cse-584-HW1

Jiamu Bai

September 2024

1 Paper 1: Bayesian Active Learning for Classification and Preference Learning

1.1 What problem does this paper try to solve?

The paper addresses the challenge of active learning in probabilistic models, particularly in complex tasks like classification with nonparametric models such as Gaussian Process Classifier (GPC). Traditional active learning methods can be computationally intensive and rely on approximations that can reduce their effectiveness. The motivation is to develop a more efficient active learning algorithm that makes minimal approximations while providing optimal information gain in learning with the least amount of data.

1.2 How does it solve the problem?

The authors propose an approach that expresses information gain in terms of predictive entropies. They apply this method to the GPC, allowing for more accurate active learning by maximizing the expected reduction in uncertainty about the model parameters. The algorithm leverages the mutual information between the unknown output and model parameters, computing entropies in the output space, which simplifies calculations. This is referred to as Bayesian Active Learning by Disagreement (BALD). They extend this method to preference learning, transforming it into a classification problem and applying their active learning framework to it.

1.3 A list of novelties/contributions

- The paper presents an active learning algorithm that makes fewer approximations than previous methods, making it the most exact and fastest way to perform full information-theoretic active learning in non-parametric models.

- They develop a novel application of information-theoretic active learning to Gaussian Process Classification, which is challenging due to its infinite-dimensional parameter space.

- The paper extends the algorithm to preference learning, a complex task where data consists of pairs of items with binary labels indicating preference.

1.4 Downsides of the Work

Although the method reduces the computational demands compared to other approaches, it still requires complex calculations and might not scale well to very large datasets.

2 Paper 2: Deep Bayesian Active Learning with Image Data

2.1 What problem does this paper try to solve?

The paper addresses the challenges of applying active learning to deep learning, particularly with high-dimensional data like images. Traditional active learning methods often require models to learn from small amounts of data, but deep learning models usually rely on large datasets and do not natively represent model uncertainty, which is crucial for active learning. The motivation is to integrate Bayesian approaches into deep learning to create an effective active learning framework for image data, which has been challenging due to the complexity and high dimensionality of images.

2.2 How does it solve the problem?

The paper proposes using Bayesian Convolutional Neural Networks (BCNNs) to represent model uncertainty in deep learning. These networks utilize dropout as a form of approximate Bayesian inference, allowing them to capture uncertainty in predictions. The framework employs acquisition functions that leverage the uncertainty information provided by the BCNNs to select the most informative data points for labeling. The paper tests this approach on image datasets, including MNIST and a skin cancer diagnosis task.

2.3 A list of novelties/contributions

- The paper combines Bayesian deep learning with active learning to handle high-dimensional data, specifically using Bayesian CNNs for image data.

- It develops a practical framework that uses model uncertainty for effective data selection, leading to significant improvements in labeling efficiency.

2.4 Downsides of the Work

The approach involves Bayesian inference, which can be computationally intensive. Training and performing MC dropout multiple times to capture uncertainty might not be feasible for very large datasets or models with limited computational resources.

3 Paper 3: Batch Active Learning at Scale

3.1 What problem does this paper try to solve?

The paper addresses the challenge of efficiently training complex machine learning models, particularly deep neural networks, which require large amounts of labeled data. Acquiring this data can be costly and time-consuming. The paper focuses on batch active learning, which involves issuing large batches of queries to a labeling oracle to mitigate the costs associated with frequent model retraining and labeling iterations. The motivation is to develop an active learning algorithm that scales to large batch sizes, ensuring the model is trained efficiently with minimal redundancy in the selected samples.

3.2 How does it solve the problem?

The authors propose the "Cluster-Margin" algorithm, which combines uncertainty and diversity in sampling. The algorithm uses Hierarchical Agglomerative Clustering (HAC) to preprocess and diversify batches of examples where the model shows the least confidence. Cluster-Margin initializes by training on a small seed set, clusters the entire dataset using HAC, and then iteratively selects batches of the least confident examples, ensuring diversity by sampling across clusters. This method requires only one clustering step as preprocessing, making it computationally efficient.

3.3 A list of novelties/contributions

- Cluster-Margin is designed to handle very large batch sizes (100K to 1M), which is orders of magnitude larger than previous studies in active learning.
- It leverages HAC for clustering only once during preprocessing, making it more efficient compared to other methods that require clustering or diversification at each iteration.

3.4 Downsides of the Work

Batch active learning, by nature, has reduced adaptivity compared to sequential active learning, which might lead to sampling redundant or less informative examples within a batch, especially for very large batch sizes.