2025

CS 4407
Data Mining & Machine Learning

LEARNING JOURNAL UNIT 3

SANA UR REHMAN

INSTRUCTOR: NIRMAL ADHIKARI

Introduction

This week's focus on regression deepened my understanding of supervised learning and how past data can be used to predict future outcomes. I explored linear, curvilinear, and logistic regression, learning how each method handles different relationships between variables. Through R programming exercises, I practiced plotting data points, fitting lines, and minimizing error to build accurate predictive models. The lessons reinforced key concepts such as model assumptions and the importance of validating predictions within the observed data range (James et al., 2021).

Activities Performed

I began by reviewing course materials on regression fundamentals and their practical applications. For the discussion activity, I analyzed the limitations of regression analysis, highlighting issues like assumption violations, extrapolation risks, and multicollinearity, and proposed strategies for dynamic model updates in fast-changing environments.

The programming assignment required implementing multiple linear regression in R. I calculated parameter estimates and 95% confidence intervals, then reduced the model by removing a non-significant variable (x2). Residual analysis confirmed normality, and I created fitted line plots with confidence and prediction intervals. I also ran a simulation with small datasets, testing simple linear models that showed weak predictive power. Finally, I completed the self-quiz to reinforce key concepts.

Difficulties Faced

The most challenging part was interpreting diagnostic plots and understanding how assumption violations affect model reliability. It took several attempts to correctly interpret residual patterns and confirm normality. Additionally, deciding when to reduce a model without losing important predictors required careful consideration of p-values and confidence intervals.

Reactions and Feedback

I felt both challenged and motivated during the R programming tasks. The assignment's hands-on approach clarified abstract concepts from the readings. Feedback from peers on my discussion post emphasized the importance of considering real-world data complexity and supported my point about the dangers of extrapolation beyond observed ranges. Their comments helped me see how regression models can be strengthened by combining different approaches (Kutner et al., 2005).

Reflections and Learning

I was surprised by how sensitive regression models are to small changes in data and assumptions. Understanding the difference between linear, curvilinear, and logistic regression expanded my perspective on selecting the right model for a dataset. I recognize that I am improving in R coding and statistical interpretation. This week reinforced that I learn best by applying concepts through practice and reflection.

One important realization is the need for critical thinking when using regression for prediction. The work highlighted that even strong R^2 values can be misleading if the underlying assumptions are not met (James et al., 2021). I can apply these ideas to future data analysis projects by rigorously testing model assumptions and carefully defining prediction ranges.

Conclusion

Week 3 strengthened my ability to implement and evaluate regression models. I improved my R skills, gained confidence in interpreting outputs, and developed a deeper appreciation for the limitations of predictive modeling. Going forward, I will focus on validating models and considering alternative approaches when assumptions are questionable.

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.

Wordcount: 485