International Journal of Management, IT & Engineering
Vol. 14 Issue 01, January 2024
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Measuring and increasing the quality of Human Labels

Deepanjan Kundu[**]

Keywords:

Machine Learning;
Artificial Intelligence;
Human Labels;
Golden Data;
Label Quality

Abstract

In this article, we look at the fundamental issue of label quality in machine learning (ML) projects, focusing on supervised learning where human-generated labels are considered as ground truth. The quality of these labels directly impacts the performance and reliability of ML models. This study explores the methods and challenges involved in ensuring high label quality. It demonstrates the concept of inter-annotator disagreement as a crucial metric for assessing label consistency and discusses statistical measures like Fleiss' kappa for a more nuanced evaluation. The article emphasizes on the role of "Golden Data" in setting benchmarks for label accuracy and consistency. Additionally, it presents various strategies to improve label quality, such as the implementation of effective tracking systems, the refinement of labeling instructions, and adherence to best practices in data labeling. Strong correlation between label quality and the overall success of ML models, underscores the importance of rigorous data labeling processes. By providing insights into practical measures for enhancing label quality, the article contributes significantly to the field of machine learning, offering guidelines and a framework that can be adopted by data scientists and ML practitioners to ensure the collection of reliable and accurate data for their models.

Author correspondence:

Deepanjan Kundu,
Independent Researcher,
Applied Machine Learning Expert
USA
Email:
kundu.deepanjan@gmail.com

1. Introduction

In machine learning, a label is a categorical or numerical value assigned to a data point that serves as the target variable for a predictive model. The process of labeling involves manually or automatically assigning these values to the training data, which is then used to train a machine learning algorithm to predict the labels of unseen data. Labels can be binary, multi-class, or continuous values, depending on the type of problem and the nature of the data. Most of the content-based end tasks in applied ML use human labeling programs to collect their labels. We will be looking into human labels in this study.

Labels are the bread and butter of supervised machine learning problems. When one looks at a supervised machine learning problem, they look at the data set and assume that label is the ground truth. But how do you ensure that human labels reflect the ground truth? These ML models are delivered to clients or end users and hence hold critical value. Data scientists need to be sure that the labels are of good quality and are a

[*] Author has 7 years of experience in applied machine learning, and the AI industry mainly focused on text classification, ranking and personalization. Currently the author is working as a Senior Software Engineer in Palo Alto, USA.

reflection of the problem they are solving. Accurately labeling data is a crucial step in the machine learning pipeline as it directly impacts the performance and generalization ability of the model. In this study, we will explore and compare different methods to evaluate quality of human labels, and propose best practices to increase and maintain label quality.

2. How can we measure label quality?

2.1 Inter annotator disagreement

It is considered best practice to have human labelers provide multiple labels for each data point. Different human labelers provide labels for the same data point, and then the final label is aggregated from each human label. If all the labelers provide the same response for a given data point, the confidence in the label for such data points would be high. But as is common in most of the tasks, a significant portion of the data points have labeler disagreement. This can be measured using inter annotator disagreement or labeler disagreement, which assesses the extent to which multiple human labelers disagree on the labels assigned to the same data point.

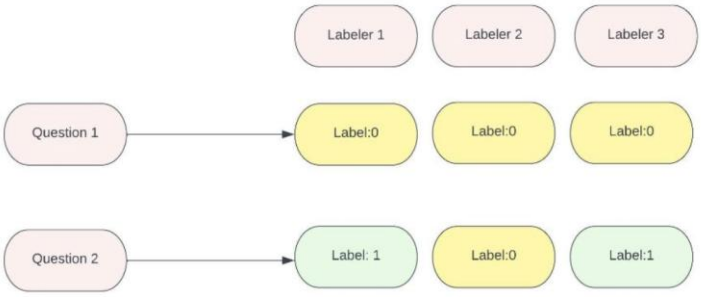


Figure 1. Different human labelers provide different labels for same data point

We will be using the term “labeler disagreement” for the purposes of our study. This also acts as an indication of the quality of the labels collected. Suppose the labeler disagreement is high for a large number of data points, which indicates a higher chance of collecting lower-quality labels. Here are a few methods by which the labeler disagreement rate is measured.

2.1.a. % of data points with disagreement

For each data point, check if each label response from all human raters is exactly the same. In this method, disagreement would be represented by the ratio of the number of data points that have any disagreement and the total number of data In points. Let's look at an example below:

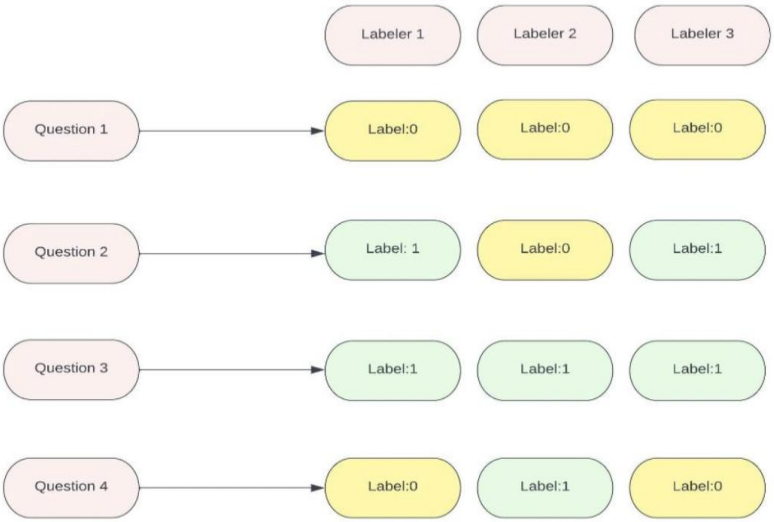


Figure 2. In this example, disagreement Percentage = 50% as Question 2 and Question 4 have disagreement

2.1.b. Fleiss' kappa

Fleiss' kappa [1] is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. This helps to incorporate the relative weight of scenarios with partial agreement along with partial disagreement among labelers. If the raters are in complete agreement, then the Kappa is 1. If there is no agreement among the raters (other than what would be expected by chance), then Kappa ≤ 0.

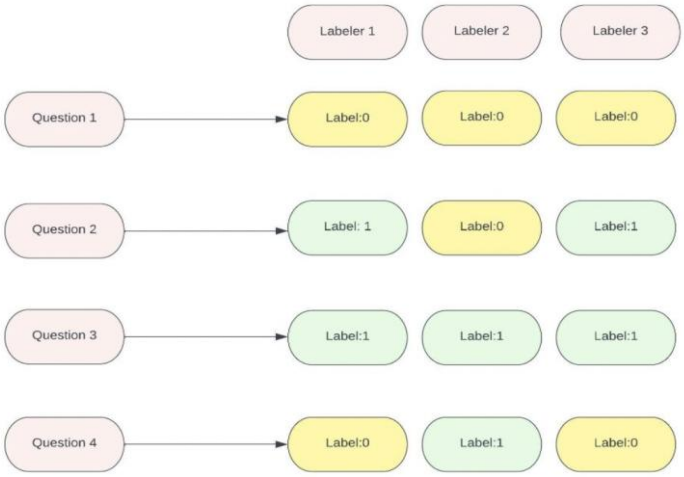
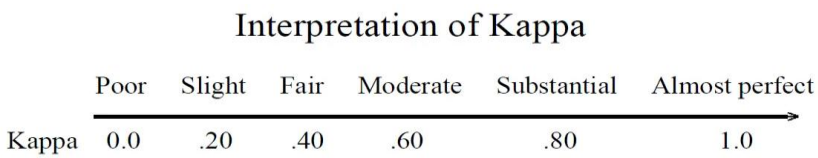


Figure 3. In this example, Fleiss' Kappa = 0.33



Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Figure 4. Interpreting Fleiss' kappa [2]

2.2 Golden Data

Golden Data is a collection of data points with labels that act as a definitive reference for the problem/labeling task at hand. For the use cases of content-based tasks, golden data is usually a collection of data that is manually curated with highly precise labels. They aim to cover the different dimensions of the problem and multiple edge cases. The goal is not to be a reflection of true distribution but to cover obvious and borderline cases across all the labels for the task. The size of these datasets is usually in the low hundreds and is fairly static.

The way to measure label quality is to intermittently mix a subset of the golden dataset with each batch of labeling tasks. The aim would be to have about 50 of these golden data points in each batch of dataset to be sent for labeling. If the labels are collected daily, then the golden data could be mixed once a week. A lower frequency of golden data is important to ensure golden data points are not leaked to the raters. Here is a diagram demonstrating the mixing process:

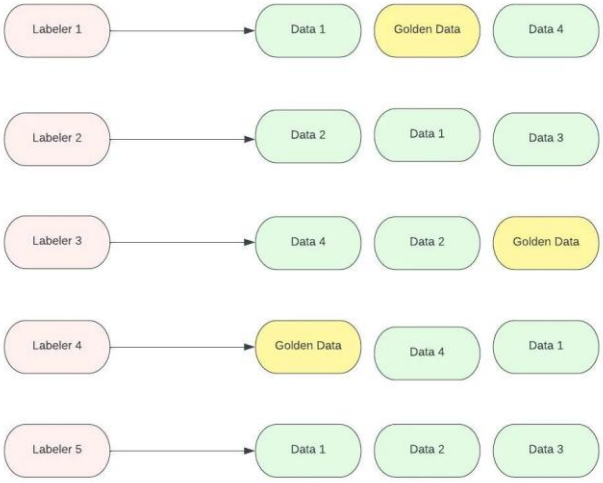


Figure 5. This figure shows how the golden data is mixed with existing data.

The example is a scenario in which 5 labelers were used to get 3 labels for a total of 5 data points, one of which was a golden data point. The accuracy [4], precision[4,5], and recall[5] of the labels from the human raters vs. the "golden labels" will be the key metrics to measure label quality.

3. How can we increase label quality?

3.1 Track the quality

It is important to track label quality and be alerted when the label quality drops below a certain threshold. Each of the measures mentioned in the above sections has its own pros and cons, and it would be advisable to track all of them. In an example in Figure 5, we look at the Fleiss' Kappa measured week over week for the labels collected. For the weeks 2, 6, 7, 9, 11, 12 and 13 the kappa falls below zero highlighting intermittent irregularities in the labeling process and the need for immediate intervention.

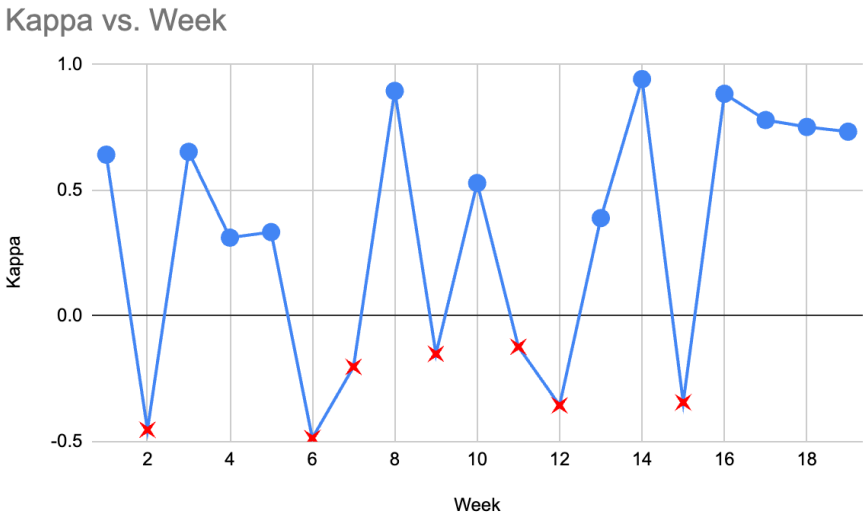


Figure 6. Tracking Fleiss' Kappa for Label Quality

3.2 Reiterate on labeling instructions

Labeling instructions should be the usual suspect when irregularities in label quality are detected. It is important to identify if there are any holes in the template. Two key methods using which you can identify

and fill those gaps are the following: One, the accuracy of the labels collected using the instructions on golden data should be high (>95%). The way to ensure this is to run pilot programs with the initial versions of the rating template on Golden Data and tune the rating template till you achieve the desired accuracy of human labeling. Second, for batches that have high labeler disagreement, go through the data points with labeler disagreement and manually identify gaps in the instructions for the data points that might be leading to confusion for labelers.

3.3 Best Practices for labeling process

There are a few recommended practices to follow while collecting labels for your data that improve the label quality:

- **Replication**: A standard practice is to collect multiple labels for the same data to ensure better quality labels. An aggregate of these labels is used as the final label. For classification, the median of the labels collected could be used, and for regression, the average of the labels could be used.

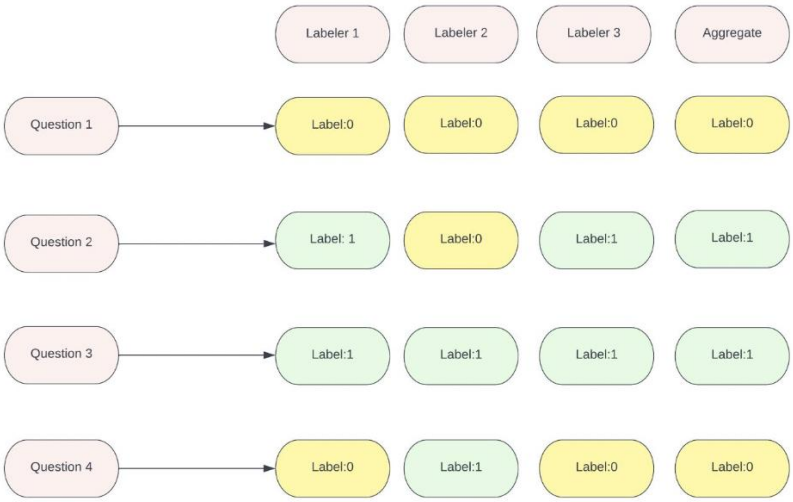


Figure 7. Median aggregation of labels for Classification task into classes 1 and 0

- **Limit the number of tasks per labeler:** It is important to minimize human bias in the labels being collected. If all the tasks are executed by the same labeler, there is a higher chance of human bias in the data. Limiting the number of tasks per labeler is really important. A simple way to achieve this is by increasing the number of distinct labelers. You can use Golden Data to identify good vs. bad labelers and assign more tasks to good labelers and fewer tasks to bad labelers.

- **Appropriate amount of time for the labelers:** It is important to measure the amount of time taken by the labelers to optimize the balance between label accuracy and costs. You should run multiple experiments to ensure labelers have enough time to read the rules and label the data. The only guidance here is the quality of the labels. The minimum amount of time required to achieve the desired level of accuracy and labeler disagreement should be the goal. If the average time taken by the labelers is closer to the total time provided, it does not necessarily indicate efficiency. The labelers could be wasting time, and it would seem that they are using up the entire time. It is important to track the accuracy and labeler disagreement.

4. Conclusion

In the realm of machine learning, data labeling is often underestimated, yet it plays a pivotal role in the success of ML models. This article provides insightful methodologies for assessing and enhancing the quality of human labels in ML tasks. Key methods for measuring label quality include monitoring labeler disagreement and employing statistical tools like Fleiss' kappa. The concept of "Golden Data" is introduced as a standard against which label quality is evaluated. We emphasize that improving label quality is a multifaceted process. It involves diligent tracking of label quality, revisiting and refining labeling instructions, and adhering to best practices in the data labeling process. These practices ensure the collection of high-quality data, which is fundamental for the development of effective and reliable ML models. By prioritizing label quality, data scientists and ML practitioners can devote more resources to designing and building innovative ML models, secure in the knowledge that the underlying data is of the highest quality.

Overall, the article underscores the significance of label quality as the cornerstone of successful machine learning projects.

References

[1] McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276–282.

[2] Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Family medicine, 37(5), 360–363.

[3] Plank, B. (2022). The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Conference on Empirical Methods in Natural Language Processing.

[4] Wikipedia contributors. (2023, September 26). Accuracy and precision. In Wikipedia, The Free Encyclopedia. Retrieved 09:05, November 28, 2023, from https://en.wikipedia.org/w/index.php?title=Accuracy_and_precision&oldid=1177199912

[5] Wikipedia contributors. (2023, November 27). Precision and recall. In Wikipedia, The Free Encyclopedia. Retrieved 09:04, November 28, 2023, from https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=1187084495