STA380 Project Proposal

Monte Carlo Study of the Bias–Variance Decomposition in k-NN

Jiachen Chen, Yiwen Zhao, Yuxi Ren, Jintong Li

-"jiachenz.chen@mail.utoronto.ca, ywen.zhao@mail.utoronto.ca"
-"yuxi.ren@mail.utoronto.ca, jintong.li@mail.utoronto.ca"

1 Project Topic

Monte Carlo Study of the Bias–Variance Decomposition in k-Nearest Neighbors Regression. This project builds on the bias–variance framework discussed in classical statistical learning literature (James et al. 2013).

2 Simulation vs. Dataset

This project is based entirely on simulated data. Using simulation allows us to specify the true data-generating mechanism and directly evaluate the bias, variance, and mean squared error of k-NN regression estimators under controlled settings. In particular, simulation makes it possible to isolate the effect of the neighborhood size k, sample size, and noise level on the bias and variance components of the prediction error, which would not be directly observable using real-world datasets (Voss 2013).

3 Project Details

Our project implements a Monte Carlo framework to dissect the Mean Squared Error (MSE) of k-NN. The simulation is structured as follows:

- **Data Generating Process (DGP):** We define $Y = f(X) + \epsilon$, where $X \sim U(0, 1)$ and $\epsilon \sim N(0, \sigma^2)$. The uniform distribution of X provides a standardized domain for evaluating neighborhood density without edge-case distortion (Rizzo 2019).
- **True Functions ($f(X)$):** To maximize the contrast in dimensionality, we utilize two primary functions:
  - **Baseline (1D):** $f(x) = \sin(2\pi x)$, allowing for a clear visualization of the bias and variance decomposition in a simple setting.
  - **Dimensionality Extension (2D):** $f(x_1, x_2) = \sin(\sqrt{x_1^2 + x_2^2})$ to illustrate why k-NN stability degrades as the feature space becomes sparse (OpenAI 2026).
- **MSE Evaluation (Monte Carlo vs. Theoretical):** To address the significance of the "Optimal k," our Shiny app will output two distinct MSE curves (R Core Team, Team, et al. 2024; Chang et al. 2012):
  1. **Monte Carlo MSE:** Calculated by averaging results across $B = 500$ independent simulations.

2. **Theoretical (True) MSE:** Derived from the known DGP components: $MSE = \text{Bias}^2 + \text{Variance} + \sigma^2$.
   - **Significance:** Comparing these two lines demonstrates how Monte Carlo estimates converge to the theoretical truth. The **Optimal** k identified is value that minimizes the total prediction error by decomposing it into bias and variance components.

The following outputs and justifications are provided:

- **Sample Size ($n = 200$):** Chosen to ensure enough local density for k-NN while remaining computationally efficient for the Shiny interface.
- **Repetitions ($B = 500$):** We use 500 independent datasets to compute the expected prediction $E[\hat{f}(x)]$, separating "Bias" from "Variance" (Rizzo 2019).
- **Optimal k Evaluation:** We will plot the total MSE curve to identify the k that reaches the global minimum.

4 User Inputs (Shiny Components)

User will be able to modify the following parameters to observe real-time changesin the bias and variance components of the prediction error:

1. **Selection of the seed:** to ensure reproducibility of specific noisy realizations.
2. **Adjustment of Simulation Parameters:** specifically the number of neighbors k, the noise standard deviation σ, and the number of repetitions B.
3. **Select Dimension:** Users can choose between the univariate function and the bivariate function to examine how model performance changes with dimensionality.
4. **Visual Output Selection:** Ability to choose between:

- The MSE breakdown plot (Bias² vs. Variance),
- The comparison of Monte Carlo MSE vs. True MSE,
- The visual comparison of the estimated fit $\hat{f}(x)$ against the true DGP $f(x)$.

5. **Plot Customization:** Modification of colors for the different error components to enhance clarity.

References

Chang, Winston, Joe Cheng, Joseph J Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2012. "Shiny: Web Application Framework for r." *(No Title)*.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. "An Introduction to Statistical Learning with Applications in r."

OpenAI. 2026. "ChatGPT." https://chat.openai.com.

R Core Team, R, R Core Team, et al. 2024. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2012."

Rizzo, Maria L. 2019. *Statistical Computing with r*. Chapman; Hall/CRC.

Voss, Jochen. 2013. *An Introduction to Statistical Computing: A Simulation-Based Approach*. John Wiley & Sons.