Spline Model Notes and Observations

Ben Browder

December 13, 2021

1 Notes and Observations

- I spent Thursday, Friday, and part of Saturday on the last weekend working on the code searching for a Spline method or concept that could help me classify the Holling Type II datasets. I wasn't able to find anything that could help me, so if someone were to create a method or model to use Splines to classify function-valued traits, they would most likely be the first to present it, at least on the internet. Since I wasn't able to find anything on the internet I shifted my focus on simply trying to classify the simulated datasets without worrying to much about the limitations the model would face in the real world.

- There are several limitations within the model that are unrelated to the unsupervised dataset. Possibly the most glaring limitation is that the dataset that was trained on was split 50% 50% between the control and experimental datasets, in the real world, it would more likely be a 90% and 10% split, which would make the method in which was I calculated the average Spline coefficients and threshold very skewed towards the 90%. The model still could possibly be used to classify. If there were a way in which someone could accurately create the Spline coefficients and threshold to classify, the model could be very effective.

- One interesting thing I noticed when playing around with the model and tests is I figured that the average intercept value of 0.3 didn't have much of an effect when classifying the datasets since the p-value of the intercept in the linear regression table was 0.9. However, when I took out the intercept, the accuracy scores went down to 50%. The intercept scores increased as the curve coefficients increased as well. However, the intercepts would vary from positive and negative values for all curves. Since the average intercept was 0.3, it would classify all values below 0.3 and negative as traits affecting survival. I didn't have time to look into why this was the case considering the p-value was 0.9. It might be interesting for someone looking at a different way of classifying the dataset using the Spline coefficient to look into the effects of the intercept.

- Another topic I should address is Dr. Childs's question in one of the client meetings on this type of Spline classification being a form of looking at times of death to classify. Generally, the organisms who didn't live long would have a linear plot, and those who lived longer would have a polynomial plot, both of which would be represented by the Spline coefficients. However, because our simulated dataset model used stochastical modeling, the growth of each organism's Holling Type II curve would vary depending on the a and b values. For example, a comparison of two polynomial curves could be if there was an organism who had a rapid rate of growth for the trait and died at timestep 20, compared to an organism who had a slower rate of growth who died at timestep 30 and finished with a lower final trait value than the first organism. With Linear Interpolation applied to the plot, it would seem that the first organism lived a longer lifespan since it had a more obvious Holling Type II curve and finished with a higher function valued trait when in reality, the second organism lived longer. This type of example happened throughout the dataset, which made the comparison of Spline coefficients not exactly correlated to the times of death.

- I finished the Spline model on Saturday night with the expectation that I would stop coding on Sunday, which didn't give me time to attempt a model for Sinusoidal data. The things that would need to be changed is to have the degrees of freedom be chosen independently for each organism instead of having a base degree of freedom of 4 for the entire dataset since the number

of knots for each organism would vary depending on how long the organism lived. Even with that implemented, it would still be difficult to compare the coefficients since the organisms would have a different amount of coefficients. Maybe someone could compare the number of knots and degrees of freedom between the organisms in order to classify?

- If given more time, I would have attempted to find a pattern between the Spline coefficients within the datasets to determine a threshold amount instead of just getting the percentage amount above the average Spline intercept and coefficient values. Just by looking at the plots and coefficients there might be a pattern around the 20% mark that could be used to classify. If a pattern was found it would be a bit more flexible than the method I'm currently using since it wouldn't rely on the training data and instead a certain pattern in Holling Type 2 curves that indicate whether the dataset was or wasn't affected by the trait. I can't say whether this could possibly be done or if it would be applicable in the real world due to the unsupervised dataset, but it would be interesting to look at.