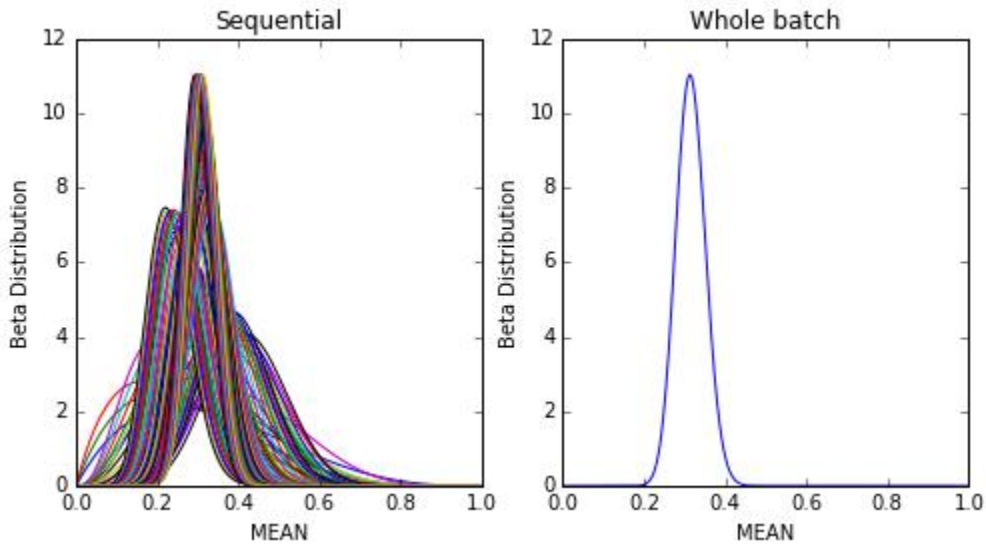
CS F320: Foundations of Data Science
Assignment 1 Report

Siddarth Gopalakrishnan - 2017B3A71379H
Rohan Maheshwari - 2017B4A70965H
Satvik Vuppala - 2017B4A71449H

Parameters:

As stated in the assignment, we perform a total of 160 coin tosses, such that the mean of the distribution doesn't lie between 0.4 to 0.6. For this purpose, we chose a distribution with 50 heads (1) and the remaining 110 tails (0) and shuffled the data for random distribution. Thus, our m=50 and N-m=110, making the mean of the distribution as m/N = 50/160 = 0.3125, which is less than 0.4.

We are also required to choose such parameters for beta distribution such that the mean of the prior = 0.4. To accomplish this, we select the hyperparameters, alpha and beta as 2 and 3 respectively, because we know that the mean of beta distribution with parameters alpha and beta is alpha/(alpha+beta). Thus, with alpha and beta values as 2 and 3, respectively, the value of the mean of prior turns out to be 2/(2+3) = 2/5 = 0.4.



Comparing Bob's and Lisa's approaches:

The above image shows the distributions for Bob's sequential approach(left) and Lisa's whole batch approach(right). In Bob's sequential approach, we adopt a Bayesian viewpoint. The learning in each iteration is independent of the prior and likelihood and only depends on the assumption of independent and identically distributed data, thus, in each iteration, the previous observations are discarded. As

we can observe from the above figure, the peak of the beta distribution becomes sharper as the number of observations increases. From this, we can conclude that as the number of observations increases, the uncertainty represented by the posterior decreases steadily.

In Lisa's whole batch approach, the hyperparameters are updated by the number of ones and zeros present in the distribution (50 and 110 respectively). Thus, alpha = alpha+50 = 52, and beta = beta+110 = 113, thus making the posterior value alpha/(alpha+beta) = 52/165 = 0.315 which lies between the likelihood mean (0.3125) and the prior mean (0.4) as expected in the case of a finite dataset.

The similarity between Bob's and Lisa's approaches is that the final posterior distribution for both the approaches is the same (11.0310962699) with hyperparameters alpha=52 and beta=113.

If more data points are added, the alpha and beta hyperparameters will increase as the number of ones and zeros increase. The curve will become more sharply peaked as the variance of the distribution decreases. This is because, as per the expression of the variance of the beta distribution, the variance decreases as alpha and beta increases. If the dataset is infinitely large, variance almost goes to zero and our posterior becomes equal to our likelihood estimator. Thus, we can conclude that as we get a larger dataset, we are more confident of our prediction and hence the variance decreases.

If the mean of maximum likelihood estimator is kept as 0.5, the mean of the posterior will shift more towards the right (it will be somewhere between 0.4 and 0.5). Earlier the mean of the posterior was somewhere between 0.3125 and 0.4, now the prior mean is 0.4 while the likelihood mean is changed from 0.3125 to 0.5. For a dataset of 160 points likelihood of 0.5 will imply the dataset having 80 ones and 80 zeros and the posterior mean will become 0.497, as alpha=82 and beta=83.

In the case of real time data, as there's a steady inflow of data, it would be better to use Bob's sequential approach. As mentioned earlier, the sequential approach works on iid data, so the previous observations are discarded in current iteration so, they can be used for large amounts of data as we need not load the entire data into the memory. For large real time data, it would be better to use a mixture of both these

approaches by considering one batch of data at a time. This way, the number of computations would also reduce as compared to Bob's approach.

If instead of Beta distribution, we use some other distribution for prior such as Gaussian or Gamma or Pareto distribution, the computation for calculating posterior becomes difficult. The posterior distribution takes the form of the prior distribution at the end of every iteration, from Bayes' theorem, in the sequential approach and acts as the prior for the next set of data. The prior distribution must, therefore, represent the distribution followed by the mean. Beta distribution for prior thus acts as a conjugate prior, as it has simple analytical properties and is of the form proportional to the powers of mean and 1-mean. This leads the posterior to have same form as the prior, thus leading to a simplified Bayesian analysis by maintaining conjugacy properties. Thus, Beta distribution is better suited for binomial distributions. We generally use Gaussian prior for data which follows a distribution of continuous variables, Dirichlet for multinomial distributions, Gamma for exponential distributions, etc.