A Project Report

On

Beta Distribution using Bayes' Theorem

BY

Kumar Pranjal (2018A7PS0163H)

Sneh Lohia (2018A7PS0171H)

Abhishek Mishra (2018A7PS0019H)

Under the supervision of

Dr NL Bhanu Murthy

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(OCTOBER, 2020)

Assignment 1: Beta Probability Distribution

- The following formula gives the beta distribution:

$$Beta(x|\ a, b) = \Gamma(a + b)\ x^{a-1}\ (1 - x)^{b-1}\ /\ \Gamma(a)\ \Gamma(b)$$

- For this assignment, the prior probability distribution used in beta distribution with parameter values of a and b being 2 and 3 respectively. These values were used to ensure that the prior mean becomes 0.4 (a/(a+b)).

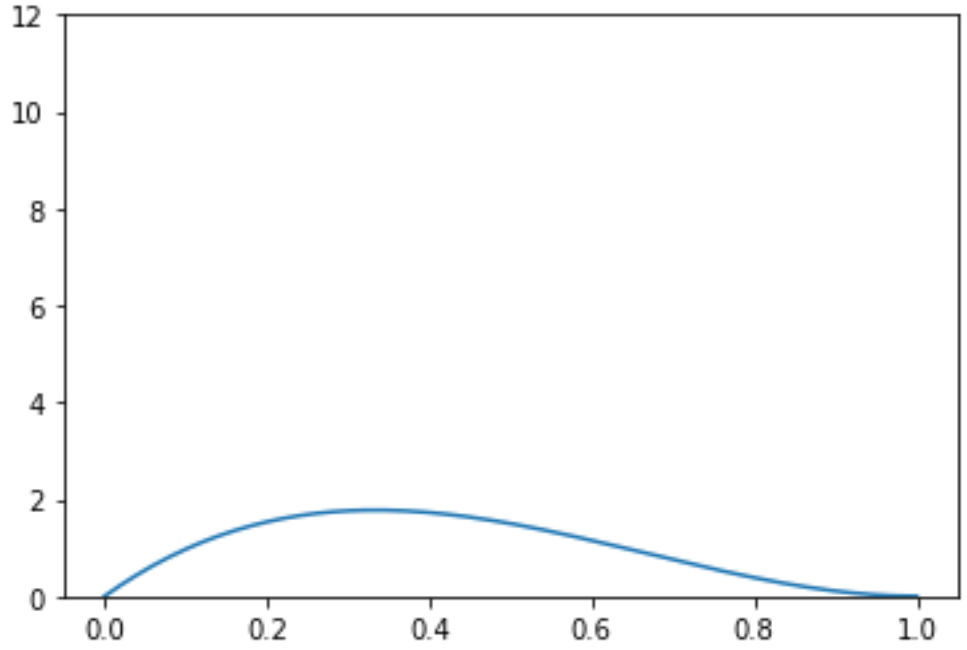
- The dataset used has 160 entries consisting of 112 instances of 1 (denoting heads) and 48 instances of 0 (denoting tails). The maximum likelihood estimator of the mean (μ_{ML}) is 0.7. All these entries are shuffled using NumPy's shuffling method (np.random.shuffle()). This dataset ensures that the maximum likelihood estimator of the mean does not lie in the range (0.4, 0.6). The shuffling is done to ensure the random distribution.

Steps in building the dataset:

Mentioned below are the steps involved in building the dataset. They are:

1. **Array Generation:** A NumPy array is generated with a given size and given mean. It consists of 1s and 0s sequentially.
2. **Shuffling:** The generated array is shuffled to simulate the random distribution of the data.

The prior distribution function for the mean is as follows :



Building the model:

Bob's Approach

To find the maximum likelihood estimator, Bob used the sequential learning method. He is taking one example at a time, and since the dataset consists of 160 entries, the whole process will be repeated 160 times.

- For each head/tail in the dataset, the values of a and b (initially 2 and 3) will be updated.
- The posterior becomes the prior for the next example after viewing the first example.
- Using this approach, the final plot for posterior curve obtained is as follows:

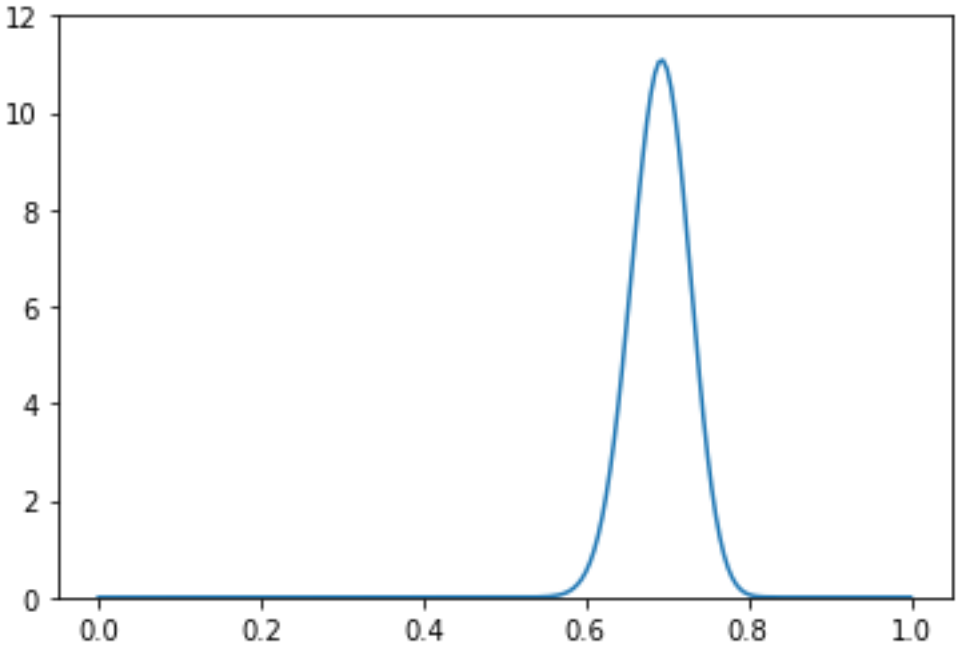


Fig. Posterior Curve for Bob

- With each data point read, the mean of data changes as follows:

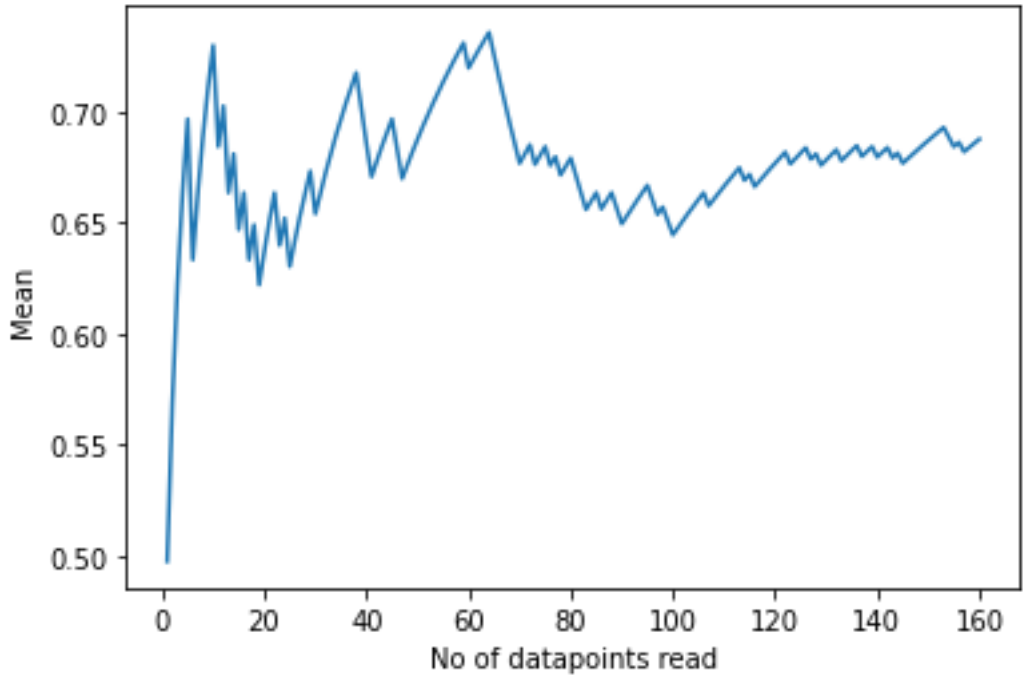


Fig. Variation of mean

Lisa's Approach

Lisa uses the complete dataset at once.

- In this case, the likelihood will be over the entire dataset. Hence, the values of a and b directly get updated to a+number of heads and b+number of tails respectively. The new values are then used to calculate the posterior probabilities.
- Using this approach, the final plot for posterior curve obtained is as follows:

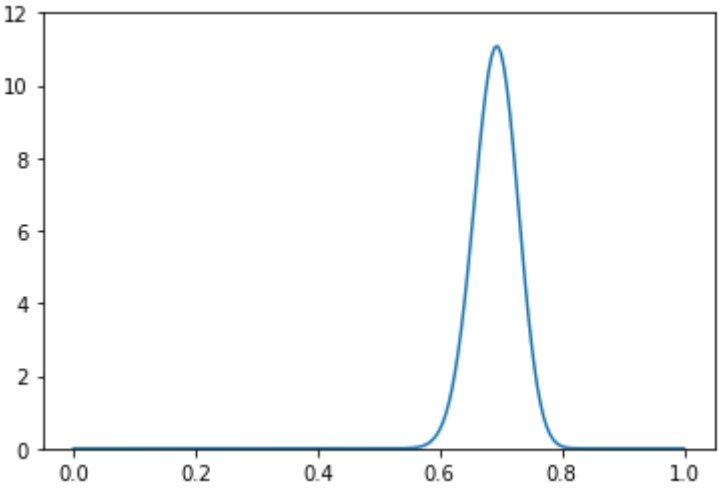


Fig. Posterior Curve for Lisa

Comparing the Approaches

In Bob's approach (sequential approach), observations are recorded one at a time and deleted before recording the next observation, and hence it can be used for extrapolation of data. Also, this method is slower for large datasets because the whole dataset is not read at once, and calculation is done for each data point. As visualised in the GIF, the sharpness of peak increases on increasing the number of observations. It indicates that increasing the number of known data decreases the uncertainty in posterior probability.

In Lisa's approach, the whole dataset becomes available at once. Hence it becomes computationally easier to solve such problems. The final curve for posterior is the same for both the approaches, as can be seen in the following figure.

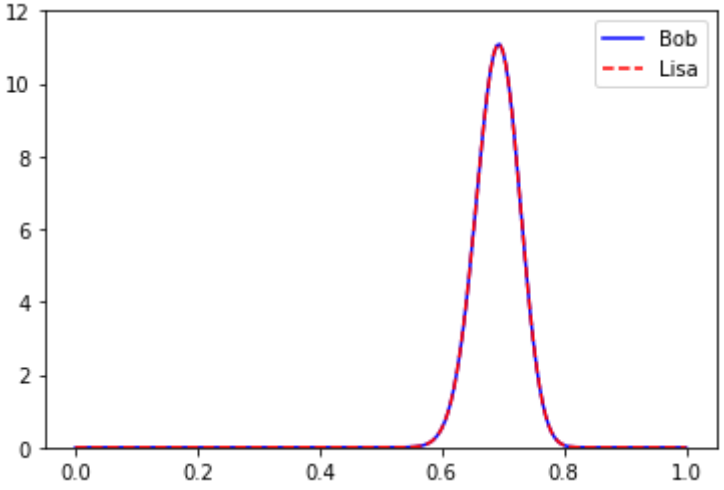


Fig. Comparison of both the curves

Questions to Ponder On

Q1. The size of the dataset has been restricted to 160 data points. What happens if more points are added (say ~10^5)? What would the posterior distribution look like if $\mu_{ML} = 0.5$? Which model, Bob's or Lisa's, would be more helpful and easier while working with large real-time data and why?

- Suppose more points are added to the dataset. In that case, the parameter values a and b will change accordingly, and the peak of the curve will become sharper, because of lesser variance in the beta distribution.
- If μ_{ML} is set to be 0.5, the mean of posterior function will become less. Hence the curve will shift leftwards.
- The prior mean is 0.4, while the likelihood mean is changed from 0.7 to 0.5. For our dataset consisting of 160 entries, the likelihood of 0.5 will imply an equal number of ones and zeros(80 each) and the posterior mean will attain a value equal to 0.488.
- Bob's model will be more useful for working with real-time data as it does not require all the data at once and updates itself with every new data point in the data set. So for large real-time data, rather than importing the whole data again and again, Bob's model will import only new data and change the function accordingly.

Q2. What if another distribution like Gamma, Gaussian or Pareto were to be chosen as the prior? Would the posterior computation be easier or difficult, and why?

- If another distribution like Gamma, Gaussian or Pareto were to be used as the prior, the curve would have been changed, and the curves for both the approaches may not overlap.
- Also, it would have been computationally difficult if other functions were to be used since for beta distribution, prior and posterior distributions are similar in terms of powers of x.