHODS

Machine learning incorporates a vast array of algorithmic implementations, not all of which can be classified as deep learning. For example, singular algorithms, including statistical mechanisms like Bayesian algorithms, function approximation such as linear and logistic regression, or decision trees, while powerful, are limited in their application and ability to learn massively complex data representations [1].

The Several applications mentioned earlier suggest considerable advancement so far in ML algorithms and their fundamental theory. The discipline is divulging in several direction, probing a range of learning problems. ML is a vast discipline and over past few decades numerous researchers have added their works in this field. The enumeration of these works is countable infinite and mentioning every work is out of the scope of this paper. However this paper describes the main research questions that are being pursued at present and provide references to some of the recent notable works on that task [2].

With the advancement of computing technologies, the implementation of large collections of neurons was possible, giving rise to neural networks. Indeed, though neural networks are becoming commonplace, they are actually an old technology] that fell out of favor because of complexity and computing deficiencies. Nonetheless, this has clearly changed, thanks in no small part to the applications at which neural networks have excelled [6].

Examples include winning the ImageNet object recognition competition [3], in which neural networks can exceed even human accuracy, or beating humans in the game of Go without having received any direct input or game sessions against human players [10].

Optimization and machine learning is complicated by the fact that machine learning mixes modeling and methods. In that respect, ML is much like operations research (OR). Mathematical programming/optimization is historically a subfield of OR. Here OR is concerned with modeling a system. Mathematical programming is concerned with analyzing and solving the model [17]. In ML, generalization is the most essential property used to validate a novel approach. For a practical ML problem, the ML analyst might pick one or more families of learning models and an appropriate training loss/regularization function, and then search for an appropriate model that performs well according to some estimate of the generalization error based on the given training data [15].

This search typically involves some combination of data preprocessing, optimization, and heuristics. Yet every stage of the process can introduce errors that can degrade the quality of the resulting inductive functions [12].The paper “Kernel-Based Learning of Hierarchical Multilabel Classification Models” [5] provides a more efficient framework for scenarios where the vector output describes a hierarchical relationship. Their formulation requires the solution of a large scale quadratic program. This method’s efficiency relies on a decomposition of the core problem into single variable sub problems and the use of a gradient-based approach [14].

Moreover, the optimization is enhanced by a dynamic program that computes the best update directions in the feasible set. The paper "Structured Prediction, Dual Extra gradient and Bregman Projections" [12] proposes simple scalable maximum margin algorithms for structured output models including Markov networks and combinatorial models. The problem is to take training data of instances labeled with desired structured outputs and a parametric scoring function and learn the parameters so that the highest scoring outputs match as closely as possible the desired outputs. Prior maximum margin approaches produced QP models [11]. By thinking of the problem one level up as a convex concave saddle point model, the authors can capitalize on the recent advances in optimization on extragradient methods [10]. The extragradient approach produces a simple algorithm consisting of a gradient and projection step. For the class of models considered, the projection requires solution of dynamic program or network flow models for which very efficient algorithms exist. The method is regularized by early stopping. Interestingly the path of the extragradient algorithm corresponds closely to the parametric solution path of the regularized margin methods in their experiments.

4. IMPORTANCE OF OPTIMIZATION IN ML

The most important optimization algorithms currently are those that can be used to solve constrained non-linear, non-smooth large-scale optimization problems as these challenging problems are of increasing importance in modern ML.These are mainly first-order (i.e. gradient-based) methods that solve a relaxed problem such as the augmented Lagrangianfunctions; some of these methods can be seen from a dual perspective:

Alternating Directions Methods/Coordinate Descent with the Augmented Lagrange Multipliers methodsproximal gradient methods such as ISTA and more importantly the (Nesterov) accelerated versions such as FISTA. Some other gradient-based algorithms are also extremely important, the one that comes to mind first is stochastic gradient descent and its applications in neural networks training and gradient boosted trees (see Gradient Boosted Machines) [13].

As others have said, Bayesian optimization and other global optimization techniques are also very important. They can be used for optimizing difficult objective functions, for instance when optimizing the hyper-parameters of a model.

Desirable properties of an optimization algorithm from the ML perspective are-

- Good generalization
- Scalability to large problems
- Good performance in practice in terms of execution times and memory requirements,
- Simple and easy implementation of algorithm,
- Exploitation of problem structure • fast convergence to an approximate solution of model,
- Robustness and numerical stability for class of machine learning models attempted,
- Theoretically known convergence and complexity.

5. FUTURE WORKS & CONCLUSIONS

Research in ML and research in MP have become increasingly coupled. ML researchers are making fuller use of the branches of the MP modeling tree. In this issue we see MP researchers using convex optimization methods including linear, nonlinear, saddle point, semi-infinite, second order cone, and semi-definite programming models. The availability of general MP models, along with robust general purpose solvers; provide tools for ML researchers to explore new ML problems. The resulting ML models challenge the capacity of general purpose solvers resulting in the development of novel special purpose algorithms that exploit problem structure.

This paper covers various Optimisationmethods used in Machine learning. In future work we will presents an efficient optimization method for machine learning and compare its results with various existing methods.

REFERENCES

1. M. *Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In Advances in Neural Information Processing Systems 25. 2010.*
2. *E. Crosson and A. W. Harrow. Simulated quantum annealing can be exponentially faster than classical simulated annealing. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 714–723, Oct 2016.*
3. *Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.*
4. *Andrew Saxe, Pang Wei Koh, Zhenghao Chen, ManeeshBhand, Bipin Suresh, and Andrew Ng. On random weights and unsupervised feature learning. In Proceedings of the 28th International Conference on Machine Learning, 2011.*

5. *Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009.*

6. *Adam Coates and Andrew Y. Ng. Selecting receptive fields in deep networks. In Advances in Neural Information Processing Systems 25. 2011.*

7. *Cornell University Library. "Breiman: Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". Retrieved 8 August 2015.*

8. *Dan ClaudiuCiresan, Ueli Meier, and JurgenSchmidhuber. Multi-column deep neural net- ¨works for image classification. In Computer Vision and Pattern Recognition, 2012.*

9. *Zhang, Jun; Zhan, Zhi-hui; Lin, Ying; Chen, Ni; Gong, Yue-jiao; Zhong, Jing-hui; Chung, Henry S.H.; Li, Yun; Shi, Yu-hui (2011). "Evolutionary Computation Meets Machine Learning: A Survey" (PDF). Computational Intelligence Magazine. 6 (4): 68–75.*

10. *Sarikaya, Ruhi, Geoffrey E. Hinton, and AnoopDeoras. "Application of deep belief networks for natural language understanding." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 22.4 (2014): 778–784.*

11. *Tillmann, A. M. (2015). "On the Computational Intractability of Exact and Approximate Dictionary Learning". IEEE Signal Processing Letters. 22 (1): 45–49*

12. *Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207.*

13. *Bose I. and Mahapatra R. K. (2001). "Business data mining a machine learning perspective", Information & Management, Vol. 39(3), pp. 211-225.*

14. *Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, May 26-28, 1993, 207-216.*

15. *Agrawal R, Srikant R. Mining sequential patterns. In Proceedings of the 11th IEEE ICDE International Conference on Data Engineering, pages 3-14, 1995.*

16. *Ma Q, Wang J T L. Biological data mining using Bayesian neural networks: A case study. International Journal on Artificial Intelligence Tools, Special Issue on Biocomputing, 1999, 8(4), 433-451.*

17. *Hirsh H, Noordewier M. Using background knowledge to improve inductive learning of DNA sequences. Proceedings of the 10th IEEE Conference on Artificial Intelligence for Applications, 2014, 351-357.*