Machine Learning - Homework 6

Tim Kmecl, 22.5.2024

Implementation

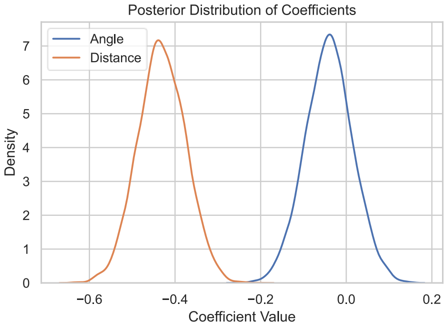
I used python libraries *pymc* and bayesian logistic regression model from *bambi*. Default priors are normal with standard deviation of 2.5. I used markov chain monte carlo with 4 chains, each with 1000 tuning samples and then 2500 draws for a total of 10,000 final samples. The data was standardized beforehand both for the whole dataset and the small subset.

Prior opinion

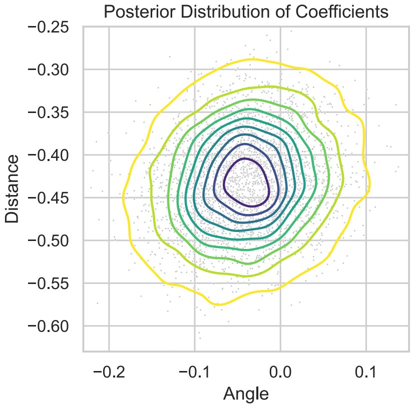
In basketball, statistics for players usually show higher percentage of 2pt shots made than 3pt, which are thrown from larger distance. We may assume that shots made from smaller distance are easier and therefore more likely to succeed, which implies a negative coefficient for distance. There is, however, a possibility that it could be positive, since shots directly under the basket (too close) are less likely to suceed, moreover some players might practice free throws a lot making them better at that distance, and worse at slightly smaller ones. I would say a probability of coefficient being positive is at most a few percents, but not zero. The coefficient also shouldn't be too large, because there is still a large amount of chance and different external circumstances involved when making shots. My opinion for coefficient on standardized data is normal distribution with mean -0.3 and standard deviation 0.15 (which makes probability of it being positive about 2%).

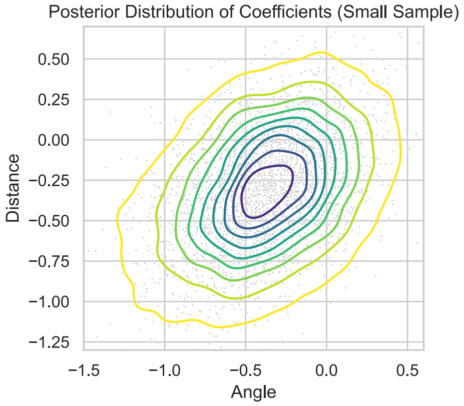
Questions

Let β_1 be the coeficitent for angle and β_2 for distance. We can answer the question whether the distance is more important for shot success than angle by checking what the probability $P(|\beta_2| > |\beta_1|)$ and if it is larger than 0.5. In this case it is about 1.00, which means distance is basically certainly more important tha angle. To check if shot success increases with angle, calculate the probability $P(\beta_1 > 0\ )$ of coefficient for angle being positive. In my case it is 0.23, so it is not likely that success increases with angle. Both of these can be seen in the plot below.



Plots





In the plots above, we can see that samples are situated in the similar general area, however they are way more spread out for the smaller datatest (note the different axes). This is due to smaller sample size making our estimates less certain. Probability of specific coefficients is more focused in a single area for larger data size. The aree with the highest posterior density is also not exactly at the same place for both, probably because the smaller dataset is too small and therefore not really representative of the whole data. In general, more data allows us for less uncertainty.