Siyu
March 5th, 2024

Hi everyone, I'm excited to share that I'm diving into learning Stan for data mining! Stan is definitely a unique tool within the Python scientific stack compared to other ML libraries I've encountered. In a way, it feels more like statistical modeling with detailed control over parameters and functions, as opposed to traditional machine learning.

I build models directly from mathematical equations, which involves a cool translation process between models and those equations! I also appreciate how Stan allows us to incorporate various distributions into priors and utilize efficient sampling through Markov Chain Monte Carlo (MCMC).

My first modeling project focuses on estimating the weights of female and male golden retrievers. We have weight measurements for 500 dogs, but unfortunately, with one crucial piece of information missing: we don't know the genders (sex) of the dogs. This makes it impossible to determine the number of females and males in the sample.

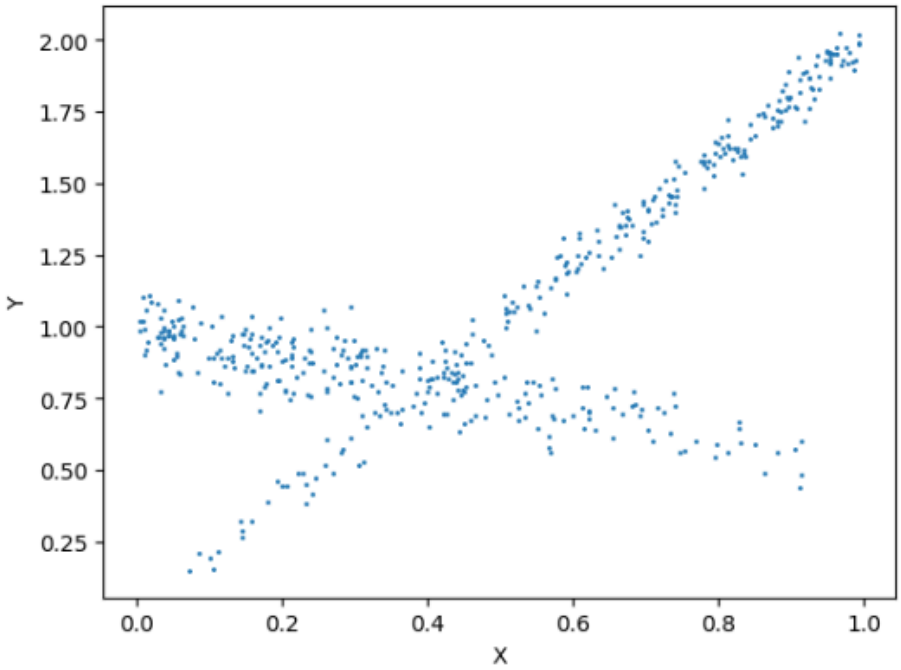
I'm developing a Stan model for this dataset, assuming that the weights of both female and male dogs follow Gaussian distributions. Additionally, the model aims to achieve the following:

- Estimate weights for female and male golden retrievers with 95% confidence intervals.
- Provide an estimate for the ratio of female and male dogs in the sample.

The Stan model I'm building assumes that the observed weights come from a mixture of two different normal distributions, potentially representing the two subpopulations of female and male dogs. The model will infer the mean, standard deviation of weights, and the proportion of the overall population each subpopulation represents.

When considering broader applications, for example, in the medical field, it's important to recognize ethical considerations regarding data collection. However, if anonymized data, such as heart disease rates, can be collected regardless of their ethical origin, similar models could be used to infer disease rates among different populations, ultimately aiding public health initiatives.

The second project focuses on translating functions that can capture an abnormal data pattern with mathematics notations into Stan code.



Pretty weird. However, the following model should fit this quite well:

$$
\begin{aligned}
y &\sim \begin{cases} w_1 & \text{if } z = 0 \\ w_2 & \text{if } z = 1 \end{cases} \\
w_1 &\sim \text{Normal}(\beta_1 \cdot x + \alpha_1, \sigma_1) \\
w_2 &\sim \text{Normal}(\beta_2 \cdot x + \alpha_2, \sigma_2) \\
z &\sim \text{Bernoulli}(x) \\
\beta_1, \beta_2, \alpha_1, \alpha_2 &\sim \text{Normal}(0, 1) \\
\sigma_1, \sigma_2 &\sim \text{InverseGamma}(1, 1)
\end{aligned}
$$

where we assumed that $x \in (0, 1)$.

This assignment has made me realize the power of Stan in statistical modeling. I can mirror the same structure in priors and functions (as much as we want)into Stan code and use MCMC for efficient sampling.