MATHEMATICAL FOUNDATIONS OF SUPERVISED LEARNING METHODS AND APPLICATIONS ON DISEASE MODELING

Math 490 - Graduation Project

2024-2025 Spring Semester

Author
Zeynep Nur KARABAY

Supervisor
Assoc. Prof. Dr Özlem DEFTERLİ

Department of Mathematics

Çankaya University

Abstract

In this study, we investigate the mathematical principles that underpin supervised learning methodologies and their real-world applications in disease modeling, with a specific emphasis on predicting measles cases. We implement and compare several regression-based models, including linear regression, polynomial regression, random forest, and XGBoost models, to evaluate their performance against historical measles incidence data and to assess their effectiveness in forecasting future disease outbreaks. Our results emphasize both the strengths and limitations of each model. The results illustrate the challenges inherent in modeling noisy, real-world disease data while highlighting the advantages and disadvantages of each method from both predictive and theoretical perspectives. This study aims to emphasize a deeper understanding of how mathematical principles support the use and performance of supervised learning algorithms in disease modeling for measles cases.

Keywords: Supervised Learning, Mathematical Foundations, Compare Models, Disease Modeling, Time Series Forecasting, Regression Analysis, Measles, XGBoost, Random Forest, Polynomial Regression, Linear Regression

Contents

List of Figures

1 Introduction

1.1 Motivation and Research Objectives

Greater access to large datasets and advances in computational methods have transformed the field of disease modeling [1][1] However, it brings challenges in analyzing large datasets. With the help of Supervised learning techniques, which allow for the prediction of outcomes grounded on historical data, these tools have become essential in epidemiology for forecasting disease incidence and supporting public health decision-making.[2][2]This is the prior motivation for this study. Measles is an infectious illness brought on by a contagion. Although it is a disease that can be prevented by vaccination, it spreads easily when someone who is infected coughs, breathes, or sneezes. It can lead to serious illness, complications, and potentially death. Measles remains to pose influential public health challenges worldwide due to factors such as variability in vaccination scope and occasional outbreaks. Accurate modeling and prediction of measles case counts can prop in resource allocation, early warning systems, and plan targeted interventions.[3] [3].

This study examines the mathematical foundations of several supervised learning methods, including Linear Regression, Polynomial Regression, Random Forest, and XGBoost. It applies them to the problem of disease modeling using real-world measles case data. The study is not limited to implementation but also emphasizes the comparative evaluation of these models in a time-series forecasting context.

The primary objectives of this work are as follows:

- To provide a comprehensive mathematical background for each supervised learning algorithm.
- To implement and compare these models using historical measles incidence data.
- To evaluate the performance of each supervised learning method in terms of predictive accuracy, interpretability, and suitability for epidemiological forecasting.

[1]X. Li, J. Qiu, L. Lin, and B. Yin, "Machine learning in epidemiology: applications and challenges"

[2]A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare"

[3]Who measles data

- To discuss practical considerations in modeling time series data, including the incorporation of lag features.
- To forecast future measles case numbers (e.g., 2027) and analyze model strengths and weaknesses.

1.2 Overview of Supervised Learning in Disease Modeling

Supervised learning has become a robust method in predictive analytics, allowing the development of models that correlate input features with target outputs using labeled training data. When supervised learning techniques applied to disease modeling, these methods facilitate the learning of historical epidemiological trends and the prediction of future developments, which in turn guides public health policy and intervention strategies.[1][4] [2][5] [4][6]

Regression-based supervised learning algorithms are particularly valuable for modeling continuous variables, such as disease incidence counts. Linear regression provides a straightforward and easy-to-understand method approach based on the connection of a linear relationship between predictors and the outcome. In contrast, polynomial regression improves this model by representing non-linear patterns through the inclusion of higher-order terms. Furthermore, ensemble methods like Random Forest [5][7] and gradient boosting techniques, such as XGBoost [6][8], employ several decision trees to enhance predictive accuracy and robustness.

Consider for temporal relationships is crucial when implementing supervised learning techniques to time-series disease modeling. One common approach involves the use of lag features—past observations used as input for current predictions—which help capture underlying temporal patterns more

[4]X. Li, J. Qiu, L. Lin, and B. Yin, Machine learning in epidemiology: applications and challenges

[5]A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare"

[6]D.Bzdok "statistics versus machine learning"

[7]L. Breiman, "Random forests"

[8]T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system"

effectively.[7][9] [8][10] , [9][11]

Despite the natural noise and inconsistencies present in actual health data, these methods have reliably shown robust performance in different epidemiological contexts, which includes forecasting infectious disease outbreaks.

1.3 Structure

This study is structured to progressively develop both the theoretical and experimental aspects of supervised learning methods applied to disease modeling for measles cases.[10][12][3] [13] The chapters are organized as follows:

- **Chapter 1** Introduces the research motivation, objectives, and provides a brief overview of supervised learning in the context of epidemiology.
- **Chapter 2** Presents the mathematical foundations of the selected supervised learning algorithms, including linear regression, polynomial regression, random forest, and XGBoost, along with commonly used assessment measures such as Mean Squared Error (MSE) and the coefficient of determination (R^2).
- **Chapter 3** Describes the dataset utilized in this research. Including its source and the preprocessing steps applied. Details the implementation of each model, the use of lag features for time series considerations, model evaluation, and comparison of predictive performances. This chapter also includes forecasting of future measles cases (e.g., for the year 2027).
- **Chapter 4** Wraps up the study by summarizing the main findings.

This research provides an in-depth understanding of the mathematical foundations of supervised learning models and their applications in real-world disease prediction scenarios.

[9]Box, G.E.P., Jenkins, G.M., Reinsel, G.C., and Ljung, G.M. (2015)"Time Series Analysis: Forecasting and Control"

[10]P. J. Brockwell and R. A. Davis."Introduction to Time Series and Forecasting"

[11]Hyndman, R.J., Athanasopoulos, G. (2018) "Forecasting: Principles and Practice"

[12]T. Hastie, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"

[13]Who measles data

2 Mathematical Foundations of Supervised Learning

This section provides an overview of the supervised learning methods used in this study. Each algorithm will be presented along with its mathematical formulation and underlying assumptions. We also present the evaluation metrics utilized to assess their performance.

2.1 Linear Regression

Linear regression is one of the most fundamental and widely used regression techniques. It assumes a linear correlation between the predictor (independent) variable and the response (dependent) variable. By reducing the sum of squared residuals, it determines the coefficients that best fit the data. Linear regression is especially effective when the relationship among variables is linear and there are no significant outliers [11][14].

The equation of the regression line can represent the model:

$$\hat{y}_i = \beta_0 + \beta_1 k_i, \tag{2.1}$$

where $k_i \in \mathbb{R}$ is the independent variable (predictor), and $\hat{y}_i \in \mathbb{R}$ is the predicted value of the dependent variable (response).

2.1.1 Model Parameters and Sample Means

To compute the parameters β_0 and β_1, we use the sample means and covariances:

- $\hat{y}_i$ is the predicted value of the response variable,
- $\beta_1 = \frac{S_{ky}}{S_{kk}} = \frac{\sum(k_i-\bar{k})(y_i-\bar{y})}{\sum(k_i-\bar{k})^2} = \frac{\sum k_i y_i - \frac{\sum k_i \sum y_i}{n}}{\sum k_i^2 - \frac{(\sum k_i)^2}{n}}$,
- $\beta_0 = \bar{y} - \beta_1 \bar{k}$,
- $\bar{k} = \frac{\sum k_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}$,

[14] A. Hasudungan, "A Comprehensive Analysis of Regression Algorithms in Machine Learning"

where n is the number of observations [12][15].

The underlying statistical model assumes a linear relationship with an error term ϵ accounting for the deviation:

$$y_i = \beta_0 + \beta_1 k_i + \epsilon_i, \tag{2.2}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is assumed to be i.i.d. noise [13][16].

In linear regression models, the error terms ϵ_i are commonly assumed to be independently and identically distributed (i.i.d.) random variables. This implies two key statistical properties:

- **Independence**: Each error term ϵ_i is statistically independent from others, i.e., $\epsilon_i \perp \epsilon_j$ for $i \neq j$.
- **Identical Distribution**: All error terms follow the same probability distribution, typically assumed to be Gaussian with zero mean and constant variance: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

2.1.2 Least Squares Estimation

The Ordinary Least Squares (OLS) technique makes estimates regarding the coefficients β_0 and β_1 with reducing the Residual Sum of Squares (RSS), which defined as:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 k_i)^2.$$

Let the residual be $e_i = y_i - \hat{y}_i$. Then,

$$e_i = y_i - (\beta_0 + \beta_1 k_i).$$

To find the optimal parameters, we take partial derivatives of RSS with respect to β_0 and β_1 and set them to zero:

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 k_i) = 0,$$

[15] Newcastle University, "Simple Linear Regression"

[16] Statstutor, "Simple Linear Regression"

$$\frac{\partial RSS}{\partial \beta_1} = -2\sum_{i=1}^{n} k_i(y_i - \beta_0 - \beta_1 k_i) = 0.$$

Solving this system gives the least squares estimators:

$$\beta_1 = \frac{\sum(k_i - \bar{k})(y_i - \bar{y})}{\sum(k_i - \bar{k})^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{k}.$$

The Ordinary Least Squares (OLS) estimators provide the Best Linear Unbiased Estimators (BLUE) of the regression coefficients under the classical linear regression assumptions, including linearity, independence, and normality of errors [10][17] [14][18] [15][19]. While these assumptions hold well for linear regression, more complex supervised learning methods such as Random Forest and XGBoost do not rely on BLUE properties but instead optimize prediction accuracy via different approaches. Hence, BLUE is primarily relevant in the mathematical foundation of linear regression models.

2.2 Polynomial Regression

Polynomial regression is built upon linear regression by incorporating polynomial terms to represent non-linear interactions between variables more effectively. Polynomial regression approach offers enhanced flexibility for modeling complex data patterns. However, it may be sensitive to overfitting, particularly with high-degree polynomials. Therefore, it is crucial to apply regularization techniques to address this problem.[11][20] Polynomial regression models the relationship between a scalar predictor variable z and the response variable y as an n-th degree polynomial:[10][21]

$$\hat{y} = b_0 + b_1 z + b_2 z^2 + \cdots + b_n z^n = \sum_{i=0}^{n} b_i z^i [13][22]$$

[17] T. Hastie, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"

[18] D. C. Montgomery, E. A. Peck, and G. G. Vining,"Introduction to Linear Regression Analysis"

[19] W. H. Greene,"Econometric Analysis"

[20] A. Hasudungan, "A Comprehensive Analysis of Regression Algorithms in Machine Learning"

[21] T. Hastie, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"

[22] Statstutor, "Simple Linear Regression"

The goal is to reduce the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} \left(y_i - \sum_{j=0}^{n} b_j z_{ij} \right)^2$$

This can be represented using matrix notation as:

$$Z = \begin{bmatrix} 1 & z_1 & z_1^2 & \cdots & z_1^n \\ 1 & z_2 & z_2^2 & \cdots & z_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_m & z_m^2 & \cdots & z_m^n \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

Then, the optimal coefficient vector $\vec{b}$ that minimizes the standard equation gives the squared error:

$$\vec{b} = (Z^\top Z)^{-1} Z^\top \vec{y} \text{ [16]}^{23}$$

This closed-form solution is identical to the one utilized in linear regression, where polynomial terms are considered separate features.

2.3 Random Forest Regression

Random forest regression is a group technique that merges several decision trees to produce forecasts. By averaging the predictions made during training along with outputs from each of the individual trees, random forest regression declines overfitting and enhances prediction accuracy. Even, it offers measures of feature importance, facilitating variable selection. Nevertheless, random forest regression can experience significant computational demands and may lack interpretability. [5][24].

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x)$$

where $T_m(x)$ is the prediction of the m-th regression tree.

This method is especially effective in managing high-dimensional datasets and capturing non-linear relationships without requiring extensive hyperparameter tuning.

[23] C. E. Rasmussen and C. K. I. Williams, "Gaussian Processes for Machine Learning"
[24] L. Breiman, "Random forests"

2.4 XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is a scalable and regularized gradient-boosting framework designed for high performance and efficiency [6][25]. It builds models in an additive fashion, optimizing a regularized loss function:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(m_i), \quad f_m \in \mathcal{F}$$

where $\mathcal{F}$ is the space of regression trees. The overall objective function is defined as:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{m=1}^{m} \Omega(f_m)$$

with a regularization term $\Omega(f_m) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ to control model complexity and prevent overfitting.

2.5 Evaluation Metrics (MSE, R^2)

To evaluate the effectiveness of regression models, the subsequent metrics are commonly used [10][26]:

- **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{k} \sum_{j=1}^{k} (y_j - \hat{y}_j)^2 [17][27]$$

  where y_j is the actual value, $\hat{y}_j$ is the predicted value, and n is indicates the count of observations. A lower MSE signifies greater predictive accuracy, as it penalizes larger errors more heavily.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system"

[26] T. Hastie, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"

[27] Yu, W., Song, W., Xue, B., Zhang, M. (2023). "Surrogate-assisted Multi-objective Optimization via Genetic Programming Based Symbolic Regression"

- **Coefficient of Determination (R^2)**:

$$R^2 = 1 - \frac{\sum_{j=1}^{k}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{k}(y_j - \bar{y})^2}[17]^{28}$$

where $\bar{y}$ is the mean of the actual values. R^2 computes the proportion of variance in the predictor variable y that is predictable from the response variable. An R^2 value that approaches 1 signifies a more accurate fit.

[28] Yu, W., Song, W., Xue, B., Zhang, M. (2023). "Surrogate-assisted Multi-objective Optimization via Genetic Programming Based Symbolic Regression"

5 References

References

[1] X. Li, J. Qiu, L. Lin, and B. Yin, "Machine learning in epidemiology: applications and challenges," *Computers in Biology and Medicine*, vol. 103, pp. 1–11, 2018.

[2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[3] World Health Organization, "Measles – Global situation and challenges," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/measles

[4] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nature Methods*, vol. 15, no. 4, pp. 233–234, 2018.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., and Ljung, G.M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.

[8] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 3rd ed., Springer, 2016.

[9] Hyndman, R.J., Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts. Retrieved from https://otexts.com/fpp2/

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[11] A. Hasudungan, "A Comprehensive Analysis of Regression Algorithms in Machine Learning," 2024. [Online]. Available:http:/andrehasudungan.blog.uma.ac.id/2024/02/05/a-comprehensive-analysis-of-regression-algorithms-in-machine-learning/

[12] Newcastle University, "Simple Linear Regression," *Maths Support – Statistics Resources*, [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/simple-linear-regression.html.

[13] Statstutor, "Simple Linear Regression," *Statstutor resources*, [Online]. Available: https://www.statstutor.ac.uk/resources/uploaded/simplelinearregression.pdf.

[14] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed., Wiley, Hoboken, NJ, 2012.

[15] W. H. Greene, *Econometric Analysis*, 8th ed., Pearson, 2018.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006. [Online]. Available: https://gaussianprocess.org/gpml/chapters/RW.pdf

[17] Yu, W., Song, W., Xue, B., Zhang, M. (2023). Surrogate-assisted Multi-objective Optimization via Genetic Programming Based Symbolic Regression.In *Genetic Programming* (pp. 199–216). Springer, Cham. https://doi.org/10.1007/978-3-031-27250-9_13

[18] Our World in Data. (2025). Reported cases of measles. *Our World in Data*. Retrieved from https://ourworldindata.org/grapher/reported-cases-of-measles