國立臺灣大學理學院應用數學科學研究所

碩士論文

Graduate Institute of Applied Mathematical Sciences
College of Science
National Taiwan University

Master Thesis

統計機器學習在最佳成長投資組合之應用

Statistical Machine Learning in Growth Optimal Portfolio

蘇俊德

Chun-Te Su

指導教授：王藹農 博士

Advisor: Ai-Nung Wang, Ph.D.

中華民國 110 年 1 月

January 2021

doi:10.6342/NTU202100742

目　錄

doi:10.6342/NTU202100742

國立臺灣大學碩士學位論文
口試委員會審定書

統計機器學習在最佳成長投資組合之應用
Statistical Machine Learning in Growth Optimal Portfolio

本論文係蘇俊德君（R02246008）在國立臺灣大學應用數學科學研究所完成之碩士學位論文，於民國 108 年 7 月 26 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

王若雲 （簽名）

（指導教授）

系主任、所長 （簽名）

（是否須簽章依各院系所規定）

doi:10.6342/NTU202100742

致謝

首先誠摯地感謝指導教授王藹農老師，讓我得以選擇感興趣想做的題目當作碩士論文，提供我相當大的彈性並且解決我在過程中遇到的問題，讓我在這篇論文研究的日子裡獲益匪淺，老師對做研究的熱情和嚴謹態度更是我學習的榜樣。

碩士班的日子裡，感觸真是一言難盡，研究室好夥伴蘇偉宏、陳柏傑、陳冠中、陳健樺，彼此互相幫忙、分享生活與共同打拼的革命情感。另外也特別感謝周謀鴻老師，讓我擔任老師的微積分班助教，認識了許多科系大學部的學生，也在教學中找到了自己的價值，發現自己對教學與教育充滿了熱忱與想法，獲得同學們的肯定得以通過多次傑出 TA 遴選，這段日子是改變我未來人生道路選擇的關鍵時期!

而最後也是最重要的是，家人在背後關心與默默支持我做的決定，是我前進的最大動力，沒有家人的體諒與包容，相信我的研究生生活將是很不一樣的光景。

doi:10.6342/NTU202100742

中文摘要

最佳成長投資組合（GOP）是在任何時間範圍內都具有最大期望增長率的投資組合，隨著時間範圍的增加，這種投資組合肯定會勝過其他任何不同的投資策略。在本文中，我使用了非母數統計方法和統計機器學習工具來估計市場數據的分佈。此外，模擬了平穩型數據和非平穩型數據以表示市場數據，通過了解市場分佈，我建立了最佳成長投資組合。GOP 的文獻綜述在第 1 節中、理論研究在第 2 節中介紹、方法將在第 3 節中簡要介紹且在第 4 節中做出模擬結果。

關鍵字：Kelly 公式、最佳成長投資組合、核迴歸、局部多項式估計、支持向量迴歸

doi:10.6342/NTU202100742

Abstract

The growth-optimal portfolio (GOP) is a portfolio which has a maximal expected growth rate over any time horizon. As a consequence, this portfolio is certain to outperform any other significantly different strategy as the time horizon increases. In this thesis, I used nonparametric statistic and machine learning tools to estimate the distribtuion of market data. Also, simulated both stationary and nonstationary data to represent market data. Through understanding the distribution of market, I built up the growth-optimal portfolio. The literature reviewing of GOP are in section 1. The theoretical studieso are presented in section 2. Methodology will be briefly introduced in section 3. Simulation results are in section 4.

Key words: Kelly Formula, Growth Optimal Portfolio, kernel regression, local polynomial, support vector regression

doi:10.6342/NTU202100742

1 Introduction

1.1 Literature Review

The growth-optimal portfolio (GOP) is a portfolio which has a maximal expected growth rate over any time horizon. As a consequence, this portfolio is certain to outperform any other significantly different strategy as the time horizon increases. The origins of the GOP have usually been tracked to the paper [Kelly(1956)], hence the name "kelly criterion", which is used synonymously.

Kelly's motivation came from gambling and information theory, and his paper derived a striking but simple results: there is an optimal gambling strategy, such that, with probability one, this optimal gambling strategy will accumulate more wealth than any other different strategy. However, GOP is a portfolio with several aspects, one of which is the maximization of the geometric mean. In this respect, the history might be said to have its origin in [Williams (1936)], who considered speculators in a multi-period setting and reached the conclusion that, due to compounding, speculators should consider the geometric and not the arithmetic mean. [Breiman (1960, 1961)] expanded the analysis of [Kelly (1956)] and discussed applications for long-term investment and gambling in a more general mathematical setting.

Calculating the GOP is generally very dicult in discrete time and is treated in [Bellman and Kalaba (1957)], [Elton and Gruber (1974)] and [Maier et al. (1977b)], although the diculties disappear whenever the market is complete. This is similar to the case when jumps in asset prices happen at random. In the continuous-time continuous-diusion case, the problem is much easier and was solved in [Merton (1969)]. Today, solutions to the problem exist in a semi-explicit form and, in the general case, the GOP can be characterized in terms of the semimartingale characteristic triplet.

1.2 Growth-Optimal Portfolios

In this project, we focus on the discrete type GOP. Consider a market consisting of a nite number of non-dividend paying assets. The market consists of d+1 assets, represented by a d+1 dimensional vector process, S, where

$$S = \{S(t) = (S^{(0)})(t), \cdots, S^{(d)}(t), t \in \{0, 1, \cdots, T\}\}$$

Define the return process

$$R = \{R(t) = (R^0(t), \cdots, R^d(t)), t \in \{1, 2, \cdots, T\}\}$$

by $R^i(t) = \dfrac{S^{(i)}(t)}{S^{(i)}(t-1)} - 1$. Often, it is assumed that returns are independent over time, and, for simplicity, this assumption is made in this section.

Investors in such a market consider the choice of a strategy

$$\delta = \{\delta(t) = (\delta^{(0)})(t), \cdots, \delta^{(d)}(t), t \in \{0, \cdots, T\}\}$$

where $\delta^{(i)}(t)$ denotes the number of units of asset i that are being held during the period $(t, t+1]$. We need to give some rational constraints on strategy. Firt, the strategy cannot use future information. Second, we require investor to remain solvent. Third, we request investor re-invest all money in each time step. This kind of strategy is defined as below:

doi:10.6342/NTU202100742

Definition A trading strategy, δ, generates the portfolio value process $S^\delta(t) = \delta(t)S(t)^T$. The strategy is called admissible if it satisfies the three conditions below.

- Non-anticipative: The process δ is adapted to the filtration $\mathcal{F}$, meaning that $\delta(t)$ can only be chosen based on information available at time t.
- Limited liability: The strategy generates a portfolio process $S^\delta(t)$ which is nonnegative.
- Self-financing: $\delta(t-1)S(t) = \delta(t)S(t), t \in \{1, \cdots, T\}$ or equvalently $\triangle S^\delta(t) = \delta(t-1) \triangle S(t)$.

Consider an investor who invests a dollar of wealth in some portfolio. At the end of period T, his wealth becomes

$$S^\delta(T) = S^\delta(0) \prod_{i=1}^{T} (1 + R^\delta(i)),$$

where $R^\delta(t)$ is the return in period t. If the portfolio fractions are xed during the period, the right hand side is the product of T independent and identically distributed (i.i.d.) random variables. The geometric average return over the period is then

$$(\prod_{i=1}^{T} (1 + R^\delta(i)))^{\frac{1}{T}}$$

Because the returns of each period are i.i.d., this average is a sample of the geometric mean value of the one-period return distribution. For discrete random variables, the geometric mean of a random variable X taking (not necessarily distinct) values $x_1, \cdots, x_S$, with equal probabilities, is dened as

$$G(X) = (\prod_{s=1}^{S} x_s)^{\frac{1}{S}} = (\prod_{k=1}^{K} \tilde{x_k}^{f_k}) = exp(E[log(X)])$$

where $\tilde{x_k}$ are the distinct values of X and f_k is the frequency at which $X = x_k$, that is $f_k = P(X = x_k)$. In other words, the geometric mean is the exponential function of the growth rate $g^\delta(t) = E[log(1 + R^\delta(t))]$ of some portfolio. Generally, one defines the geometric mean of an arbitrary random variable by

$$G(X) = exp(E[log(X)])$$

assuming the mean value $E[log(X)]$ is well-dened. Over long stretches, intuition dictates that each realized value of the return distribution should appear, on average, the number of times dictated by its frequency, and hence, as the number of periods increase, it would hold that

$$(\prod_{i=1}^{T} (1 + R^\delta(i)))^{\frac{1}{T}} = exp(\frac{\sum_{i=1}^{T} log(1 + R^\delta)(i)}{T}) \to G(1 + R^\delta(1))$$

As $T \to \infty$. This states that the average growth rate converges to the expected growth rate. Hence, we can define our growth-optimal portfolio by the following.

doi:10.6342/NTU202100742

Definition A solution of $\sup_{S^{(\delta)}(T)\in\Theta} E[log(\frac{S^{(\delta)}(T)}{S^{(\delta)}(0)})]$ is called a GOP.

The objective given above is often referred to as the geometric mean criteria. It is convenient to infer some properties of the GOP strategy by viewing it as a logarithmic return. Hence, the theorem below is nature.

Theorem 1 The GOP strategy has the following properties:

- The fractions of wealth invested in each asset are independent of the level of total wealth.
- The invested fraction of wealth in asset i is proportional to the return on asset i.
- The strategy is myopic

The r-st part is to be understood in the sense that the fractions invested are independent of current wealth. Moreover, the GOP strategy allocates funds in proportion to the excess return on an asset. Myopic means shortsighted and implies that the GOP strategy in a given period depends only on the distribution of returns in the next period. Hence, the strategy is independent of the time horizon. Despite the negative connotations the word myopic can be given, it may, for practical reasons be quite convenient to have a strategy which only requires the estimation of returns one period ahead. It seems reasonable to assume that return distributions further out in the future are more uncertain. To see why the GOP strategy depends only on the distribution of asset returns one period ahead, note that

$$E[log(S^{(\delta)}(T))] = log(S^{(\delta)}(0)) + \sum_{i=1}^{T} E[log(1 + R^{(\delta)}(i))]$$

In general, obtaining the strategy in an explicit closed form is not possible. This involves solving a non-linear optimization problem. Since, by Theorem 1, the GOP strategy is myopic and the invested fractions are independent of wealth, one needs to solve the problem

$$\sup_{\delta(t)} E_t[log(\frac{S^{(\delta)}(t+1)}{S^{(\delta)}}(t))]$$

for each $t \in \{0, 1, \cdots, T-1\}$, where E_t denotes the conditional expectation with respect to $\mathcal{F}_t$. Using the fractions $\pi^i_\delta(t) = \dfrac{\delta^{(i)}(t)S^{(i)}(t)}{S^{(\delta)}(t)}$, the problem can be written

$$\sup_{\pi_\delta(t)\in R^d} E[log(1 + (1 - \sum_{i=1}^{n} \pi^i_\delta)R^0(t) + \sum_{i=1}^{n} \pi^i_\delta(t)R^i(t))]$$

Through this way, the problem becomes to solve the optimal weight of logarithmic return.

doi:10.6342/NTU202100742

2 Emperical Log-Optimal Portfolio Selection

In this section, we will show that under memoryless assumption about the underlying process generating the asset prices, the best rebalancing is the log-optimal portfolio, which achieves the maximim asymptotic average growth rate. We will discuss the different between constant-rebalanced and dynamic portfolio section. Under the stationary assumption, we will show it is the same thing.

Consisder a market consisting of d assets. The evolution of the market in time is represented by a sequence of price vectors $S_1, S_2, \cdots \in R_+^d$, where

$$\mathbf{s_n} = (s_n^{(1)}, \cdots, s_n^{(d)})$$

such that the j-th component $s_n^{(j)}$ of s_n denotes the price of the j-th asset on the n-th trading period. In order to normalize, put $s_0^{(j)} = 1$. $\{s_n\}$ has an exponential trend:

$$s_n^{(j)} = e^{nW_n^{(j)}} \approx e^{nW^{(j)}}$$

with average growth rate (average yield)

$$W_n^{(j)} = \frac{1}{n} \ln s_n^{(j)}$$

and with asymptotic average growth rate

$$W^{(j)} = \lim_{n\to\infty} \frac{1}{n} \ln s_n^{(j)}$$

The static portfolio selection is a single-period investment strategy. A portfolio vector is denoted by $\mathbf{b_n} = (b^{(1)}, \cdots, b^{(d)})$. The j-th component $b^{(j)}$ of $\mathbf{b}$ denotes the proportion of the investors capital invested in asset j. We assume that the portfolio vector $\mathbf{b}$ has nonnegative components and sum to 1. The set of portfolio vectors is denoted by

$$\triangle_d = \{\mathbf{b} = (b^{(1)}, \cdots, b^{(d)}); b^{(j)} \geq 0, \sum_{j=1}^{d} b^{(j)} = 1\}$$

The aim of static portfolio selection is to achieve $\max_{1\leq j\leq d} W^{(j)}$.

2.1 Constantly-Rebalanced Portfolio Selection

In case of constantly-rebalanced portfolio, we fix a portfolio vector $b \in \triangle_d$, i.e., we are concerned with a hypothetical investor who neither consumes nor deposits new cash into his portfolio, but reinvests his portfolio each trading period. Note that, in this case, the investor has to rebalance his portfolio after each trading day to corrigate the daily price shifts of the invested stocks.

Let S_0 denote the investors initial capital. Then, at the beginning of the r-st trading period, $S_0 b^{(j)}$ is invested into asset j, and it results in return $S_0 b^{(j)} x_1^{(j)}$. Therefore, at the end of the r-st trading period, the investors wealth becomes

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x_1^{(j)} = S_0 < \mathbf{b}, \mathbf{x_1} >$$

doi:10.6342/NTU202100742

For the second trading period, S_1 is the new initial capital

$$S_2 = S_1 < \mathbf{b}, \mathbf{x_2} >= S_0 < \mathbf{b}, \mathbf{x_1} >< \mathbf{b}, \mathbf{x_2} >$$

By induction, for the trading period n, the initial capital is S_{n-1}. Therefore,

$$S_n = S_{n-1} < \mathbf{b}, \mathbf{x_n} >= S_0 \prod_{i=1}^{n} < \mathbf{b}, \mathbf{x_i} >$$

The asymptotic average growth rate of this portfolio selection is

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n = \lim_{n\to\infty} (\frac{1}{n} \ln S_0 + \frac{1}{n} \sum_{i=1}^{n} \ln < \mathbf{b}, \mathbf{x_i} >) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ln < \mathbf{b}, \mathbf{x_i} >$$

If the market process $\{X_i\}$ is memoryless, i.e., it is a sequence of independent and identically distributed (i.i.d.) random return vectors, then we show that the best constantly-rebalanced portfolio (BCRP) is the log optimal portfolio

$$\mathbf{b}^* = arg \max_{\mathbf{b}\in\triangle_d} E[\ln < \mathbf{b}, \mathbf{X_1} >]$$

This optimality means that, if $S_n^* = S_n(\mathbf{b}^*)$ denotes the capital after day n achieved by a log optimal portfolio strategy $\mathbf{b}^*$, then, for any portfolio strategy $\mathbf{b}$ with finite $E[(\ln < \mathbf{b}, \mathbf{X_1} >)^2]$ and with capital $S_n = S_n(\mathbf{b})$, and for any memoryless market process $\{\mathbf{X_n}\}_{-\infty}^{\infty}$,

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n \leq \lim_{n\to\infty} \frac{1}{n} \ln S_n^* \qquad almost \quad surely$$

and maximal asymptotic average growth rate is

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n^* = W^* = E\{\ln < \mathbf{b}^*, \mathbf{X_1} >\} \qquad almost \quad surely$$

2.2 Time-Varying Portfolio Selection

For a general dynamic portfolio selection, the portfolio vector may depend on the past data. As before, $\mathbf{x_i} = (x_i^{(1)}, \cdots, x_i^{(d)})$ denotes the return vector on trading period i. Let $\mathbf{b} = \mathbf{b_1}$ be the portfolio vector for the r-st trading period. For initial capital S_0 , we obtain

$$S_1 = S_0 < \mathbf{b_1}, \mathbf{x_1} >$$

For the second trading period, S_1 is the new initial capital, the portfolio vector is $\mathbf{b_2} = \mathbf{b}(\mathbf{x_1})$, and

$$S_2 = S_0 < \mathbf{b_1}, \mathbf{x_1} >< \mathbf{b}(\mathbf{x_1}), \mathbf{x_2} >$$

For the n-th trading period, the portfolio vector is $\mathbf{b_n} = \mathbf{b}(\mathbf{x_1}, \cdots, \mathbf{x_n}) = \mathbf{b}(\mathbf{x_1^{n-1}})$ and

$$S_n = S_0 \prod_{i=1}^{n} < \mathbf{b}(\mathbf{x_1^{i-1}}), \mathbf{x_1} >= S_0 e^{nW_n(B)}$$

doi:10.6342/NTU202100742

with average growth rate

$$W_n(B) = \frac{1}{n}\sum_{i=1}^{n} \ln < \mathbf{b}(\mathbf{x_1^{i-1}}), x_i >$$

With the assumption of stationary process, we can use the conditionally log optimal portfolio to show that $\mathbf{B}^* = \{\mathbf{b}^*()\}$ is the best possible choice. More precisely, on trading period n, let $\mathbf{b}^*()$ be such that

$$E\{\ln < \mathbf{b}^*(\mathbf{X_1^{n-1}}), \mathbf{X_n} > | \mathbf{X_1^{n-1}}\} = \max_{\mathbf{b}()} E\{\ln < \mathbf{b}(\mathbf{X_1^{n-1}}), \mathbf{X_n} > | \mathbf{X_1^{n-1}}\}$$

If $S_n^* = S_n(B^*)$ denotes the capital achieved by a log optimal portfolio strategy $\mathbf{B}^*$, after n trading periods, then, for any other investment strategy $\mathbf{B}$ with capital $S_n = S_n(\mathbf{B})$ and with

$$\sup_n E\{(\ln < \mathbf{b_n}(\mathbf{X_1^{n-1}}), \mathbf{X_n} >)^2\} < \infty$$

and for any stationary and ergodic process $\{\mathbf{X_n}\}_{-\infty}^{\infty}$,

$$\limsup_{n\to\infty}(\frac{1}{n}\ln S_n - \frac{1}{n}\ln S_n^*) \leq 0 \qquad almost \quad surely$$

and

$$\lim_{n\to\infty}\frac{1}{n}\ln S_n^* = W^* \qquad almost \quad surely$$

, where

$$W^* = E\{\max_{\mathbf{b}()} E\{\ln < \mathbf{b}(\mathbf{X_{-\infty}^{-1}}), \mathbf{X_0} > | \mathbf{X_{-\infty}^{-1}}\}\}$$

is the maximum possible growth rate of any investment strategy. Note that, for memoryless marlets, $W^* = \max_b E\{\ln < \mathbf{b}, \mathbf{X_0} >\}$ which shows that, in this case the log optimal portfolio is the BCRP.

The optimality relations above give rise to the following denition:

Denition An empirical (data driven) portfolio strategy B is called universally consistent with respect to a class $\mathcal{C}$ of stationary and ergodic processes $\{X_n\}_{-\infty}^{\infty}$ if, for each process in the class,

$$\lim_{n\to\infty}\frac{1}{n}\ln S_n(\mathbf{B}) = W^* \qquad almost \quad surely$$

With the definition above, let us recapitulate the definition of log optimal portfolio:

$$E\{\ln < \mathbf{b}^*(\mathbf{X_1^{n-1}}), \mathbf{X_n} > | \mathbf{X_1^{n-1}}\} = \max_{\mathbf{b}()} E\{\ln < \mathbf{b}(\mathbf{X_1^{n-1}}), \mathbf{X_n} > | \mathbf{X_1^{n-1}}\}$$

For a fixed integer $k \geq 0$ large enough, we expect that

$$E\{\ln < \mathbf{b}(\mathbf{X_1^{n-1}}), \mathbf{X_n} > | \mathbf{X_1^{n-1}}\} \approx E\{\ln < \mathbf{b}(\mathbf{X_1^{n-1}}), \mathbf{X_n} > | \mathbf{X_1^{n-1}}\}$$

and

$$\mathbf{b}^*(\mathbf{X_1^{n-1}}) \approx \mathbf{b_k}(\mathbf{X_{n-k}^{n-1}}) = \max_{\mathbf{b}()} E\{\ln < \mathbf{b}(\mathbf{X_1^{n-1}}), \mathbf{X_n} > | \mathbf{X_1^{n-1}}\}$$

doi:10.6342/NTU202100742

Because of stationary,

$$\begin{aligned}\mathbf{b_k}(\mathbf{x_1^k}) &= \arg\max_{\mathbf{b}()} E\{\ln < \mathbf{b}(\mathbf{X_{n-k}^{n-1}}), \mathbf{X_n} > | \ \mathbf{X_{n-k}^{n-1}} = \mathbf{x_1^k}\} \\ &= \arg\max_{\mathbf{b}()} E\{\ln < \mathbf{b}(\mathbf{x_1^k}), \mathbf{X_{k+1}} > | \ \mathbf{X_1^k} = x_1^k\} \\ &= \arg\max_{b} E\{\ln < \mathbf{b}, \mathbf{X_{k+1}} > | \ \mathbf{X_1^k} = x_1^k\}\end{aligned}$$

which is the maximization of the regression fuction

$$m_b(x_1^k) = E\{\ln < \mathbf{b}, \mathbf{X_{k+1}} > | \ \mathbf{X_1^k} = x_1^k\}$$

Thus, a possible way for asymptotically optimal empirical portfolio selection is, based on past data, sequentially estimate the regression function $m_b(x_1^k)$ and choose the portfolio vector, which maximizes the regression function estimate.

3 Methodology

In this section, we will briefly introduce three methods that we use in our project. The first and second method we used are traditional nonparametric methods – kernel smoothing and local polynomial. The third method we use is a popular machine learning method – support vector regression. We compare three methods and show the results in next section, before that, we want ot mention some coditions of this kinds of method. Because we focus on time series data, the data we use is not independent. If we want to apply the method, we need to put some mixing conditions on data. That makes us guarantee the asymptotic properties.

3.1 Mixing condtions

The classical asymptotic theory in statistics is built on the central limit theory and law of large numbers for the sequence of independent random variables. A mixing time series can be viewed as a sequence of random variables for which the past and distant future are asymptotic independent. For mixing sequences, both the law of large numbers and central limit theorem can be established. In this section, we only introduce mixing condition for strictly stationary processes. The idea is to define mixing coefficients to measure the strength of dependence for the two segments of a time series that are apart of each other in time. For $n = 1, 2, \cdots$, define

$$\begin{aligned}\alpha(n) &= \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_n^\infty} | \ P(A)P(B) - P(A\cap B) \ | \\ \beta(n) &= E\{\sup_{B\in\mathcal{F}_n^\infty} | \ P(B) - P(B \mid X_0, X_{-1}, X_{-2}, \cdots) \ |\} \\ \rho(n) &= \sup_{X\in\mathcal{L}^2(\mathcal{F}_{-\infty}^0) Y\in\mathcal{L}^2(\mathcal{F}_n^\infty)} | \ Corr(X,Y) \ | \\ \varphi(n) &= \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_n^\infty} | \ P(B) - P(B \mid A) \ |\end{aligned}$$

where $\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{X_t, i \le t \le j\}$ and $\mathcal{L}^2(\mathcal{F}_i^j)$ consist of $\mathcal{F}_i^j$-measure random variables with finite second moment. When at least one of the mixing coefficients converges to 0 as $n \to \infty$, we may say the process $\{X_t\}$ is asymptotically independent.

doi:10.6342/NTU202100742

3.2 Kernel Regression Smoother

Consider the pair data $(X_1,\ Y_1), \ldots, (X_n,\ Y_n)$, which form an independent and identically distributed sample from a population $(X,\ Y)$. We now regard the data as being generated from the model:

$$Y = m(X) + \varepsilon,$$

where $E(Y|X = x) = 0$, $Var(Y|X = x) = \sigma^2(x)$, and X and ε are independent. Of interest is to estimate the regression function

$$m(x_0) = E(Y|X = x_0).$$

Without assuming a specific form of the regression function m, a datum point remote from x carries little information about the value $m(x)$. Thus, an intuitive estimator for the conditional mean function $m(x)$ is the running local average. An improved version of this is the locally weighted average

$$\begin{aligned}\hat{m}(x_0) =& \arg\min_{m(x_0)} \sum_{i=1}^{n} (Y_i - m(x_0))^2\, w(X_i,\ x_0) \\ =& \frac{\sum_{i=1}^{n} w(X_i,\ x_0) Y_i}{\sum_{i=1}^{n} w(X_i,\ x_0)},\end{aligned}$$

where $w(X_i,\ x_0)$ is the weight function.

Nadaraya-Watson esimator

Let K be a real-valued function assigning weights. The function K is usually a symmetric probability density and is called a kernel function. Let h be a bandwidth (also called a smoothing parameter), which is a nonnegative number controlling the size of the local neighborhood. Denote $K_h(\cdot) = K(\cdot/h)/h$. The Nadaraya-Watson kernel regression esimator is given by

$$\hat{m}_h(x_0) = \frac{\sum_{i=1}^{n} K_h(X_i - x_0) Y_i}{\sum_{i=1}^{n} K_h(X_i - x_0)}.$$

Three common examples are the box kernel:

$$K(u) = \frac{1}{2} \cdot \mathbf{1}(|u| \leq 1),$$

the Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\{-u^2/2\},$$

and the Epanechnikov kernel:

$$K(u) = \frac{3}{4}(1 - u^2) \cdot \mathbf{1}(|u| \leq 1).$$

Bandwidth selection

A natural question is how wide the local neighborhood should be so that the local model is valid. This is equivalent to asking how large the bandwidth parameter h should be taken. If we take a very small h, the modelling bias (approximation error) will be small. However,

doi:10.6342/NTU202100742

since the number of data points falling in this local neighborhood is also small, the variance of the estimated local parameter $\hat{m}_h(x_0)$ will be large. On the other hand, if the bandwidth h is large, it can create a large modelling bias and small variance.

As illustrated above, there is a bias and variance trade-off in selecting the bandwidth h. A data-analytic approach is to let data choose a bandwidth that minimizes the risk of an estimated curve. Since

$$E\left[(Y_i - \hat{m}_h(X))^2\right] = E\left[(Y_i - m(X) + m(X) - \hat{m}_h(X))^2\right]$$
$$= E\left[(Y_i - m(X))^2\right] + 2E\left[(Y_i - m(X))(m(X) - \hat{m}_h(X))\right] + E\left[(m(X) - \hat{m}_h(X))^2\right],$$

where $E\left[(Y_i - m(X))^2\right]$ is independent of h, and $E\left[(Y_i - m(X))(m(X) - \hat{m}_h(X))\right]$
$= EE\left[(Y_i - m(X))(m(X) - \hat{m}_h(X))|X\right] = E\left[(m(X) - \hat{m}_h(X))E\left[(Y_i - m(X))|X\right]\right] = 0.$

Hence, the data-driven criterion is defined by optimize the following cross-validation function $\hat{h}_{opt} = \arg\min_h \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_h^{(-i)}(X_i)\right)^2$.

3.3 Local Polynomial Regression

Consider the same model as in Section 3.1, of interest is to estimate the regression function $m(x_0) = E[Y|X = x_0]$ and its derivatives $m'(x_0)$, $m''(x_0)$, $\ldots$, $m^{(p)}(x_0)$. Suppose that the $(p+1)^{th}$ derivative of $m(x)$ at the point x_0 exists. We then approximate the unknown regression function $m(x)$ locally by a polynomial of order p. A Taylor expansion gives, for x in a neighborhood of x_0,

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p.$$

Let $\beta_r = \frac{m^{(r)}(x_0)}{r!}$. This polynomial is fitted locally by a weighted least squares regression problem: minimize

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\beta_j(X_i - x_0)^j\right)^2 K_h(X_i - x_0),$$

where, as in Section 3.1, h is a bandwidth controlling the size of the local neighborhood, and $K_h(\cdot) = K(\cdot/h)/h$ with K a kernel function assigning weights to each datum point.

Denote by $\hat{\beta}_j$, $j = 1, \ldots, p$, the solution to the least squares problem above. It is clear that $\hat{m}_r(x_0) = r!\hat{\beta}_r$ is an estimator for $m^{(r)}(x_0)$, $r = 0,\ 1,\ \ldots,\ p$. To estimate the entire function $m^{(r)}(\cdot)$ we solve the above weighted least squares problem for all points x_0 in the domain of interest. It is more convenient to work with matrix notation. Denote by $\mathbf{X}$ the design matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix},$$

doi:10.6342/NTU202100742

and put

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ and } \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}.$$

Further, let $\mathbf{W}$ be the $n \times n$ diagonal matrix of weights:

$$\mathbf{W} = diag\{K_h(X_i - x_0)\}.$$

Then the weighted least squares problem can be written as

$$\min_{\beta}(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{W}(\mathbf{y} - \mathbf{X}\beta),$$

with $\beta = (\beta_0, \ \ldots, \ \beta_p)^\top$. The solution vector is provided by weighted least squares theory and is given by

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{y}.$$

Hence, the estimator for $m(x_0)$ is given by

$$\hat{m}_h(x_0) = \hat{\beta}_0.$$

3.4 Support Vector Regression

Suppose we have are given training data $\{(x_1, y_1), \cdots, (x_l, y_l)\}$, $X \subseteq R$, where $\mathcal{X}$ denotes the space of the input patterns.

Our goal is to find a function $f(x)$ that has at most deviation from the actually obtained targets y_i for all the training data, and at the same time, is as flat as possible. In other words, we do not care about errors as long as they are less than ϵ, but will not accept any deviation larger than this.

For pedagogical reasons, we begin by describing the case of linear functions f ,taking the form

$$f(x) =< w, x_i > +b, \quad with \quad w \in X, b \in R$$

Formally we can write this problem as a convex optimization problem by requiring:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \parallel w \parallel^2 \\ \text{subject} \quad \text{to} \quad & \begin{cases} y_i - < w, x_i > -b \leq \epsilon \\ < w, x_i > +b - y_i \leq \epsilon \end{cases} \end{aligned}$$

Sometimes, however, this may not be the case, or we also may want to allow for some errors. Analogously to the soft margin loss function, one can introduce slack variables ξ_i to cope with otherwise infeasible constraints of the optimization.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \\ \text{subject} \quad \text{to} \quad & \begin{cases} y_i - < w, x_i > -b \leq \epsilon + \xi_i \\ < w, x_i > +b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned}$$

doi:10.6342/NTU202100742

The constant C determined the trade off between the flatness of f and the amount up to which deviations larger than ϵ are tolerated.

The key idea to slove the otimization problem is to construct the Lagrange function from both the objective function and the coresponding constraints. By introducing a set of varaibles, it can be showed that this function has a saddle point with respect to the primal and dual variables at the optimal solution. As follow:

$$L := \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) - \sum_{i=1}^{l} \alpha_i (\epsilon + \xi_i - y_i + < w, x_i > + b)$$
$$- \sum_{i=1}^{l} \alpha_i^* (\epsilon + \xi_i^* + y_i - < w, x_i > - b) - \sum_{i=1}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*)$$

After using the lagrange function, we can transfer it to a dual optimization problem.

$$\text{maximize} \begin{cases} \frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) < x_i, x_j > \\ -\epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \end{cases}$$
$$\text{subject to} \begin{cases} \sum (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

The next step is to make the SV algorithm nonlinear. This, for instance, could be achieved by simply preprocessing the training patterns x_i by a map $\Phi : \mathcal{X} \mapsto \mathcal{F}$ into some feature space F, and then applying the standard SV algorithm. The dual optimization problem becomes :

$$\text{maximize} \begin{cases} \frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \mathbf{k}(\mathbf{x_i}, \mathbf{x_j}) \\ -\epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \end{cases}$$
$$\text{subject to} \begin{cases} \sum (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

After the procedure, we can get our estimate:

$$\hat{f}(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b$$

4 Simulation

In this section, we simulate three different kinds data. One is classical time series AR(1). Another is brownian motion. The other is nonlinear time series data. Although brownian motion

doi:10.6342/NTU202100742

is not strictly stationary process, we want to test how our methods perform in nonstationary series. Three kinds of data are specialized as follow:

$$X_t = 0.5X_{t-1} + \epsilon$$
$$X_t = X_{t-1} + \epsilon$$
$$X_t = (0.3 + X_{t-1})e^{-4X_{t-1}^2} + \epsilon$$

We simulate 1000 data of each series, use the former 800 to be the training data and the latter 200 be testing data. The badwith we selected is using cross validation. The results are follow:

The first row of graph is using kernel smoothing, the second raw is using local polynomial and the last raw is using support vector regression. The data of second column is brownian motion, it is clear that three methods perform worse when the data is not stationary. Below is the residual sum of square.

	AR process	brownian motion	nonlinear time series
Kernel smoothing	207.7414	non	208.7118
Local polynomial	202.8692	non	208.9454
Support Vector Regression	88.34536	59817.76	240.3199

We can see the results that support vector regression perform well in linear model, lose in nonlinear time series data. It may because the data is not well separated in nonlinear time series and the parameter we selected may also affect the accurancy in support vector regression. Anyway, according to the results, we suggest when facing nonlinear time series data, local polynomial using cubic spline may be a good estimation method; However, facing the AR process data, Support Vector Regression may be a good method.

doi:10.6342/NTU202100742

參考文獻

1. J. L. Kelly, A new interpretation of information rate, Bell System Techn. Journal. 35, 917–926, (1956).
2. J. B. Williams, Speculation and the carryover, The Quarterly Journal of Economics. 50(3), 436–455, (1936).
3. L. Breiman, Investment policies for expanding businesses optimal in a longrun sense, Naval Research Logistics Quarterly. 7(4), 647–651, (1960).
4. L. Breiman, Optimal gambling systems for favorable games, 4th Berkeley Symposium on Probability and Statistics. 1, 65–78, (1961).
5. R. Bellman and R. Kalaba, Dynamic programming and statistical communication theory, Proceedings of the National Academy of Sciences of the United States of America. 43(8), 749–751, (1957).
6. E. J. Elton and M. J. Gruber, On the maximization of the geometric mean with log-normal return distribution, Management Science. 21(4), 483–488, (1974).
7. S. Maier, D. Peterson, and J. V. Weide, A strategy which maximizes the geometric mean return on portfolio investments, Management Science. 23 (10), 1117–1123, (1977).
8. R. Merton, Lifetime portfolio selection under uncertainty: The continuous time case, Review of Economics and Statistics. 51(3), 247–257, (1969).
9. T. Cover, An algorithm for maximizing expected log investment return, IEEE Transactions on Information Theory. 30(2), 369–373, (1984).
10. Bierens, Herman J, The Nadaraya–Watson kernel regression function estimator. *Topics in Advanced Econometrics*. New York: Cambridge University Press. pp. 212–247, (1994).
11. Cleveland, William S.; Devlin, Susan J, Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*. 83(403): 596–610, (1988)
12. D. Basak, S. Pal and D. Patranabis, Support vector regression, Neural Information Processing – Letters and Reviews 11(10) (2007).

doi:10.6342/NTU202100742