Differential Privacy & Approximate Bayesian Inference

Mrinank Sharma

Supervisor: Dr Richard E. Turner

Department of Engineering
University of Cambridge

This report is submitted for the degree of
Master of Engineering

Pembroke College May 2019

Abstract

ms2314: Need to write this

Table of contents

Chapter 1

Introduction

Machine learning methods are trained on large quantities of data, leveraging information within the dataset in order to make predictions about previously unseen data and make decisions. Recently, such methods have found use in scenarios where the data used in personal and sensitive, one example being the use of human genomic data to predict drug sensitivity [1]. Typically, data relating to individuals in training datasets are *anonymised*, for example, by removing all identifiable information (such as names, addresses, etc) and replacing this information with an anonymous identifier. However, Narayanan and Shmatikov show that anonymisation is insufficient, partially due to the availability of *auxiliary information* i.e. additional, publicly available information. When Netflix released a dataset in 2006 containing movie ratings for approximately 500000 subscribers with names replaced with identifiers, the data could be combined with public ratings on Internet Movie Database to identify movie ratings of two users. [2] Intuitively, data-points about individuals are highly dimensional meaning that anonymisation is insufficient, a further example being that even when sharing DNA sequence data without identifiers, it is possible to recover particular surnames using additional metadata. [3]

ms2314: Look into these papers more

Whilst a particular user's film ratings are not particularly sensitive, a lack of privacy in the areas of healthcare and public policy are critical.

Fredrikson et. al have shown that *model inversion attacks* are possible, where an adversary seeks to learn information about training data given model predictions. In particular, a neural network for facial recognition which returned confidence values was exploited in order to recover the image of a training set participant. Whilst more sophisticated attacks will be required for algorithms which do not provide confidence information, it is therefore possible for an adversary to recover anonymised training data-points and then de-anonymise this information using auxiliary information. [4]

The increasing availability and affordability of mobile smart-phones means that the *federated learning* context, where data across a number of clients is used to train a global model

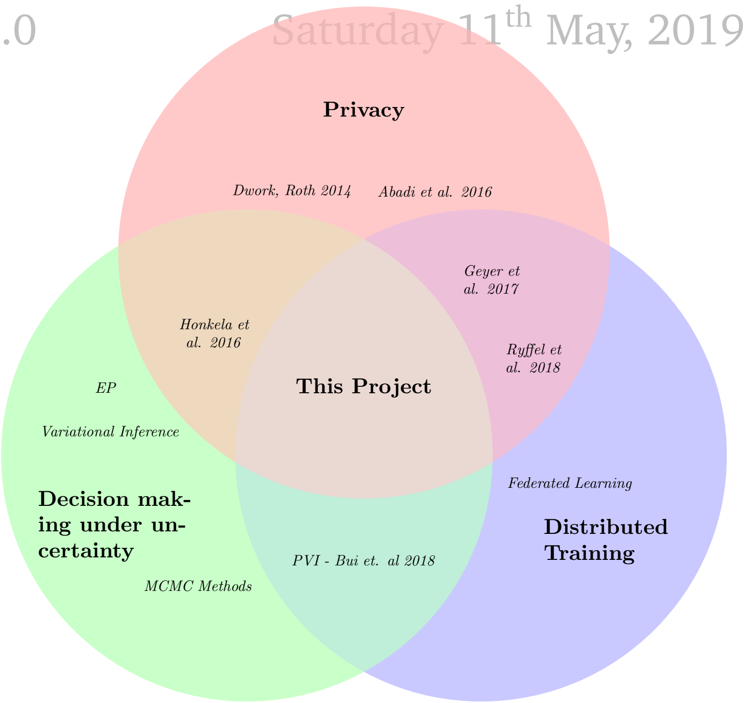

Fig. 1.1 Project aims, including other work within this area.

without transferring local data to a central server, is particularly interesting. Additionally, federated learning schemes reduce power consumption and intuitively give stronger privacy by removing the requirement of transferring entire local datasets. [5] Additionally, in order to make optimal decisions, it is important to model the uncertainty in what is known. This corresponds to performing Bayesian inference and producing a *posterior distribution* over unknown variables.

ms2314: Update Venn Diagram - also add these papers in the references of this report

This project aims to develop a generalised method to enable Bayesian inference to be performed in contexts where data is distributed over a number of clients whilst also providing privacy guarantees for each client.

Chapter 2

Literature Review

2.1 Differential Privacy

Differential privacy is a mathematical technique which formalises privacy and is able to numerically quantify the level of privacy that some method provides. This is particularly useful not only from the point of view of a designer who is able to compare techniques formally but also from the point of view of a client whose data we seek to protect as this formalism enables them to choose particular settings corresponding to the level of privacy that they seek.

ms2314: Perhaps add diagram showing privacy barrier / we generally assume that the adversary is able to do anything to the data after some sort of model has been released.

Definition 2.1.1 (ϵ-Differential Privacy) *A randomized algorithm, $\mathcal{A}$, is said to be ϵ-differentially private if for any possible subset of outputs, S, and for all pairs of datasets, $(\mathcal{D}, \mathcal{D}')$, which differ in one entry only, the following inequality holds:*

$$Pr(\mathcal{A}(\mathcal{D}) \in S) \leq e^{\epsilon} Pr(\mathcal{A}(\mathcal{D}') \in S) \tag{2.1}$$

Thus, differential privacy provides privacy in the sense that the output probability densities ought to be similar (i.e. bounded by e^{ϵ}) for datasets which are also similar (i.e. differ in only entry only). ϵ, a positive quantity, quantifies the level of privacy provided; large values of epsilon allow the resulting output densities to differ significantly whilst small values of epsilon mean that the output densities are similar. [6]

The key intuition behind differential privacy, noting the requirement that the algorithm is **randomized**, is to introduce noise which obscures the contribution of any particular data-point meaning that any adversary is unable to determine whether a particularly output was simply due to noise or due to a specific data-point.

Often, the above definition of differential privacy is slackened by introducing an extra privacy variable.

Definition 2.1.2 ((ϵ, δ)-Differential Privacy) . *A randomized algorithm, $\mathcal{A}$, is said to be (ϵ, δ) differentially private if for any possible subset of outputs, S, and for all datasets, $(\mathcal{D}, \mathcal{D}')$, which differ in one entry only, the following inequality holds:*

$$Pr(\mathcal{A}(\mathcal{D}) \in S) \leq e^{\epsilon} Pr(\mathcal{A}(\mathcal{D}') \in S) + \delta \tag{2.2}$$

Similar to ϵ, larger values of δ correspond to weaker privacy guarantees and $\delta = 0$ corresponds to a pure ϵ-differentially private algorithm. Note that it can be shown that (ϵ, δ)-DP provides a probabilistic ϵ-DP guarantee with probably $1 - \delta$. Additionally, it can be shown that differential privacy is immune to *post-processing* i.e. without additional knowledge, it is not possible to reduce the level of privacy provided by a differentially private algorithm. [6]

A useful quantity relating to a function of some dataset is the ℓ_2 sensitivity.

Definition 2.1.3 (ℓ_2 Sensitivity) *The ℓ_2 sensitivity of function $f : \mathcal{D} \rightarrow \mathbb{R}^n$, is denoted as $\Delta_2(f)$ and is defined as:*

$$\Delta_2(f) = \max_{\mathcal{D}, \mathcal{D}'} ||f(\mathcal{D}) - f(\mathcal{D}')||_2 \tag{2.3}$$

where $\mathcal{D}$ and $\mathcal{D}'$ differ in one entry only.

2.2 Federated Learning

Chapter 3

Design of Distributed DP Mechanisms

Chapter 4

Results & Discussion

Chapter 5

Conclusions

References

[1] Teppo Niinimäki, Mikko Heikkilä, Antti Honkela, and Samuel Kaski. Representation transfer for differentially private drug sensitivity prediction. *arXiv preprint arXiv:1901.10227*, 2019.

[2] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, May 2008. doi: 10.1109/SP.2008.33.

[3] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[4] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1322–1333, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813677. URL http://doi.acm.org/10.1145/2810103.2813677.

[5] Federated learning: Collaborative machine learning without centralized training data, Apr 2017. URL https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

[6] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3&#8211;4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.