Kernel Regression

Kyle Del Balso \* Swornim Bhandari \* Zack Starrett

# November 21, 2022

**1 Introduction**

Data that assumes a non-linear relationship can be handled through non-parametric regression,

which will help accurately find the conditional expectation between two random variables.

Kernel regression is a popular form of non-parametric regression. Specifically, Kernel regression

uses a kernel smoothing method to estimate a curve for non-parametric data. Kernel estimator

or smoothing is done by taking the weighted average of training samples that surround a

targeted y value. Inside this smoothing parameter is distance-based kernel function, which

helps attach weights to the observations surrounding the y value. In essence, the curve is

estimated by a kernel estimator of the conditional mean. There is also a scale factor h, known

as the bandwidth, which determines area of points surrounding the response value that will be

used for the weighted local averaging. The development and critique of non-parametric kernel

regression estimation is ongoing.

The history of kernel estimators continues to be researched. Nadarayan-Watson Kernel

Estimator, which was founded in 1964, has been considered a very effective approach in

accurately modeling real world data. In 1998, Richard Blundell implements kernel methods in

microeconomics data, including analysis on food and alcohol expenditure for households in the

United Kingdom (Blundell). In 2017, this approach was used to predict monthly electricity

demand production, which aids with appropriately scheduling upkeeping as providing

important information during contract negotiations between energy companies and customers

(Dudek). However, latest works indicate kernel regression can provide meaningful results in

classification scenarios. Support Vector Machines (SVM) is a type of supervised machine

learning method that has been shown to have better performance measurements (sensitivity,

specificity, Area Under the Curve, etc.) than other machine learning methods such as the

Decision Tree and Neural Network models. SVMs solve binary classification problems by

defining a maximum margin separating hyperplane. This hyperplane is then used to classify

training and testing datapoints. To construct the best hyperplane, SVMs need support vectors,

or datapoints around the boundary. For non-linear data, kernels SVMs allow the hyperplane to

be mapped on a higher dimensional space (Awad).

To achieve a classifier, kernel SVM standard model selection practice requires us to select a regularization parameter, C, and a parameter governing the sensitivity of the kernel, γ. Since we do not know the true values of γ and c, we can only achieve the optimal estimates through resampling procedures such as k-fold cross validation. It should also be k-fold cross validation is best suited when both the training and testing datasets are large, and the number of k-folds used ensure that the variance and computational costs tradeoff is minimized (Wainer).

This report presents a two-part study. The first section addresses how to determine the optimal parameters for Nadarayan-Watson Kernel estimator. We will examine the bias-variance tradeoff and conclude the optimal bandwidth using a commonly used cars dataset in R. The second section focuses on kernel SVM, answering questions about how to fine tune the regularization and sensitivity parameters discussed above. We will conclude with a comparative analysis the four different kernel functions provided in R.

**2 Methodology**

**2.1 Nadarayan-Watson Kernel estimator**

Nadarayan-Watson kernel estimator is an important nonparametric kernel estimator of a regression function which is achieved by using a fixed bandwidth. The NW method is a method for estimating unknown parameters of a regression function and is suitable for the situation where the data comes from a joint probability distribution function, f (x, y). The non-parametric regression model for NW method is

= *(1)*

*for (i = 1,……,n)*

Where (.) is unknown, is the sum of the regression function value for and predicted error value is zero.

f (x, y) is the joint density function of (X, Y) and f(x) is the marginal density function of f(x). An estimation of this regression function can be taken as

= *dy (2)*

Using kernel estimation k(.), we have nonparametric kernel estimation of regression.

*(3)*

Using bivariate kernel function and substituting different smoothing parameters (h1, h2) with single fixed smoothing parameter (h) we get the NW kernel estimator of regression.

*(4)*

Smoothing level of estimation also known as bandwidth (h) plays a very important role in the performance of the kernel estimators. The commonly used methods for deciding h are cross-validation, penalized functions, and bootstrap. (Demir & Toktamiş, 2010)

**2.2 Support Vector Machines**

The idea of SVM was established in 1958 by Rosenblatt. SVM is one of the popular and efficient classification and regression methods which separates hyperplane with maximum margin in two-dimensional space which are linearly separable. The one that maximizes the distance to the closest data points from both classes is called hyperplane with maximum margin. However, for a non-linear data SVM finds it difficult to classify the data. That’s where we can use the Kernel Trick. A Kernel trick is a method where a non-linear data is projected onto a higher dimension space to make it easier to classify the data where it could be linearly divided by a plane (Karatzoglou, Meyer, & Hornik, 2006). SVMs use an implicit mapping Φ of the input data into a high-dimensional feature space. The common classification model is

*(5)*

Where *b* is the bias of the model.

Solving the subsequent constrained optimization and creating decision boundary, we get

*(6)*

*Subject to*

Where (i = 1,…..m),

As it is constrained optimization problem it needs Lagrangian formula or Lagrangian multipliers. Firstly, we need to find decision boundary for linearly separable plane and then solve with Lagrange multipliers. We get following equation,

subject and for all *i.*

As there are different types of kernel function for different type of data, picking right kernel and parameters can be the challenge for SVMs.

**2.3 Types of Kernel function**

There are four main types of kernels which are listed below with equations.

**2.3.1 Linear Kernel Function**

The linear kernel is the basic and useful when dealing with large sparse data vectors. Linear Kernel is the inner product of two vectors. The equation is:

*(8)*

where c is the constant.

**2.3.2 Polynomial Kernel Function**

The polynomial kernel is commonly used function with SVMs. It represents kernels with a degree of more than one. It helps on learning of non-linear models and suitable for image processing. (Vidvan, 2020). There are two types of polynomials:

1. **Homogenous Polynomial Kernel Function**

*(9)*

where d is the degree of the polynomial.

1. **Inhomogeneous Polynomial Kernel Function**

*(10)*

where c is a constant and d is degree of polynomial.

**2.3.3 Gaussian RBF Kernel Function**

Gaussian RBF is used when there is no prior knowledge about the data.

*(11)*

where

**2.3.4 Sigmoid Kernel Function**

This is mainly used in neural networks. We can show this as:

*(12)*

where h and c are parameters to optimize.

**3 Experiment**

* 1. **Experiment Background**

Attempting to obtain the best estimated curve for Nadarayan-Watson Kernel estimator requires an input of the bandwidth h. A higher bandwidth will include more data points for the weighted average, while still giving the highest weights to the closer data points. Choosing a bandwidth is vital in finding a function estimate that fits the observed sample of data while providing a smooth function estimate (Helwig). Too small a bandwidth, and the estimated curve will overfit the data, which will negatively affect the accuracy of predicting future observations. Too large a bandwidth, and the estimated curve will underfit the data, which results in the estimated curve missing the true relationship between the input values. The question to analyze becomes clear; How do we optimize the bandwidth, so our estimated curve minimizes the negative tradeoff effects?

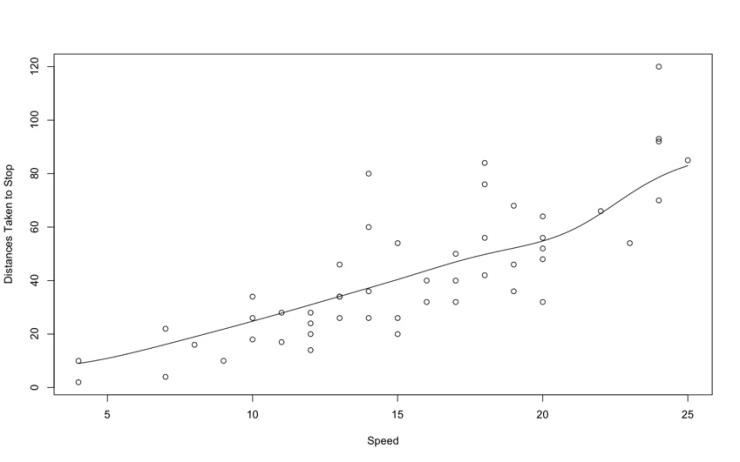
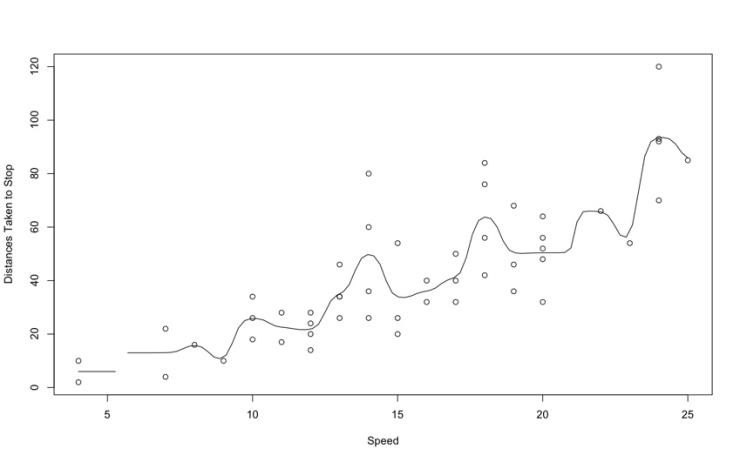
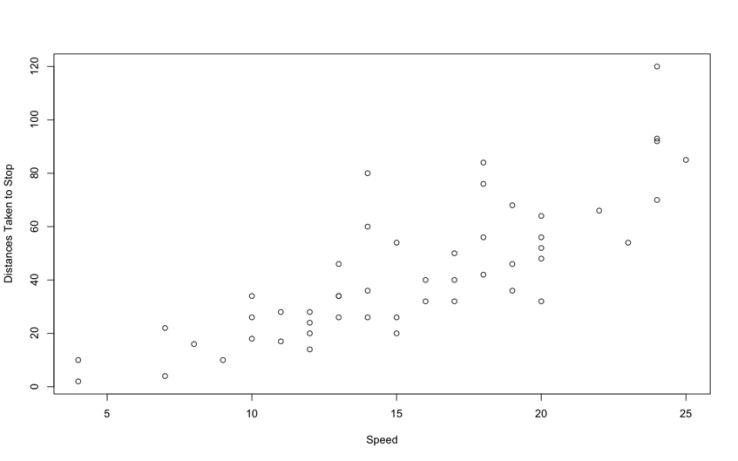
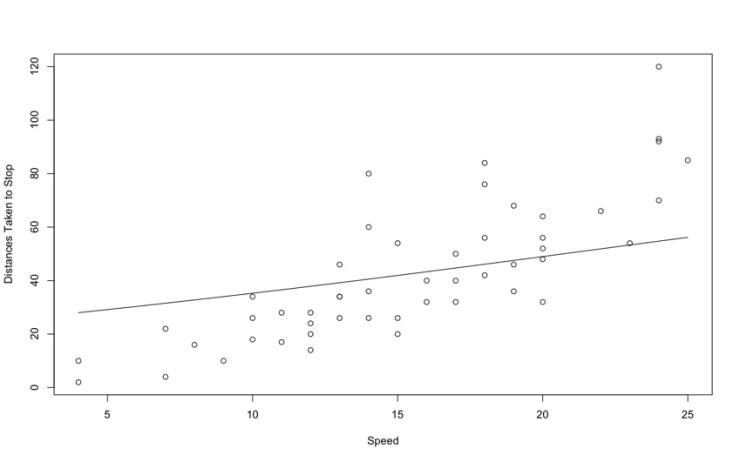
We also wish to examine the impact of kernels for classification, specifically kernel SVM models. We will use R programming language to design four kernel SVM models on the Algerian Forest Fire dataset: linear, polynomial, radial basis function, and sigmoid. We will discover the optimal parameter for each kernel using a tuning parameter in R. Once we obtain the best models for each kernel, we will run a comparative analysis to determine which is superior in its accuracy on the testing set.

* 1. **Data sets for experimentation** 
     + Cars dataset provides a 1920’s dataset containing 50 observations. There are two variables in the dataset: the X variable is the speed of cars, and the Y variable is the distance taken to stop(Ezekiel).
     + Algerian Forest Fires Dataset for the Bajaia region located in the northeast of Algeria, which includes 122 observations. The dataset has a binary response variable that indicates whether the Bajai region of Algeria experienced no fire = 0, or a fire = 1. The dataset contains 10 explanatory variables including temperature, relative humidity, wind speed, rain, fine fuel moisture code, duff moisture code, drought code, initial spread index, and fire weather index (Dua). For the Algerian Forest Fires Dataset, we will split the data into training and testing sets. Specifically, 75% of the data will go into the training set, while the remaining 25% will be used for testing the accuracy of the model.

**4 Examples and Results**

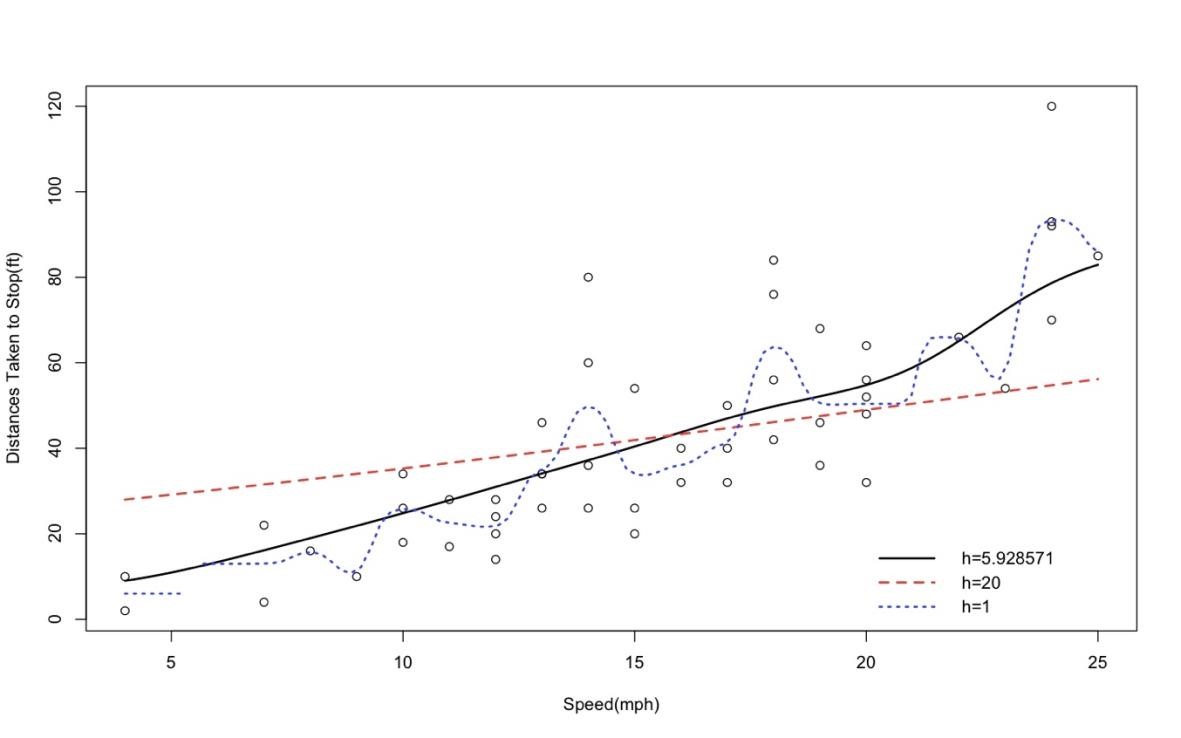
**4.1 Cars Example**

Using the cars dataset in R, we obtain four plots in Figure 4.1. Each regression line is obtained using the ksmooth() function. The top left figure is a scatter plot between the predictor variable speed and the response variable stopping distance. The data appears to be normal when speed < 15. However, the data when speed > 15 appears to have a non-linear relationship with stopping distance. Since the data illustrates a non-linear relationship, nonparametric kernel regression is used to determine the regression curve. The figure on the top right shows the regression curve estimation with a bandwidth = 20. Although this result produces a stable line with little variance, the bias tradeoff is too high as the line does not fit the data well. The figure on the bottom left shows the regression curve estimation with a bandwidth = 1. The line is too rigid and overfits the data.



# Figure 1: The graph of the data with different bandwidths(Helwig)

The optimal bandwidth is obtained through generalized cross-validation using the np.gcv() function with a selected “gaussian” kernel (Kauermann). The generalized cross-validation procedure gives us an optimal bandwidth value of 5.928571. The bottom right figure shows the regression curve estimation with the obtained optimal bandwidth = 5.928571. The line now is smooth and fits the data appropriately. Figure 4.2 shows the three chosen bandwidths together.



# Figure 4.2: The combined graph of the different bandwidths (Helwig)

The experiment on real world data is a fruitful process to analyze the impact the effect of choosing an appropriate bandwidth. In the car’s dataset, we have no assumptions about the data, so we needed to figure out a bandwidth that optimizes the variance-bias tradeoff. Too small a bandwidth and we have too much variance. Too large a bandwidth and we don’t cover all the data. Therefore, finding the optimal bandwidth is crucial in determining the best regression curve. Even though R uses a default bandwidth of h = 0.5, that does not indicate that the optimal bandwidth is so. There is no uniform way to obtain this metric, and in our experiment, we used a cross-validation technique to approximate the best regression curve (Kauermann).

**References**

Awad, M., & Khanna, R. (2015). Support vector machines for classification. In *Efficient learning machines* (pp. 39-66). Apress, Berkeley, CA.

Blundell, R., & Duncan, A. (1998). Kernel regression in empirical microeconomics. *Journal of Human Resources*, 62-87.

Demir, S., & Toktamiş, Ö. (2010). On the adaptive Nadaraya-Watson kernel regression estimators. *Hacettepe Journal of Mathematics and Statistics*, *39*(3), 429-437.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Dudek, G., & Pelka, P. (2017, May). Medium-term electric energy demand forecasting using NadarayaWatson estimator. In *2017 18th International Scientific Conference on Electric Power Engineering (EPE)* (pp. 1-6). IEEE.

Ezekiel, M. (1930) Methods of Correlation Analysis. Wiley.

Heiwig, N.(2021). Nonparametric Regression (Smoothers) in R. *Department of Psychology & School of Statistics University of Minnesota.*

Kauermann, G., & Opsomer, J. D. (2004). Generalized cross-validation for bandwidth selection of backfitting estimates in generalized additive models. *Journal of Computational and Graphical Statistics*, *13*(1), 66-89.

Wainer, J., & Cawley, G. (2017). Empirical evaluation of resampling procedures for optimising SVM hyperparameters. *Journal of Machine Learning Research*, *18*(15), 1-35.

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support Vector Machines in*R*. *Journal of Statistical Software*, *15*(9). <https://doi.org/10.18637/jss.v015.i09>

Vidvan, T. (2020, March 13). SVM Kernel Functions - 'SVM knowledge. Retrieved from TechVidvan website: https://techvidvan.com/tutorials/svm-kernel-functions/