Programming Project – 2

Experiments with Bayesian Linear Regression

Harika Kanakam

Algorithm 1 – Maximum Likelihood Algorithm:

- The weights of the train datasets are calculated by using the equation 3.28 from [B] with lambda as 0.

$$w = \lambda I + \Phi^T\Phi^{-1}\Phi^T\mathbf{t}.$$

- After calculating the weights, Mean Square Error can be calculated for the test datasets.

Algorithm 2 – Bayesian Linear Regression:

- For selecting the model, α and β must be calculated. An iterative algorithm is run to select these values.
- The iterative algorithm is initialized by selecting a random value of α and β in the range of 1-10.
- Using these initial values, m_N and S_N are calculated using the equations 3.53 and 3.54 from [B].

$$m_N = \beta S_N \Phi^T\mathbf{t}$$
$$S^{-1} = \alpha I + \beta\Phi^T\Phi$$

- Using these m_N and S_N values, α and β are calculated using the equations 3.91, 3.92 and 3.95 from [B].
- The termination condition for this iteration is when the values of α and β are just changing with minute difference (When creating the algorithm, I have considered the minute difference to be 0.01).
- Once α and β are known, λ is calculated as (α/β).
- After calculating the value of λ, weights are calculated using the same equation used in Algorithm 1.
- Once the weights of the train datasets are known, the error on the test dataset can be calculated easily.

Task 1: Comparing the Bayesian algorithm to Linear Regression with and without Regularization:

- Using the above two algorithms, the crime and housing datasets are trained with training fractions, f ∈ {0.1, 0.2, 0.3, . . ., 1.0} of the dataset size.
- After training the algorithm using train datasets with different sizes, error on test dataset is calculated for both the crime and housing datasets.

1.1

- Part 1 of the task 1 is to run the model selection algorithm and report the values of α, β and effective λ for each train size.
- The values of α, β and effective λ for each train size is represented in the below table for each dataset (i.e. Crime and Housing datasets).

********** Crime Task 1.1 Summary **********

	frequency	size	MLE MSE	alpha	beta	lambda	Bayesian MSE
0	0.1	169	1.310217	542.350572	2.346505	231.131224	0.370318
1	0.2	339	0.496743	772.872678	2.571839	300.513589	0.332878
2	0.3	508	0.456349	796.239673	2.527810	314.991849	0.333974
3	0.4	678	0.387147	819.790840	2.589339	316.602405	0.328017
4	0.5	847	0.371301	804.985451	2.691104	299.128287	0.329360
5	0.6	1017	0.364471	735.624785	2.760456	266.486704	0.324691
6	0.7	1186	0.354847	718.357090	2.895915	248.058763	0.323662
7	0.8	1356	0.348305	681.031185	2.957800	230.249238	0.325988
8	0.9	1525	0.346900	651.682984	2.901564	224.597162	0.326295
9	1.0	1695	0.345057	591.674284	2.958439	199.995428	0.329118

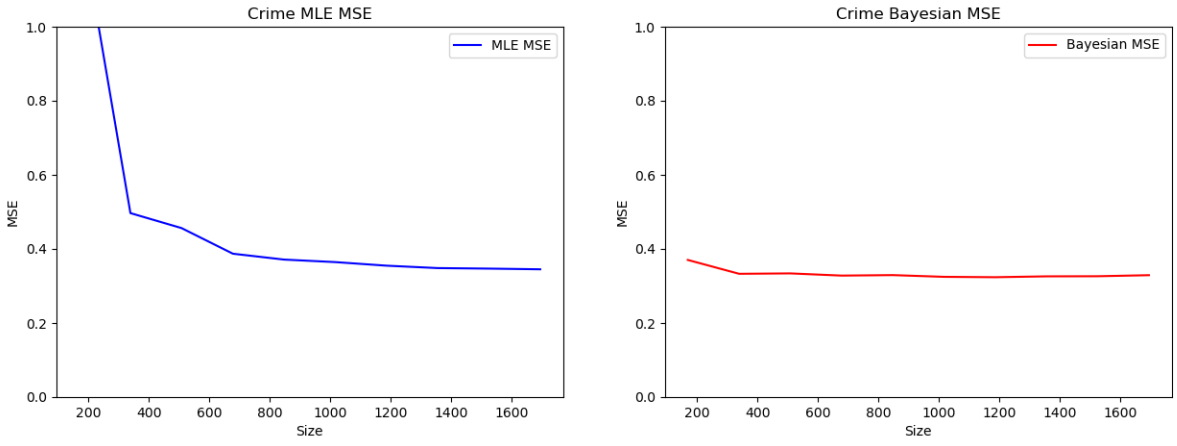
********** Housing Task 1.1 Summary **********

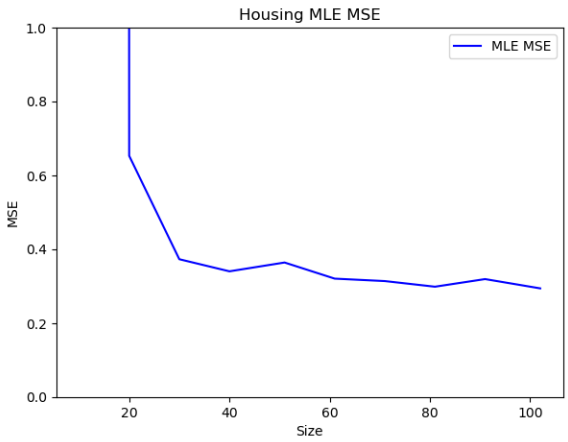
	frequency	size	MLE MSE	alpha	beta	lambda	Bayesian MSE
0	0.1	10	279.085739	81.625532	4.409624	18.510768	0.562160
1	0.2	20	0.653145	39.161003	2.124710	18.431223	0.371433
2	0.3	30	0.373280	23.444386	3.527736	6.645731	0.311566
3	0.4	40	0.340572	23.513021	4.558305	5.158281	0.305660
4	0.5	51	0.364402	22.472696	4.131391	5.439499	0.316010
5	0.6	61	0.320806	23.122948	4.511692	5.125117	0.298254
6	0.7	71	0.314182	21.568645	4.270257	5.050901	0.292834
7	0.8	81	0.298941	21.881811	4.452979	4.913971	0.286668
8	0.9	91	0.319573	18.805707	4.055429	4.637169	0.301477
9	1.0	102	0.294433	22.778575	3.983449	5.718304	0.288264

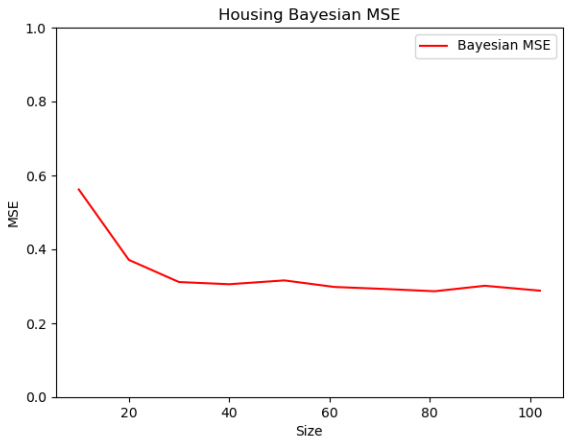
1.2

- Both the Maximum Likelihood Algorithm and Bayesian Linear Regression algorithms are run to get the mean squared error for both the crime and housing test datasets using the weights generated from each different sizes of their respective training datasets.

Below are the plots of their test set MSE as a function of their training set size. Limiting the Y-axis of the plot to the range [0,1] to ensure visibility of differences.







- From the above plots, it is observed that, for both the datasets, error is minimum using the Bayesian algorithm.
- In the maximum likelihood algorithm, the value of λ is 0. Hence the model results in more error. Here for both the datasets, when the train size frequency is 0.1, MLE is outside the limit.
- Whereas in the Bayesian Algorithm, λ is generated using the model selection.
- This phenomenon is observed because MLE is the point estimate whereas Bayesian is the PDF. Because of this, even with the small training size, Bayesian algorithm generates a reasonable error.
- Also, it is clearly visible that with the increase in the size of the training dataset, the error is decreasing. This happens as the size of the training size in increasing, both models are trained well as there is more data.
- The results are expected due to the above 2 reasons.

1.3

- In this the task 1.2 is repeated but with different λ values which are 1.0, 33.0, 100.0, 1000.0.
- Here, for each training size, MSE is calculated using the Bayesian Algorithm and with each value of λ.
- The summary of this task is as below.

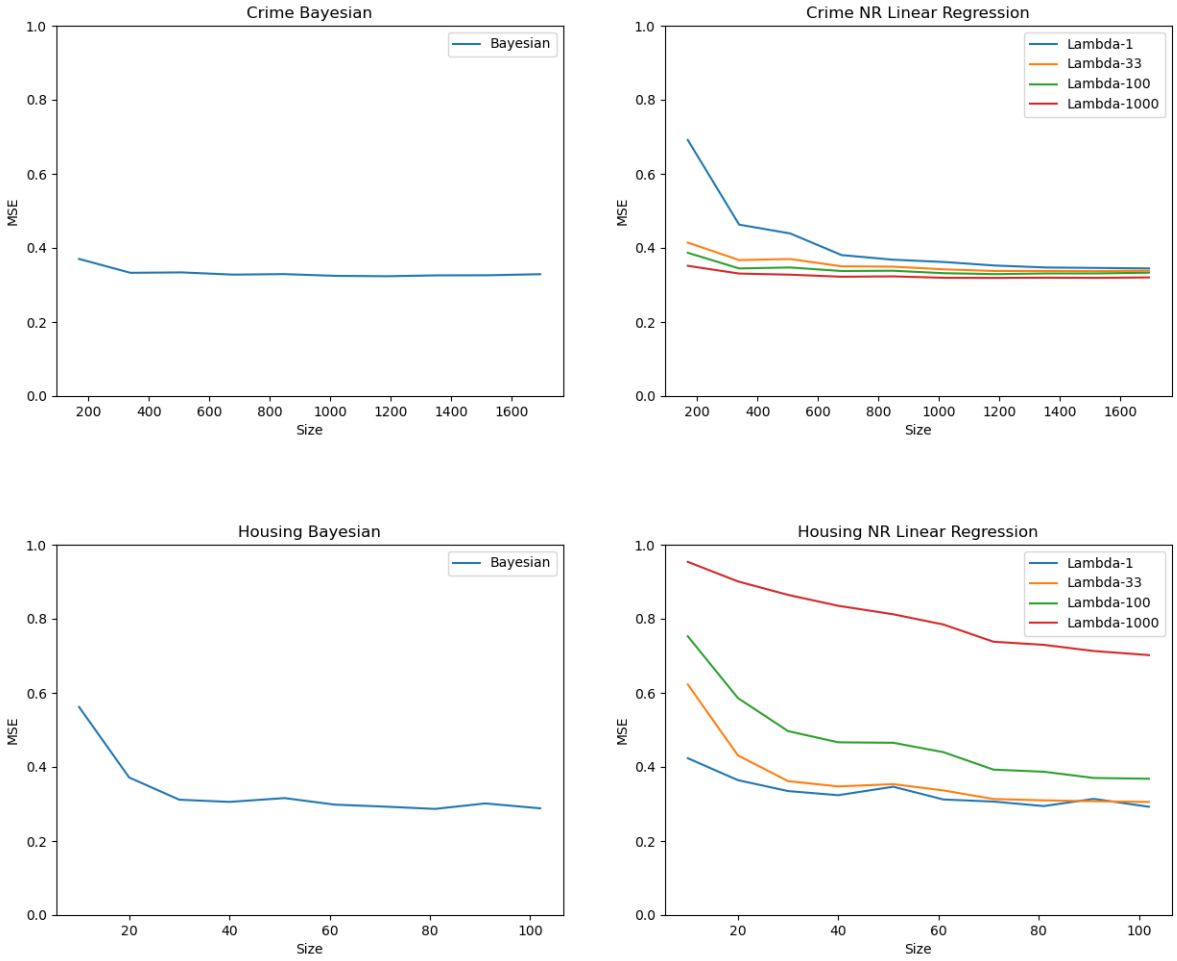
********** Crime Task 1.3 Summary **********

	Frequency	Size	Bayesian MSE	lambda 1	lambda 33	lambda 100	lambda 1000
0	0.1	169	0.370305	0.691600	0.414450	0.386958	0.351757
1	0.2	339	0.332878	0.462915	0.367220	0.344642	0.330831
2	0.3	508	0.333974	0.439131	0.370078	0.347218	0.327807
3	0.4	678	0.328017	0.380736	0.350368	0.337721	0.322280
4	0.5	847	0.329360	0.368317	0.349493	0.338453	0.323217
5	0.6	1017	0.324691	0.362030	0.342152	0.331806	0.319495
6	0.7	1186	0.323662	0.352454	0.337760	0.329351	0.319143
7	0.8	1356	0.325988	0.347213	0.337913	0.331136	0.319703
8	0.9	1525	0.326295	0.345970	0.337386	0.331233	0.319271
9	1.0	1695	0.329118	0.344762	0.338535	0.333185	0.320424

********** Housing Task 1.3 Summary **********

	Frequency	Size	Bayesian MSE	lambda 1	lambda 33	lambda 100	lambda 1000
0	0.1	10	0.561967	0.423567	0.622894	0.752847	0.953728
1	0.2	20	0.371440	0.364347	0.431200	0.586031	0.900934
2	0.3	30	0.311566	0.334882	0.361677	0.496773	0.864581
3	0.4	40	0.305660	0.323709	0.347422	0.466524	0.835124
4	0.5	51	0.316009	0.346568	0.353528	0.465232	0.812189
5	0.6	61	0.298254	0.312135	0.336586	0.439911	0.784576
6	0.7	71	0.292834	0.306493	0.313402	0.392739	0.738035
7	0.8	81	0.286678	0.294359	0.309810	0.387159	0.729700
8	0.9	91	0.301477	0.313780	0.307367	0.370237	0.713008
9	1.0	102	0.288264	0.292436	0.305441	0.368111	0.701970

Below are the plots of both the algorithms for both the datasets as a function of their training set size. Limiting the Y-axis of the plot to the range [0,1] to ensure visibility of differences.



- Using the given λ values, after training the models and calculating the MSE, it is observed from the graph that with increasing λ value for crime dataset, error on the test dataset is decreasing for each train size. But whereas, with increasing λ value for housing dataset, error on the test dataset is also increasing for each train size.
- Hence, a single universal value for λ can't be used.

- For Bayesian algorithm, as λ value is generated using the model selection, the algorithm is more reliable. Also, with increasing train size, the error on test set decrease.
- Hence, Bayesian algorithm is successful in selecting a good value.

Task 2: Bayesian model selection for Parameters and Model order:

- In this task, models are trained using the f3 and f5 train datasets and tested on their respective test datasets.
- The train datasets have only x values. The algorithms are run using the polynomial degrees d {1, 2, ..., 10}.
- For each degree d, Bayesian Model selection is run to calculate α, β and effective λ. Using the α and β values, log evidence is calculated.
- Also, MLE algorithm is run for each train dataset and MSE is calculated for their test datasets.
- Below are the summary details of this task.
-

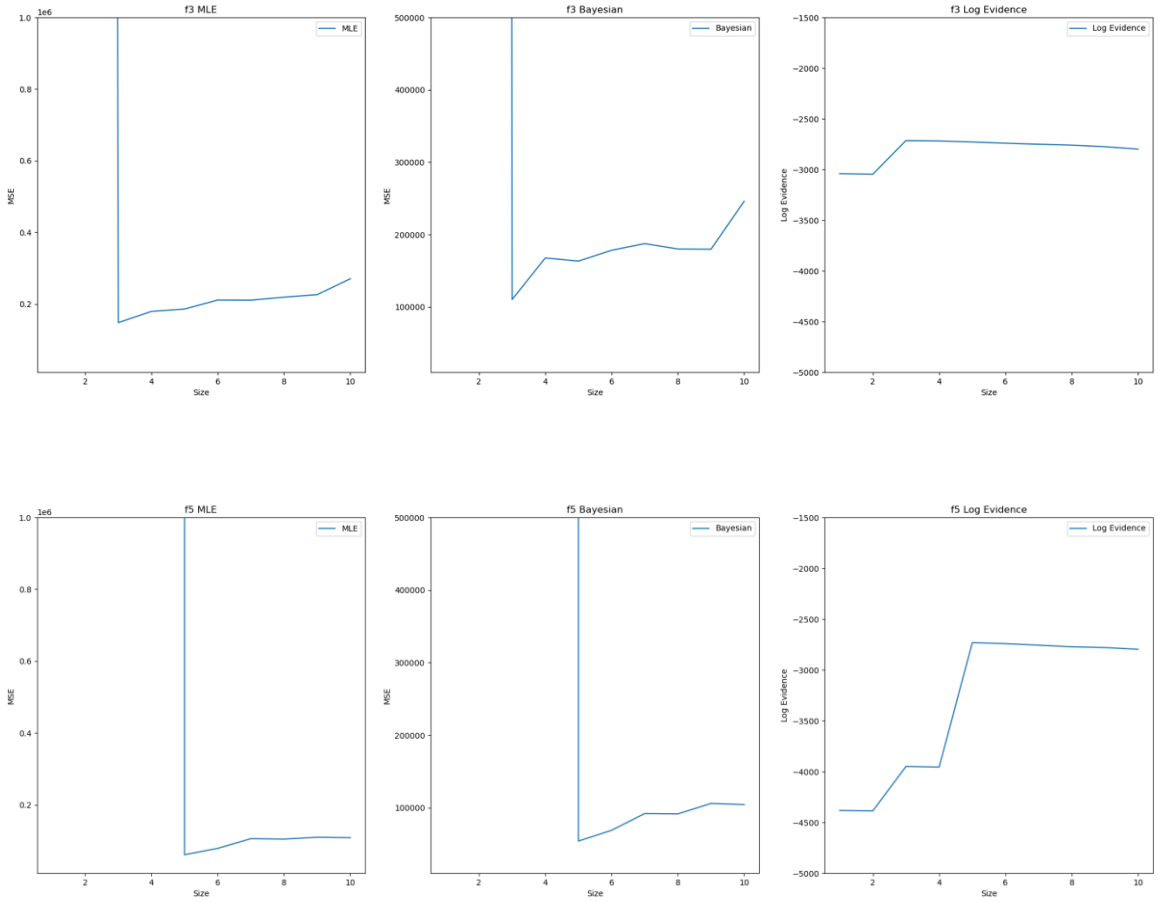
********** f3 Task 2 Summary **********

	Dimension	MLE MSE	Bayesian MSE	Log Evidence
0	1	3.938914e+07	3.937623e+07	-3041.401249
1	2	3.949576e+07	3.952914e+07	-3046.882511
2	3	1.484294e+05	1.104733e+05	-2715.216592
3	4	1.796275e+05	1.675996e+05	-2718.780643
4	5	1.862636e+05	1.633963e+05	-2728.683358
5	6	2.113706e+05	1.780880e+05	-2740.828219
6	7	2.110301e+05	1.875683e+05	-2750.924313
7	8	2.192299e+05	1.801470e+05	-2759.828832
8	9	2.263459e+05	1.797993e+05	-2775.536960
9	10	2.707748e+05	2.461465e+05	-2799.375097

********** f5 Task 2 Summary **********

	Dimension	MLE MSE	Bayesian MSE	Log Evidence
0	1	3.416427e+11	3.414064e+11	-4382.518102
1	2	3.411956e+11	3.413401e+11	-4386.909345
2	3	1.746560e+10	1.750096e+10	-3949.774297
3	4	1.743566e+10	1.748018e+10	-3956.007247
4	5	6.137535e+04	5.423972e+04	-2732.237580
5	6	7.904305e+04	6.907680e+04	-2741.372054
6	7	1.060316e+05	9.215726e+04	-2756.884585
7	8	1.048404e+05	9.174164e+04	-2772.596661
8	9	1.100656e+05	1.060432e+05	-2780.163590
9	10	1.088325e+05	1.044787e+05	-2797.345466

Below are the plots of log evidence and both the algorithms for both the datasets as a function of their training set size. Limiting the Y-axis of the log evidence plot to the range [-5000, -1500] and Y-axis of both algorithms to [10000, 1000000] to ensure visibility of differences.



- For f3 dataset, the MSE values of both algorithms are less for the dimension 3 and log evidence is high for the dimension 3.
- For f5 dataset, the MSE values of both algorithms are less for the dimension 3 and log evidence is high for the dimension 5.
- Therefore, log evidence can be successfully used to select the values of α, β and d for Bayesian method.
- Even the non-regularized models fare in these results in selecting these values.