Devamsh Kondragunta (203125890)

Professor Virginia Mushkatblat

COMP 680 Advanced Topics SWE

March 30, 2023.

**Effectiveness metrics for using ML and other AI methods**

**Abstract:**

Machine learning optimization approaches are facing more and more difficulties as a result of the exponential expansion in data volume and the rise in model complexity. Many efforts have been made over time to solve optimization issues or enhance optimization techniques in machine learning. Metrics are quantitative metrics that aid in calculating progress toward particular aims or goals. These indicators can be used to monitor performance over time, pinpoint problem areas, and make fact-based choices in developing an AI model. We provide an overview of the advancements and other applications of optimization metric techniques in a few well-known machine learning domains.

**Introduction:**

To assess the usefulness of Machine Learning (ML) and other Artificial Intelligence (AI) approaches in various applications, their performance must be evaluated. These models' performance is often evaluated using a variety of metrics that are selected based on the features of the dataset and the particular application. It is crucial to analyze and summarize optimization techniques methodically from a machine learning standpoint since this can provide direction for future work in both machine learning and AI optimization. ​​In order to guarantee the dependability of AI systems, let’s characterize the techniques that are applied in real-world situations.

**Industry Overview:**

Due to its capacity to automate processes and enhance decision-making, machine learning (ML) and other AI techniques have grown in popularity in recent years. Several industries, including banking, healthcare, and marketing, use these techniques. Yet as the usage of these techniques spreads, it's critical to provide measures that can be used to judge how beneficial they are. The most direct method for system users to express their wants and priorities to system designers is by defining the metrics that will be used to measure their performance. We must be aware of any general constraints of this experimental design and be able to relate them to potential dangers if we utilize measurement of our intended metrics on the test dataset to infer the characteristics of the system under deployment.

**First Approach:**

We put together a set of test data for the system to be encountered upon deployment. The metrics are measured on this dataset in order to estimate their genuine values. A model's exactness or Accuracy tells us how well it can estimate the proper result. It is calculated by partitioning the entire number of surmises by the number of forecasts that were adjusted. (Lalli Myllyaho and Mikko Raatikainen, 16). It can be calculated by separating the entire number of anticipated positives by the number of real positives. As an illustration, the system's processing timings for each datapoint in the test data might be timestamped, and the average of those processing times could be used to project how long it will typically take the system to process each datapoint once it is deployed. (Ish et al., 12)

**Other Approaches:**

Recall, F1 score, Mean Square Error and ROC curve analysis are some of the other measures that are frequently used to assess how well ML and AI algorithms work. The value of the measure should be anticipated to change more between distinct sets of data that are smaller than the data sets. The F1 score, which balances the trade-off between precision and recall. IROC curve analysis is a technique used to compare the true positive rate and false positive rate for various threshold settings. It provides a visual representation of the model's performance and can be used to select the most appropriate threshold for classifying successful outcomes. However ROC curve analysis could potentially be misleading if the dataset is imbalanced.

The AUPRC assesses the model's capability to identify good outcomes across all conceivable criteria. Since it is less affected by the ratio of positive to negative findings, it is particularly useful in datasets containing imbalances. (“(PDF) Predictive analysis using machine learning: Review of trends and methods”, 7).

A correlation coefficient known as the MCC takes into account all four findings (true positive, true negative, false positive, and false negative). It is a trustworthy statistic that is especially useful in unbalanced datasets since it is unaffected by the ratio of positive to negative outcomes.

**Comparison and Discussion:**

Although ML systems are expected to perform better on the test data than on data not used in training, it is also crucial to make sure that the test data do not reuse any data from the training process. Also, there is a significant difference between a metric's sample value and actual value. The actual value is the value that would result from an infinite quantity of test data, whereas the sample value is determined using the test data. (“(PDF) Predictive analysis using machine learning: Review of trends and methods”, 14).

Accuracy can be deceiving in datasets with imbalances, and metrics like AUPRC, balanced accuracy, and MCC may offer a more accurate assessment of the model's performance. It is crucial to remember that the quality of the data and the model's design are just as critical to an ML or AI model's performance as the selection of assessment measures.

While each metric has advantages and disadvantages of its own, To provide a more complete picture of the model's strengths and flaws, we can combine several metrics to get a more thorough assessment of the model's performance. Some metrics may be more significant than others depending on the application. The properties of the dataset should be considered when choosing the assessment metric. (Shiliang Sun, and Zehui Cao, 9).

**Conclusion:**

Certain metrics may be more suitable than others depending on the application and the dataset's properties and model design because they represent just one part of the model's total performance. To make sure that the model works as intended, it is crucial to carefully assess its performance using a variety of indicators.

The selection of evaluation criteria for ML and AI methods should be based on the particulars like types of the application and dataset. Since training a system inevitably affects its performance, evaluating systems that continue training after deployment, such as recommenders that modify their recommendations based on user interactions.

Proper data preparation and model construction should go hand in hand with the evaluation of these models' performance.To conclude, there is no statistical metric that can provide a comprehensive assessment of the effectiveness of a machine learning or artificial intelligence model.

**References:**

Ish, Daniel, et al. “Evaluating the Effectiveness of Artificial Intelligence Systems in Intelligence Analysis | RAND.” *RAND Corporation*, 2021

<https://www.rand.org/pubs/research_reports/RRA464-1.html.>

“(PDF) Predictive analysis using machine learning: Review of trends and methods.” *ResearchGate*, November 2020 <https://www.researchgate.net/publication/346266253_Predictive_analysis_using_machine_learning_Review_of_trends_and_methods.>

s

Shiliang Sun, and Zehui Cao. “A Survey of Optimization Methods from a Machine Learning Perspective.” 10 May 2019, <https://arxiv.org/pdf/1906.06821.pdf.>

Lalli Myllyaho, and Mikko Raatikainen,. “Systematic literature review of validation methods for AI systems.” *Science Direct*, 23 July 2021, <https://reader.elsevier.com/reader/sd/pii/S0164121221001473?token=DD2D5BF9233650AF109A86014778805B9B12CF4FB7CE96827A1637B0CF101E0250033EE34383FC56A1B2018E01B3CF09&originRegion=us-east-1&originCreation=20230331041250.>