Weakly Supervised LTR

Vinayak Siva Kumar
Virginia Tech
vinayaksk@vt.edu

Abstract

Weak supervision falls under an area of machine learning where labels are generated from unorganized,imprecise or noisy data. This labelled data can then be treated as a supervised learning task. This project examines examines a popular weak supervision framework and how it can be utilized in the realm of information retrieval, specifically in the fields of Learning-to-Rank and Recommendation systems.

1 Introduction

As the applications of machine learning and deep learning models have been becoming quite apparent in recent times, a key bottleneck is obtaining enough high-quality training data for these models. This is especially true when it comes to Human-annotated data. Generating and maintaining these types of datasets can be have high time and computational costs.

Apart from the mentioned costs, getting the expert subject-matter expertise on labelling the data might not always be possible in every situation, hence turning to weak and noisy labels can help tackle this issue. We might have some subject-matter to get a bit of relevant label information, but obtaining perfect human annotated data for all records in a dataset can become too time consuming and impractical in a lot of situations.

This is where weak supervision comes into play. Weak supervision involves methods that can assign noisy training labels to unlabeled data, through techniques such as using machine learning classifiers, crowd-sourcing and user-defined heuristics. The datasets chosen for weak supervision are usually large in size to make up for the low quality labels.

This project looks at a framework which allows users to generate their own labels to an unlabelled dataset using their own heuristics [2]. The system consists of the following steps: 1. Apply the labeling functions to unlabeled data. 2. Use a generative model to learn the accuracy of the labeling functions without any labeled data, and weight their outputs accordingly.3. Learn the structure of their correlations automatically. 4. The generative model outputs a set of probabilistic training labels, which we can use to train a powerful, flexible discriminative model (such as a deep neural network) that will generalize beyond the signal expressed in our labeling functions.

We will use this framework to generate our own ratings in a recommendation system. This type of system is applicable to also Learning-To-Rank problems when generating an ordered list of ratings of items for a user based on the rating scores.

2 Literature Survey

The paper in [1] presents an approach of using an unsupervised ranking model, such as BM25, as a weak supervision signal. The output from this model is used to further train a set of selective ranking models based on feed-forward neural networks. The paper checks various learning scenarios such as point-wise and pair-wise LTR models and uses different input representations as well (dense/sparse encoded vectors and word embeddings).

A framework called Snorkel is introduced in [2] which this project solely focuses on. This system allows users to generate their own weak labels using a custom labelling interface, a context hierarchical model and a generative model that gives probabilistic training data in those user-defined labels. This framework has the advantage of being easy to use in different situations and allows users to make their own custom labels fairly easily without having an in-dept knowledge of machine learning or programming. These heuristics can be then decided by domain experts based on the features of a dataset and be easily be generated using simple functions and programming.

A very similar framework called SNUBA [3] allows users to generate their own heuristics and label noisy/incomplete data with their own labels. This framework consists of a synthesizer step, which generates the labels based on the heuristics given if the heuristics (either through threshold functions,supervised or unsupervised signals) provide decent accuracy, a pruning step which chooses labels based on metrics like precision and recall, and a verifier that gives a single, probabilistic training label for each datapoint. The heuristics in this framework appear to be slightly more limited compared to the Snorkel framework mentioned above.

This system [4] looks at aggregating multiple weak supervision sources and obtaining the "latent true label", without knowing ground truth data. Unlike the prior approaches mentioned, here the downstream model is part of the framework and it allows for directly learning the downstream model by enabling an agreement with the probabilistic (pseudo/weak) labels previously generated and allows for reparameterizing prior probabilistic posteriors based on a noise-aware loss. Hence the labels are calculated based on how well the downstream model is effective with them.

3 Books Recommendations Using Weak Supervision

3.1 Overview of Snorkel

The Snorkel framework, instead of using hand-labeling training data, allows users to write labeling functions, which allow them to express various weak supervision sources such as patterns, heuristics and external knowledge bases.

Snorkel automatically learns a generative model over the labeling functions, which allows it to estimate their accuracy and correlations. This step uses no ground-truth data, learning instead from the agreements and disagreements of the labeling functions.

The output of Snorkel is a set of probabilistic labels and these labels can be used to train a powerful discriminative classifier with a large feature set that generalizes beyond the reasons directly addressed by the labelling functions. The generative model is encoded using three factors representing the labeling propensity,accuracy and pairwise correlations of labeling functions.

3.2 Dataset

This project looks at utilizing the Snorkel framework to generate ratings to give relevant book recommendations. The dataset [5] [6] consists of the user id details, books each user has read and ratings associated with them. This database also consists of reviews as well. However, only a small sample of data contains book ratings and reviews, and the rest of the data is sparse. Our objective is to predict whether a given user (represented by the set of book id's the user has interacted with) will read and like any given book. That is, we want to train a model that takes a set of book id's (the user) and a single book id (the book to rate) and predicts the rating.

3.3 Labelling Heuristics

We use a couple of simple labelling functions such as designing a label that gives a good rating if users read the same books and use a preexisting sentiment analyzer to generate polarity and subjectivity scores to predict user ratings.

These pseudo labels we use to label the sparse data is converted to probabilistic values as a result of Snorkel.

3.4 Neural Network Model Classifier

The probabilistic values is then fed into a classifier. A Neural Network is chosen for this task. The model represents the list of books the user interacted with, books ids, by learning an embedding for each id, and averaging the embeddings in book ids. There is an additional embedding for the the book to be rated. Then it concatenates the two embeddings and uses a multilayer perceptron (MLP) network to compute the probability of the rating being 1.

4 Results

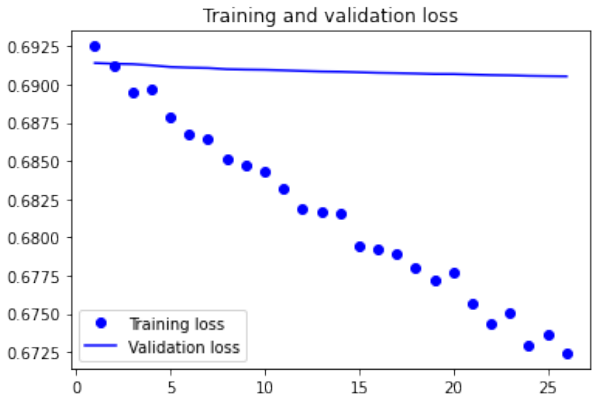

Figure 1: Training and Validation Loss

The training loss and validation loss can be seen in Figure 1. The validation loss remains constant among the epochs which might be an issue which needs to be looked at.

Accuracy	Precision	Recall
0.6484	0.3006	0.1889

Table 1: Evaluation Metrics

The evaluation metrics are given in Table 1. As seen, the accuracy lies at 0.6484 but the precision and recall are noticeable lower. This can be attributed to a lot of things such as the heuristics chosen may not be the most ideal. There is also an important aspect of weak supervision to consider, and that is the general performance of these models. Weak supervised models may not always outperform models that run on high quality datasets. It's generally an approach that people consider when the resources to process and train the data are limited. Another factor that affects the performance of the model is the dataset. For weak supervision, it is generally expected to train on a large dataset(since the data is sparse, the model needs to train on a lot more data points to get a better context).

5 Project Outcome and Further Studies

While we have been able to explore a bit in weak supervision, we still could have done a lot more and compare other approaches(such as using BM25 as a label generator and see how it compares to the above approach), and take on a more conventional LTR problemset such as a web query based one. However, research of weak supervision (especially in LTR) is a bit limited as most existing research is tailored towards text classification in the information retrieval space. Some other interesting research is this domain include Named Entity Recognition with weak supervision.

6 Conclusion

We have explored the domain of weak supervision and have seen how a popular framework that utilizes weak supervision handles in a machine learning problem. With these predicted ratings

calculated for the users, top recommendations can then be generated to the user based on the highest probability scores given in the label.

Large labeled datasets are generally important to many machine learning applications. Reducing the human effort manually label such datasets is an important step towards making machine learning more accessible. Some of the issues which may arise has been briefly discussed earlier, such as the bias or errors that may come when defining the heuristics of the system.

References

[1] Mostafa D, Hamed Z, Aliaksei S, Jaap K, and W. B. Croft. 2017. Neural Ranking Models with Weak Supervision. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17). Association for Computing Machinery, New York, NY, USA, 65–74. doi:10.1145/3077136.3080832

[2] Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid Training Data Creation with Weak Supervision. Proceedings VLDB Endowment. 2017;11(3):269-282. doi:10.14778/3157794.3157797

[3] Paroma C and Christopher R. 2018. Snuba: automating weak supervision to label training data. Proc. VLDB Endow. 12, 3 (November 2018), 223–236. DOI:https://doi.org/10.14778/3291264.3291268

[4] Rühling C, Salva Boecking, Benedikt Dubrawski, Artur. (2021). End-to-End Weak Supervision.

[5] Mengting W, Julian McAuley, "Item Recommendation on Monotonic Behavior Chains", in RecSys'18

[6] Mengting W, Rishabh M, Ndapa N, Julian McAuley, "Fine-Grained Spoiler Detection from Large-Scale Review Corpora", in ACL'19.