Extending some linear methods to nonlinear methods by positive definite kernel and RKHS

yataka

0 Motivation and Notation

0.1 Motivation

Thre are a lot of Machine Learning methods which approximate linear data. However, most of data in this World are non-linear. Thus, we show you a method, called kernel method, that embed data of input space into high dimentional vector space with inner.

0.2 Notation

$\mathcal{X}$: input space
$\mathcal{Y}$: output space
$\mathcal{L}$: Loss function
$\mathbb{K}$: field
V: vector space
$(V, \|\cdot\|)$: normed space
$(V, \langle\cdot,\cdot\rangle)$: inner space
$\mathcal{H}$: hypothesis space or Hilbert space
$\mathcal{H}_k$: RKHS with reproducing kernel k

1 Functional Analysis

Definition 1.1. (vector space)
Let $\mathbb{K}$, V be a field and a set with two operations, addition and scalar multiplication. If V statisfies

1. V becomes commutative group by addition,
2. $\forall\alpha \in \mathbb{K}, \forall u, v \in V, \alpha(u+v) = \alpha u + \alpha v$,
3. $\forall\alpha, \beta \in \mathbb{K}, \forall v \in V, (\alpha+\beta)v = \alpha v + \beta v$,
4. $\forall\alpha, \beta \in \mathbb{K}, \forall v \in V, \alpha(\beta v) = (\alpha\beta)v$ and
5. $\exists 1_{\mathbb{K}} \in \mathbb{K}$ s.t. $\forall v \in V, 1_{\mathbb{K}} v = v$.

Then, V is called **vector space** over $\mathbb{K}$.

We consider only real verctor space or complex vector space in this article. Thus, $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

Definition 1.2. (normed vector space)
Let V, $\|\cdot\|$ be a vector space over $\mathbb{K}$ and a map from V to $\mathbb{K}$. $\|\cdot\|$ is called **norm** on $\mathbb{K}$ when $\|\|$ statisfies

1. $\forall v \in V, \|v\| \geq 0,$
2. $\forall v \in V, \|v\| = 0 \iff v = 0,$
3. $\forall \alpha \in \mathbb{K}, \forall v \in V, \|\alpha v\| = |\alpha|\|v\|$ and
4. $\forall v, w \in V, \|v + w\| \leq \|v\| + \|w\|.$

A pair $(V, \|\|)$ is called **normed vector space** or **normed space**

Proposition 1.3. Suppose that $(V, \|\|)$ is normed space. Then, norm space is metric space by a distance $d : V \times V \to \mathbb{K}$ generated by norm,

$$d(x, y) = \|x - y\|.$$

Definition 1.4. (complete)
Let (X, d) be metric space. (X, d) is called **complete** when every cauchy sequence in X converges a element in X.

Definition 1.5. (compact)
Let $(X, \mathscr{O})$ be topological space. X is called compact if every open covering of X has finite open covering. i.e.

$$\forall \{O_\lambda\}_{\lambda \in \Lambda} \subset \mathscr{O}, X = \bigcup_{\lambda \in \Lambda} O_\lambda \Longrightarrow \exists \lambda_1, \lambda_n, \cdots, \lambda_n \text{ s.t. } X = \bigcup_{i=1}^{n} O_{\lambda_i}$$

Definition 1.6. (Banach space)
Let $(V, \|\|)$ be normed space. $(V, \|\|)$ is called **Banach space** when V is complete by a distance generated by norm.

Banach space is very important space in mathematics. Therefore, I show you some examples of Banach space.

Example 1.7. $(C[X])$
Let X be compact space. Suppose that $C[X]$ is a collection of all continuous functions on X. We define a norm on X below:

$$\|x\|_\infty = \max\{|x(t)| \; | t \in X\}.$$

Then, $(C[X], \|\|_\infty)$ becomes Banach space.

Example 1.8. $(\mathcal{L}^p(a, b),\, p \geq 1)$
Let (a, b) be interval. Suppose that $((a, b), \mathcal{B}(a, b), L)$ is measure space where $\mathcal{B}(a, b)$ is Borel space and L is Lebesugue measure. we define $L^p(a, b)$ as

$$L^p(a, b) := \left\{ x : (a, b) \to \mathbb{K} \mid \int_a^b |x(t)|^p dt < \infty \right\}$$

and a norm as

$$\|x\|_p = \left(\int_a^b |x(t)|^p dt \right)^{\frac{1}{p}}.$$

We consider a equivarence relation $\sim$ on $L^p(a,b)$. $x \sim y$ is defined $x(t) = y(t)$ $a.s.t \in (a,b)$. Then, a collection of all equivarence classes on $L^p(a,b)$ is donated $\mathcal{L}^p(a,b)$ and becomes a vector space under addition and scalar multiplication defined by

$$(x+y)(y) = x(t) + y(t) \qquad (\alpha x)(t) = \alpha x(t).$$

A pair $(\mathcal{L}^p, \|\|_p)$ is Banach space and called **$\mathbf{L^p}$ space**.

Definition 1.9. (inner space)
Let V be vector space over $\mathbb{K}$. Suppose that $\langle\cdot,\cdot\rangle$ is a map from $V \times V$ to $\mathbb{K}$ which statisfies

1. $\forall v \in V, \langle v, v\rangle \geq 0$,
2. $\forall v \in V, \langle v, v\rangle = 0 \iff v = 0$,
3. $\forall v, w \in V, \langle v, w\rangle = \overline{\langle w, v\rangle}$,
4. $\forall u, v, w \in V, \langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$ and
5. $\forall v, w \in V, \forall \alpha \in \mathbb{K}, \langle \alpha v, w\rangle = \alpha\langle v, w\rangle$.

Then, a pair $(V, \langle\cdot,\cdot\rangle)$ is called **inner space** and $\langle\cdot,\cdot\rangle$ is called **inner** on V.

Proposition 1.10. Let $(V, \langle\cdot,\cdot\rangle)$ be inner space over $\mathbb{K}$. we difine a map $\|\cdot\|$ from V to $\mathbb{K}$ as

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

Then, $\|\cdot\|$ is a norm on V and is called norm generated by inner.

Definition 1.11. (Hilbert space)
Let $(V, \langle\cdot,\cdot\rangle)$ be inner space. V is called **Hilbert space** if V is banach space by a norm generated by inner.

2 Positive definate kernel

Definition 2.1. (positive definate kernel)
Let $\mathcal{X}$, k be a set and a map from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{K}$. k is called **positive definate kernel** on $\mathbb{K}$ if k has two conditions below:

1. $\forall x, y \in \mathcal{X}, k(x, y) = k(y, x)$ and
2. $\forall n \in \mathbb{N}, \forall x_1, \cdots, x_n \in \mathcal{X}, \forall c_1, \cdots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

second condition is called **definiteness**. definiteness under symmetrically means that a matrix

$$\begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

is positive-semidefinite. This symmetric matrix is called **gram matrix**.

Proposition 2.2. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ be positive definate kernel. Then,

1. $\forall x \in \mathcal{X}, k(x, x) \geq 0$,
2. $\forall x, y \in \mathcal{X}, |k(x, y)|^2 \geq k(x, x)k(y, y)$ and

3. every subset $\mathcal{Y}$, $k_{|\mathcal{Y}\times\mathcal{Y}}$ is positive difine kernel too.

Proof. 1.) From definateness of k, $1 \times 1 \times k(x,x) \geq 0$ for every x in $\mathcal{X}$. Hence, $k(x,x) \geq 0$.
2.) I can't prove....
3.) It is clear. □

Proposition 2.3. Let $\mathcal{X}$, $\{k_n\}_{n\in\mathbb{N}}$ be a set and a sequence of positive definate kernal on $\mathbb{K}$. Then, $\alpha k_1 + \beta k_2 (\alpha, \beta \geq 0)$, $k_1 k_2$, $\lim\limits_{n\to\infty} k_n$ are also positive definate kernels. However, we asusme that

1. $\forall x, y \in \mathcal{X}, k_1 k_2(x,y) = k_1(x,y)k_2(x,y)$ and
2. $\{k_n\}_{n\in\mathbb{N}}$ converges a map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$.

Proof. just a moment please.... □

Proposition 2.4. Let $\mathcal{X}$ be a set.

1. A non negative constant map $k : \mathcal{X} \times \mathcal{X} \to \{c\}$ $(c \in \mathbb{R}_{\geq 0})$ is positive definate kernel.
2. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ be positive definate kernal. Then, a map $k^{'} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$, defined by

$$k^{'}(x,y) = f(x)k(x,y)\overline{f(y)},$$

is positive definate kernel for every function $f : \mathcal{X} \to \mathbb{C}$.

Proposition 2.5. (kernel tric)
Let $\mathcal{X}$, $(V, \langle\cdot,\cdot\rangle)$ be a set and inner space. a map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$, defined by

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle,$$

is positive definate kernal for every function $\Phi : \mathcal{X} \to V$.

3 Reproducing Kernel Hilbert Space

Definition 3.1. (reproducing kernel Hilbert space)
Let $\mathcal{X}$, $\mathcal{H}$ be a set and Hilbert space from some functions on $\mathcal{X}$.
$\mathcal{H}$ is called **Reproducing Kernel Hilbert Space** or **RKHS** simply when

$$\forall x \in \mathcal{X}, \exists k_x \in \mathcal{H} \text{ s.t. } \forall f \in \mathcal{H}, \langle f, k_x \rangle = f(x).$$

A map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$, defined by $k(y,x) = k_x(y)$, is called **reproducing kernel** of $\mathcal{H}$.

Proposition 3.2. Let $\mathcal{H}_k$ be a RKHS. Then, the reproducing kernal of $\mathcal{H}_k$ is positive definate kernal and is unique.

Proof. For every $x_1, x_2, \cdots, x_n \in \mathcal{X}$ and $c_1, c_2, \cdots, c_n \in \mathbb{K}$,

$$\begin{aligned}\sum_{i=1,j=1}^{n} c_i \overline{c_j} k(x_i, x_j) &= 1 \\ &= 2\end{aligned}$$

□

4 The Principle of Machine Learning

Learning Theory is devided into three parts: Supervised Learning, Unsupervised Learning and Reinforcement Learning. This article explains only Supervised Learning. Machine Learning and Deep Learning are Supervised Learning. The purpose of Supervised Learning is learning "good" functions from training data. In the next subsection, I explain basic definitions of Firstly, we define a fundamental space of Machine Learning based on Statistical Learning theory.

4.1 Fundametal space

Definition 4.1. (input space)
$\mathcal{X} := \mathbb{R}^d$ is called **input space**.

Definition 4.2. (output space)
$\mathcal{Y} := \mathbb{R}^m$ is called **output space**.

Definition 4.3. (Hypothesis space)
Let $\mathcal{X}$, $\mathcal{Y}$ be a input space and a output space respectively. **Hypothesis space**, donated $\mathcal{H}$, is a set of function from $\mathcal{X}$ to $\mathcal{Y}$ with some restrictions. i.e.

$$\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y} \mid f \text{ stisfy some conditions}\}$$

A element f of $\mathcal{H}$ is called **hypothesis**. The detail of conditions is explained later.

Definition 4.4. (Data)
Let $\mathcal{X}$, $\mathcal{Y}$ be a input space and a output space respectively. A finite subset of $\mathcal{X} \times \mathcal{Y}$ is called Data sequence or Data simply. Data is donated D.

Definition 4.5. (Training data and Testing data)
Data D is devided into two disjoint data, Training data S and Testing data T.

Definition 4.6. (Loss function)
Let $\mathcal{H}$ be a Hypothesis space and D be data. **Loss function**, donated L_D, is function from $\mathcal{H}$ to $\mathbb{R}$.

Definition 4.7. (Machine Learning space)
Let $\mathcal{X}$, $\mathcal{Y}$, D, $\mathcal{H}$, $L_D : \mathcal{H} \to \mathbb{R}$ be a input space, output space, Data, Hypothesis space and Loss function respectively. Then, the -5 tuple $(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, L_D)$ is called **Machine Learning space**.

Definition 4.8. (Learning)
Let $\mathcal{H}$ be a Hypothesis space and $L_D : \mathcal{H} \to \mathbb{R}$ be a loss function. A process that find a hypothesis $f_{opt} \in \mathcal{H}$ that minimalize $\{L_D(f) \mid f \in \mathcal{H}\}$ from Train data S is called Learning. then, $f_{opt} \in \mathcal{H}$ is called optimal hypothesis.

Attention 4.9. Generally, f_{opt} is not equal to f_{min}. (f_{min} is argment of minimam $\{L_D(f) \mid f \in \mathcal{H}\}$)

Definition 4.10. (Machine Learning)
Let $(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, L_D)$ be a Machine Learning space. A learning on $(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, L_D)$ is called Machine Learning.

4.2 Some examples

Simple Regression Analysis is the most fundametal Machine Learning Model.

4.2.1 Setting Problem

I define Machine Learning space as below: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$,

$$\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y} \mid f(x) = wx,\ w \in \mathbb{R}\},$$

$$L_D(f) = \sum_{i=1}^{N} |f(x_i) - t_i|^2 \quad (f \in \mathcal{H})$$

Then, Learning in this space is called Simple Regression Analysis.

4.2.2 Solution of this Problem

Let $f \in \mathcal{H}$ be $f(x) = wx$. then,

$$\begin{aligned}\frac{dL}{dw} &= \frac{d}{dw}\sum_{i=1}^{N}(wx_i - t_i)^2 \\ &= \sum_{i=1}^{N} 2x_i(wx_i - t_i) \\ &= 2\sum_{i=1}^{N} x_i(wx_i - t_i) = 0\end{aligned}$$

So optimal hypothesis in this problem is

$$f_{opt}(x) = \frac{\sum_{i=1}^{N} x_i t_i}{\sum_{i=1}^{N} x_i^2} x.$$

4.2.3 Improving this model

The model above has many improving points. For examples

1. It is difficult to predict non linear data.
2. In many case, there are more than one factor.

Firstly, I think about second problem. I think of some independent reasons as one vector.

4.3 Multiple Regression Analysis

4.3.1 Setting problem

I define Machine Learning space as below: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$,

$$\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y} \mid f(x) = Wx,\ W \in \mathbb{R}^{1\times d}\},$$

$$L_D(f) = \sum_{i=1}^{N} |f(x_i) - t_i|^2 \quad (f \in \mathcal{H})$$

Then, Learning in this space is called Multiple Regression Analysis.

4.3.2 Solution of this Problem

Let X, w, t be $N \times d$ Input Data matrix that i th row is ith Input data, parameter vector in $\mathbb{R}^{1\times d}$ and target vector in $\mathbb{R}^N$ respectively. i.e.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nd} \end{bmatrix},$$

$t = (t_1, t_2, \cdots, t_N)^T$ and $w = (w_1, w_2, \cdots, w_d)^T$ (T refer Transpose of vector or matrix). Optimal hypothesis in this problem is

$$f_{opt}(x) = ((X^TX)^{-1}X^Tt)^Tx.$$

5 Kernel method

I showed some methods in section 4. However, these methods is for linear data. If you have learned Machine Learning, You may know Neural Network for non linear data. However, I explain Neural Network after this section and show new method, Kernel method, for non linear data in this section.

6 Neural Network

Definition 6.1. (Shallow Neural Network for Regression)
we difine Machine Learning space as below: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$,

$$\mathcal{H}_{SNN} = \{f : \mathcal{X} \to \mathcal{Y} \mid f(x) = W_2\eta\,(W_1x - b_1) - b_2,\ W_1 \in \mathbb{R}^{L\times d}, W_2 \in \mathbb{R}^{1\times L}, b_1 \in \mathbb{R}^L, b_2 \in \mathbb{R}\},$$

$$L(f) = \frac{1}{N}\sum_{i=1}^{N} |f(x_i) - t_i|^2$$

,where $\eta : \mathbb{R}^L \to \mathbb{R}^L$ is activation function, non linear and Lipschitz continuous. A element of $\mathcal{H}_{SNN}$ is called **Shallow Neural Network for Regression**.

References

[1] https://tutorials.chainer.org/ja/index.html
[2] https://github.com/Runnrairu/machinelearning_text
[3] カーネル法入門―正定値カーネルによるデータ解析・福水健次・2010
[4] 統計的学習理論・金森 敬文・2015
[5] 函数解析 POD 版・前田 周一郎・2007