Semi-Supervised Learning Based On Nadaraya-Watson Estimator

Wang Junshi

University of Hong Kong
Faculty of Science
Department of Statistics and Actuarial Science

Name: Wang Junshi
University No.: 3035638624
Major: Statistics
Supervisor: Professor LEE, Stephen Man Sing
Research Colloquium for Science UG Students 2021-22

Abstract

In this report, the author will discuss how semi-supervised learning (SSL) can be used in kernel regression, particularly Nadaraya-Watson estimator(NW estimator). SSL here refers to the statistical approach to leverage both labeled and unlabeled to generate better results in terms of mean square error and confidence interval. More specifically, we explore how the mixture of two estimator with different convergence rate may generate a hybrid estimator of faster convergence.

Introduction

The method of SSL is powerful in that it not only focuses on predicting the unobserved points, but also lays emphasis on explore unspecified patterns (Chapelle et al., 2009). This helps boost the performance of estimators when labeled data are sparse and expensive to collect while unlabeled data can be relatively easily obtained. Under the context of NW estimator, the classical estimator and the self-supervised estimator using labeled and unlabeled data will be merged into a hybrid estimator. The asymptotic distribution, mean square error(MSE) and confidence interval(CI) of the hybrid estimator will be calculated to demonstrate the effectiveness of SSL. Finally, simulations will be carried out to visualize the performance of each estimator. We intend to show that the choice of (h, g) is of great importance and the decision depends largely on the objective of research.

Methodology

Supervised Estimator is the estimator that utilize labelled data and adopt the form of Nadaraya-Watson estimator, denoted as $\hat{m}(x)$. And x_i are *iid* random variables with density $p(x)$.

$$NW_{Labeled} = \hat{m}(x) = \frac{\hat{\alpha}(x)}{\hat{p}(x)} \tag{1}$$

$$\hat{\alpha}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} y_i K(\frac{x - x_i}{h_n}) \tag{2}$$

$$\hat{p}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x - x_i}{h_n}) \tag{3}$$

$$y_i = m(x_i) + \epsilon_i,\ \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{4}$$

Self-supervised Estimator leverages unlabelled data. The variables are denoted in a similar way as the case of NW estimator. Note that the value of w_i completely rely on the prediction of previous NW estimator. And under most circumstances, the distribution of u_i is identical to that of x_i.

$$NW_{Unlabled} = \hat{r}(x) = \frac{\hat{\beta}(x)}{\hat{q}(x)} \tag{5}$$

$$\hat{\beta}(x) = \frac{1}{mg_m}\sum_{i=1}^{n} w_i K(\frac{x - u_i}{g_m}) \tag{6}$$

$$\hat{q}(x) = \frac{1}{mg_m}\sum_{i=1}^{n} K(\frac{x - u_i}{g_m}) \tag{7}$$

$$w_i = \hat{m}(x_i) \tag{8}$$

Hybrid Estimator is formulated as a convex combination of the preceding estimators.

To compute the limiting distribution of supervised estimator(converge quicker) and self-supervised estimator(converge slower) and that their mixing actually can generate a hybrid estimator of higher convergence rate, we make use of the **Lyapunoví Central Limit Theorem for Triangular Arrays**. By quoting this theorm, we are able to extract out the asymptotic normal terms with lower order and express $\hat{m}(x)$ and $\hat{r}(x)$ as the following, where $\hat{t}(x)$ and $\hat{s}(x)$ are asymptotic normal random variables with order 1.

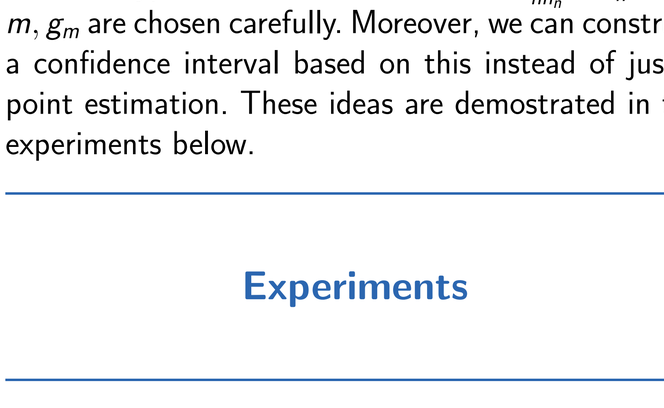
$$\hat{m}(x) = m(x) + (nh_n)^{-1/2}\hat{t}(x) \tag{9}$$

$$\hat{r}(x) = \hat{m}(x) + (mg_m)^{-1/2}\hat{s}(x) \tag{10}$$

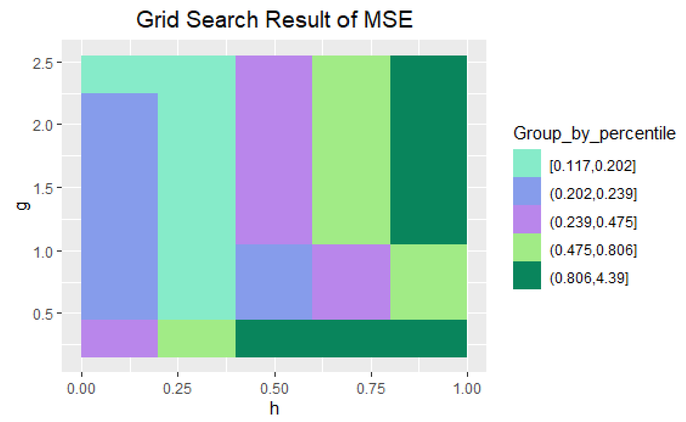
Pushing the induction further, we are able to derive the asymptotic distribution of $\hat{y_c}(x)$. Its mean square error(MSE) is of order $\frac{1}{nh_n} + \frac{h_n^4}{mg_m^3}$, which will smaller than the typical MSE of NW estimator, $\frac{1}{nh_n} + h_n^2$ when m, g_m are chosen carefully. Moreover, we can construct a confidence interval based on this instead of just a point estimation. These ideas are demostrated in the experiments below.

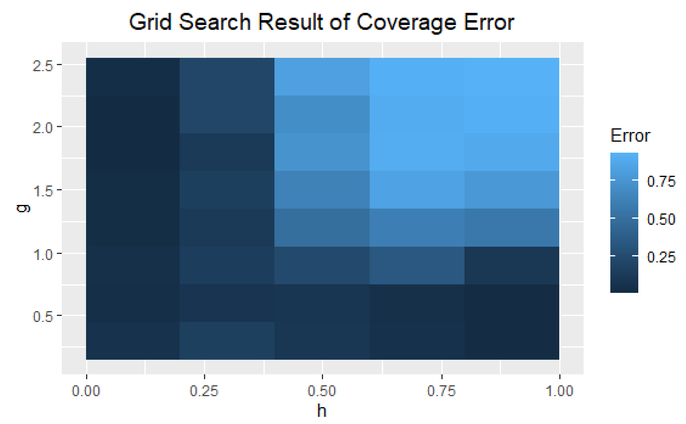
Experiments

Objective of the first experiment is to demonstrate how hybrid estimator behaves under different choice of parameters: $m(x) = x^2$, and $n = 32, 64, 128, m = n^{10/19}$. Grid search for optimal (h, g) is done to find the smallest MSE and coverage error. It seems that the optimal choice of h for hybrid estimator is in very close to that of NW estimator, but still a smaller minimal MSE. This agrees with the formula and expectation.



Just as NW estimator, (h, g) influences the performance of hybrid estimator dramatically. In fact the optimal choice of bandwidth depends on the objective of operation, i.e. whether mean square error or confidence interval is at concern. Based on two subsequent figures, conclusion can be drawn that the pair of (h, g) that provides the smallest mean square error doesn't necessarily secure the best performance in terms of coverage probability.





The simulation unveils two issues and we have made some attempts to address them. In real life situations, grid search is not practical and cross-validation may be adopted to carry out bandwidth selection. Meanwhile, normal approximation in the proposed estimator may not be satisfactory enough given its slow convergence rate. This problem may be avoided by Bootstrap.

Results and Discussion

The main findings are related to the choice of h and g. Firstly, bandwidth selection holds the key to optimal estimation in the proposed hybrid estimator as it is in kernel regression. When chosen wisely, hybrid estimator outperforms NW estimator significantly. Secondly, the choice of (h,g) varies according to the purpose of estimation. The pair of h and g that generates best MSE does not lead us directly to the optimal confidence. Nevertheless, with smallest length, the confidence interval construct using hybrid estimator still gives around 90 % of true coverage and the coverage becomes better as n and m increase.

References

Chapelle, O., Scholkopf, B., Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks, 20(3), 542-542.

Freedman, D. A. (1981). Bootstrapping regression models. The Annals of Statistics, 9(6), 1218-1228.

Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. Theory of Probability Its Applications, 10(1), 186-190.

Wand, M. P., Jones, M. C. (1994). Kernel smoothing. CRC press.

Wied, D., Weißbach, R. (2012). Consistency of the kernel density estimator: a survey. Statistical Papers, 53(1), 1-21.