Machine Unlearning with Variational Inference

February 14, 2021

Ali Faraz	alifaraz@iitk.ac.in	17807078
Jayant Ranwka	jayant@iitk.ac.in	170323
Shashank Kumar	shankk@iitk.ac.in	170652
Vishweshwar Tyagi	tyagi@iitk.ac.in	191173

Introduction

The increased use of data driven Machine Learning applications has exacerbated the issue of privacy wherein an algorithm that learns from the data fed to it, might retain it forever, unless otherwise 'unlearned', thereby putting users at the risk of all sorts of privacy breaches. In this project, we aim to tackle the problem of Machine Unlearning, viz., the problem of unlearning a subset of the dataset initially used to train a machine learning algorithm, through the framework of Variational Inference.

Prior Work

Assume D denotes the dataset initially used to train the model $M\big|_D$ with $\boldsymbol{\theta}$ denoting the collective parameters. $D_e \subset D$ denotes data subset that is to be erased, $D_r \subset D$ denotes the data subset that remains after erasing D_e, hence, $D_r = D \setminus D_e$, $D_r \cap D_e = \phi,\ D_r \cup D_e = D$. We wish to infer $M\big|_{D_r}$ from $M\big|_D$.

The naive approach to this would be to simply retrain our model from scratch using only the remaining dataset D_r. Such an approach can be forgiving when the data we're dealing with is small enough. But that is not the case with most real-world scenarios.

Alternate probabilistic approach, called **Bayesian Unlearning**, tries to recover the posterior belief $p(\boldsymbol{\theta}, D_r)$ of parameters **given remaining data** from $p(\boldsymbol{\theta}, D)$, which is the posterior belief of parameters **given full data** D. This approach is said to be **exact** wherein we recover the exact $p(\boldsymbol{\theta}, D_r)$ from exact $p(\boldsymbol{\theta}, D)$.

However, this is possible only for a select few ML models, viz., Naive Bayes, linear regression et al. Similar approach but referred to as the **approximate** version of it, tries to recover approximate $p(\boldsymbol{\theta}, D_r)$ denoted as $q(\boldsymbol{\theta}, D_r)$ from exact $p(\boldsymbol{\theta}, D)$, is applicable to a broader family of ML models. However, exact $p(\boldsymbol{\theta}, D)$ is rarely available in closed form.

Acknowledging this fact, this[1] paper approaches the problem of unlearning via approximate Bayesian unlearning along with making use of only the approximate posterior $p(\boldsymbol{\theta}, D)$ denoted as $q(\boldsymbol{\theta}, D)$.

This method is referred to as **Approximate Bayesian Unlearning using Approximate Posterior Belief**[1] which reduces the problem of unlearning to an optimisation problem and will make use of the variational inference framework.

Aim

In this project, our aim is to study the method of **Approximate Bayesian Unlearning using Approximate Posterior Belief** to tackle the problem of machine unlearning. Further, we aim to implement this approach using real-world datasets[2] on models for which exact unlearning is not possible and perform a comparative study versus the naive approach and possibly propose improvements.

Tentative Plan of Action

As we progress, we aim to break down the theory and math behind variational inference[3] and Bayesian Unlearning[1] (by 30^{th} March). We then aim to carry out our implementation by 15^{th} April and subsequently prepare our final project report and presentation. We will be maintaining a spreadsheet which will be timely updated with relevant resources that we come across as we progress.

Dataset

Many datasets are publicly available (e.g. banknote authentication dataset[2]). Other datasets will be explored as we progress.

References

[1] Q. P. Nguyen, B. K. H. Low, and P. Jaillet, "Variational bayesian unlearning," Advances in Neural Information Processing Systems, vol. 33, 2020.

[2] D. Dua and C. Graff. UCI machine learning repository, 2017.

[3] David M. Blei, Alp Kucukelbir & Jon D. McAuliffe (2017) Variational Inference: A Review for Statisticians, Journal of the American Statistical Association.