Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

Authors: John Duchi, Elad Hazan, Yoram Singer
Year: 2011
Journal: Journal of Machine Learning

Report:

1. **Abstract:**

"We present a new family of subgradient methods that dynamically incorporate knowledge of the geometry of the data observed in earlier iterations to perform more informative gradient-based learning. The adaptation, in essence, allows us to find needles in haystacks in the form of very predictive yet rarely observed features. Our paradigm stems from recent advances in online learning which employ proximal functions to control the gradient steps of the algorithm. We describe and analyze an apparatus for adaptively modifying the proximal function, which significantly simplifies the task of setting a learning rate and results in regret guarantees that are probably as good as the best proximal function that can be chosen in hindsight. We corroborate our theoretical results with experiments on a text classification task, showing substantial improvements for classification with sparse datasets."

Introduction

- The article tries to fulfill the acquirement of strong regret guarantees, which they can apply to achieve data distributions with higher performance.
- For this purpose, the control of algorithm's gradient steps is implemented by the proximal functions.
- The authors also endeavoured to achieve the predictive and frequently observed features of the data.
- This is much required in the data learning, which would increase the accuracy of the output with the decrease in computational expense.

Previous Works

- The works on Quasi-Newton stochastic gradient descent had been done former to the paper, which the article had used as base to the Adaptive Gradient Algorithm.
- But the assumptions in the former works had been very constructive which would not result more in the general cases.
- And the most of the previous literature had given the learning rate manually, which the authors in contrast created proximal function to adapt itself.

- Assigning all the frequent and infrequent features with same learning rate will result in more imprecise data. So, learning rates has to be differed based on the frequency of the features. This is one of the which had not been focused before.
- More works were also concentrated on loss functions and error thresholds. Whereas the authors overcome it through proximal function auto adaptability and considering the best bound learning rate.
- This will considerably increase the proximity of the output of the data.

Content

- It is obvious that the instances of the data will be having very high dimensional features, which are frequent.
- As mentioned in previous works, to adapt the learning rate based on the frequent features to be less, they have implemented an algorithm called as Adaptive Gradient Algorithm(ADAGRAD).
- This will result in the dynamic nature of the algorithm. The algorithm follows some of the following theory with complex mathematics.
- Initially, they have defined the function to calculate the regret value of the given convex curve.
- They have iterated over the function with increase in time to update the regret value.
- Moreover, they have declared some of the critical proofs, which they have declared to be working in the algorithms.
- The proofs define the relations between the regret variables, convex functions and the step size etc.,
- Then they have taken the optimal regret values to assign the learning rate of the data.
- The strong convex functions will result in the higher regret functions, so they even have created the mathematical function to lower the regret functions to strong convex functions.
- The authors have compared the two algorithms diagonal adaptive descent and the online gradient descent, in which both resulted a unit loss in initial some examples.
- After the considerable number of examples the online gradient descent seems to be stopped suffering from losses.
- A diagonal matrix proximal function is considered, in which it serves two major purposes.
    - In general cases, the evaluation of the function is complicated. So, the diagonal case serves for better analyzing.
    - Secondly, the diagonal proximal function decreases the computational expensiveness in higher dimensions.
- They also focused on the limiting sparsing constraints, which affect the learning algorithm effectiveness.

Results

- The performance is determined by two major characteristics:
    - Online loss or error
    - Test set performance of predictor after the single pass of training data
- They have compared around six algorithms along with their ADAGRAD, in which the ADAGRAD outperformed the remaining algorithms in the majority of the data sets.

- ADAGRAD and AROW, the adaptive algorithms has resulted in comparatively less error rates than the other algorithms like RDA, FB and PA.
- The experiments were done on four data sets, in which all the sets had lower error rate in Adaptive algorithms only.

Discussion and thoughts

- Proximal functions are the functions used to resolve non-differentiable convex shaped optimization problems. This is the basic of the article, which authors assumed to have a prior knowledge of that.
- There is no theory to estimate the efficiency of the full matrices in the proximal functions, which they had eft for future work.
- The authors have also mentioned that there is a chance to use a new algorithm instead of carrying calculations over the matrices of functions.
- The possibility to implement the KL divergence between the distributions of features of the data is also left to future work.

Conclusion

- On the whole the paper had given a paradigm which could adapt the sub gradient algorithms and derive the learning rates for the problem.
- The algorithm also successfully decreased the error rates and made the online learning had become much precise.
- Overall with the more critical mathematics and the adaptivity of proximity functions had lead to the flexibility of adaptive algorithm.
- Now this created easiness to convert the online convergence results into the convergence rate and bounds of generalization.