Depth First Learning: SVGD

Week 1: Mathematical Preliminaries and Basics

May 15, 2019

Before jumping into all the math and methodology, we have to be able to understand the basics of whats going on. Most importantly, we will review the basics of measure theory and reproducing kernel hilbert spaces. Measure theory allows us to understand the notion of discrepancy measures between distributions, which we will use later on to quantify the difference between two arbitrary distributions of interest. Our other topic, Reproducing Kernel Hilbert Spaces (RKHS), will serve as the connection between measure theory and a practical machine learning algorithm. With RKHS, we will be able to define and optimize intractable measures which previously, were only useful for theoretical analysis or a restrictive class of functions. These two together set the foundation for defining a tractable Kernelized Stein Discrepancy, which serves as the driving factor behind Stein Variational Gradient Descent.

1 Table of Contents

To keep the PDF as clean as possible, you can refer to required and optional reading on the main class site.

1.1 Objectives

What are we trying to learn? Overall, we'll cover a host of topics, but mostly center around a single paper: Stein Variational Gradient Descent [2]. We will use many other papers and resources, but our main goal is to understand how SVGD works - we will cover everything about the algorithm: from mathematical foundations and practical implementations to tricks and failure points. Understanding this single paper will allow us to enter into an exciting, (relatively) young field of research, while drawing connections with established techniques and theoretical concepts along the way.

Why do we care about this? Stein's Method is a powerful statistical method, one that is at the disposal (and the focus) of many statisticians today. Recently, Stein's Method has made its way into machine learning and has already proved to be a fruitful research area. Stein's Method has deep connections to many machine learning problems of interest, and by the end of this guide, you should be able to understand the relevant mathematics behind this powerful tool.

1.2 General Curriculum

Over the next six weeks, we'll cover the following topics, with theoretical and practical percentage splits shown as $(T|P)$ next to each topic.

1. Introduction and Basics $(90|10)$
2. Stein's Method $(100|0)$
3. Kernelized Stein Discrepancy $(80|20)$
4. Stein Variational Gradient Descent $(60|40)$
5. SVGD as Gradient Flow $(90|10)$
6. Stein's Method in Reinforcement Learning $(40|60)$

2 Measure Theory

This first week is relatively more "lecture-style", with ensuring that concepts and terms are understood.

2.1 Definitions to Know

The exercise we tried was to make sure that people could explain the following terms in their own words, either via intuition or mathematics.

- Limit
- Cauchy Sequence
- Metric Space
- Complete, Banach, Hilbert Spaces
- Measure

2.2 Questions to Test Understanding

You can find the answers to the questions at the end of this PDF.

1. "However, Cauchy sequences are not the same as convergent sequences", but a property of Cauchy sequences is that they are **bounded**. What's the difference?
2. "The open interval (0, 1) is not complete whereas the closed interval [0, 1] is complete". Why? Can we use this example to get a intuitive definition of **complete**?
3. Explain the difference between a Banach and Hilbert Space. Is every Hilbert space a Banach space?

3 Kernels

3.1 Definitions to Know

The exercise we tried was to make sure that people could explain the following terms in their own words, either via intuition or mathematics.

- Kernel, Properties of Kernels
- Feature Map
- Gram Matrix

3.2 Questions to Test Understanding

You can find the answers to the questions at the end of this PDF.

1. In Machine Learning, kernels can be thought of as a "dot product" (a kind of similarity score) in high-dimensional space. Why would this be useful? Given a feature map, do we always have a corresponding kernel? Given any kernel, can we always explicitly write out the elements of the corresponding feature map?
2. How would kernels be useful in a classification problem like the one shown in Figure 1?

4 Reproducing Kernel Hilbert Space

The exercise we tried was to make sure that people could explain the following terms in their own words, either via intuition or mathematics.

- Hilbert Space
- Inner Product, Norms
- Reproducing Property

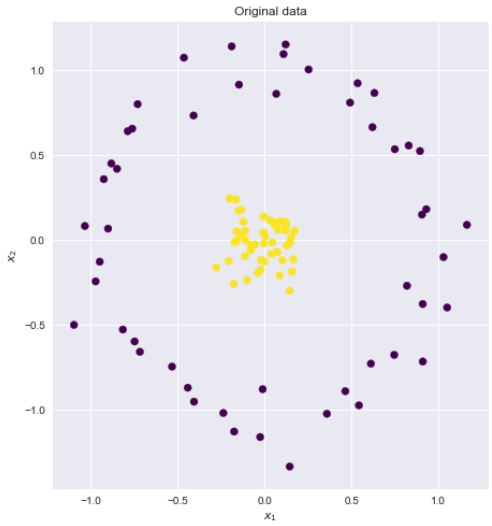

Figure 1: A Difficult Classification Problem for Linear Regression

Important Details Some of the important concepts from [1] are reproduced here. A Hilbert Space is a space $\mathcal{H}$ on which an **inner product is defined** (+ some stuff that won't be relevant here). A kernel is then defined as a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if there exists a space $\mathcal{H}$ and map $\phi : \mathcal{X} \to \mathcal{H}$ s.t k is the inner product of the two feature mapped points. Kernel functions are positive definite.

A Worked Example Here, we see how kernels and RKHS are used in practice.
Define ϕ as

$$\phi : \mathbb{R}^2 \to \mathbb{R}^3,\ \phi(x) = [x_1, x_2, x_1 x_2]^T \tag{1}$$

where kernel k is traditional dot product. If we define a function of features of x, we can define an equivalent representation for $f : \mathbb{R}^2 \to \mathbb{R}$ (since $X \in \mathbb{R}^2$).

$$f(x) + ax_1 + bx_2 + cx_1 x_2,\ f(\cdot) = [a, b, c]^T \tag{2}$$

That means that $f(x) \in \mathbb{R}$ is a function evaluated at a particular point, so we can write:

$$f(x) = f(\cdot) \cdot \phi(x) \tag{3}$$
$$f(x) = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}} \tag{4}$$

which means the evalution of $f(x)$ is an **inner product in feature space**.

Two main properties of the RKHS:

1. The feature map of every point is **in the feature space**: $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$

2. The reproducing property: $\forall x \in \mathcal{X}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$. This means that for any $x, y \in \mathcal{X}$:

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)_{\mathcal{H}}\rangle \tag{5}$$

4.1 Questions to Test Understanding

You can find the answers to the questions at the end of this PDF.

1. Explain the reproducing property in your own terms.
2. Explain how the space of all $\phi(x)$ can be smaller than $\mathcal{H}$, the space of functions, as seen in Figure 2?

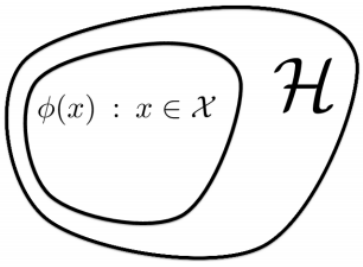

Figure 2: Feature space and mapping

5 Solutions

In this section, we provide the solutions to the questions listed above.

5.1 Solutions to Section 2.2

5.2 Solutions to Section 3.2

5.3 Solutions to Section 4.1

References

[1] A. Gretton. Introduction to rkhs, and some simple kernel algorithms, 2015.

[2] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2016.